Linear programming approach to partially identified econometric models

Andrei Voronin*

December 19, 2024

PRELIMINARY AND INCOMPLETE
Link to latest version

Abstract

The bounds on a parameter of interest in partially identified settings are often given by the values of linear programs. This paper studies estimation and inference for the value $B(\theta) = \min_{Mx \geq c} p'x$, where $\theta = (p, M, c)$ is estimated. We develop the first $\sqrt{n}-$consistent estimator that does not require additional restrictions. Unlike existing methods, our estimator remains valid under point-identification, over-identifying constraints and solution multiplicity. Exact and computationally simple inference procedure is developed. Turning to uniform properties, we prove that there exists no uniformly consistent estimator absent further conditions. We propose the '$\delta-$condition', under which our estimator is uniformly consistent. $\delta-$condition does not rule out economically relevant problematic scenarios, covers the unrestricted set of measures in the limit, and is strictly weaker than all previously proposed restrictions. We complement our estimation approach with a general identification result for models described by affine inequalities over conditional moments (AICM), potentially augmented with relevant almost sure restrictions on the potential outcomes and missing data conditions. Sharp bounds on affine treatment parameters under AICM are shown to take the form of $B(\theta)$. Our results allow applied work to employ previously intractable conditions, including arbitrary combinations of existing restrictions, and conduct sensitivity analysis. We apply our findings to estimating returns to education. For that, we develop the conditionally monotone IV assumption (cMIV) that tightens classical bounds. We argue that cMIV remains unrestrictive relative to the classical conditions and provide a formal test for it. Under cMIV, university education in Colombia is shown to increase the average wage by at least 5.91%. In contrast, classical conditions fail to produce an informative bound.

1. Introduction

Nonparametric bounds analysis offers a powerful alternative to classical causal inference methods in the absence of a credibly exogenous instrument. In many cases, sharp bounds

*Department of Economics, UCLA. Email: avoronin@g.ucla.edu. I am grateful to Andres Santos, Denis Chetverikov, Rosa Matzkin, Jinyong Hahn, Bulat Gafarov, Tim Armstrong, Kirill Ponomarev, Shuyang Sheng and Manu Navjeevan as well as to all the participants of the 2024 California Econometrics Conference and the 2024 European Winter Meeting of the Econometric Society for the valuable discussions and criticisms.

on the partially identified parameter of interest are given by the values of linear programs (LP) that depend on identified functionals of the underlying probability measure. The bound of interest is a feature of a probability measure, $B(\mathsf{P}) = B(\theta_0(\mathsf{P}))$, of form:

$$B(\theta) \equiv \min_{Mx \geq c} p\, x \tag{1}$$

where $\theta_0(\mathsf{P})$ is the true value of the parameter $\theta = (p\,, \mathrm{vec}(M)\,, c\,)$, estimated from the data via a $\sqrt{n}$-consistent estimator $\hat{\theta}_n$. In that context, $\Theta_I = \{x \quad \mathsf{R}^d : Mx \geq c\}$ is the identified set for the partially identified feature x.

The problem (1) exhibits non-regular behavior. The most problematic scenario occurs when the model is close to point-identification along some feature x. In that case the quality of existing estimators is severely reduced. This results in an unattractive tradeoff between identification power and estimation quality for partially identified models. Figure 1 illustrates this.

To address that issue, we develop *the debiased penalty function estimator* $\hat{B}(\hat{\theta}_n; w_n)$, where:

$$\hat{B}(\theta; w) \equiv \sup_{x \;\; \tilde{A}(\theta; w)} p\, x, \quad \tilde{A}(\theta; w) \equiv \arg\min_{x \;\; X} p\, x + w\iota\, (c - Mx)^+, \tag{2}$$

and $w_n \to \infty$ is the penalty parameter, with $w_n << \sqrt{n}$. Only assuming that Θ_I is non-empty and contained in a known compact X, we show that $\hat{B}(\hat{\theta}_n; w_n)$ is $\sqrt{n}-$consistent for any w_n satisfying the above conditions[1]. In contrast, other recently developed estimators are either not applicable under no further assumptions, or inconsistent for $B(\theta_0)$. For example, the plug-in estimator $B(\hat{\theta}_n)$ is not consistent and may also fail to exist with arbitrary non-vanishing probability. An alternative estimator that we develop based on set-expansion theory is only $\sqrt{n}/\kappa_n$-consistent for some $\kappa_n \to \infty$.

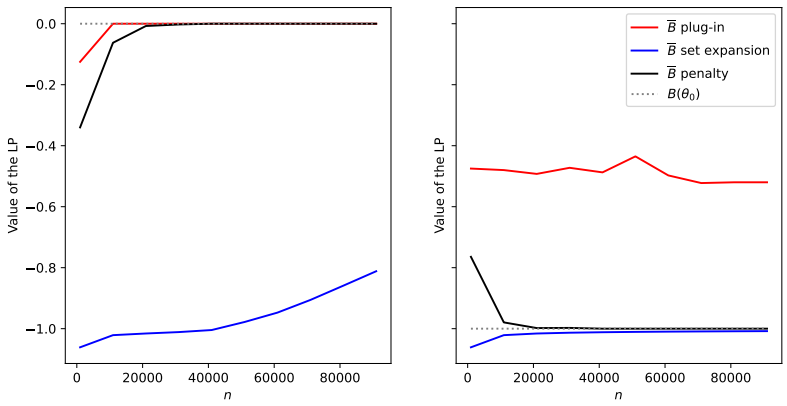


Figure 1: Comparison of estimators for two measures with the true values of 0 and −1, left to right. Average values over 400 simulations.

[1]The selection of this parameter and its relation to uniform properties of the estimator is discussed in the Appendix in great detail.

We obtain an asymptotically normal version of our estimator via sample-splitting and construct confidence regions with *exact coverage*. The previously suggested resampling methods applied to $B(\hat{\theta}_n)$ are shown to be inconsistent for the distribution even when the model is far from point-identification, while other recently developed procedures are either not applicable or asymptotically conservative.

Using an equivalence between unconstrained piecewise-linear problems and auxiliary linear programs, one can compute our exact penalty estimator in polynomial time using a LP solver. This, combined with the closed-form expression for the asymptotic variance, makes our inference procedure the most computationally efficient in the existing literature[2].

Turning to uniform asymptotic theory, we first establish that there exists no uniformly consistent estimator of $B(\mathsf{P})$ over the unrestricted set of measures $\boldsymbol{P}$. We propose to consider the '$\delta-$condition', restricting $\boldsymbol{P}$ to $\boldsymbol{P}^{\delta}$ - the measures at which the smallest singular value of a full-rank submatrix M_J of constraints binding at an optimal vertex is lower bounded by $\delta > 0$. This condition is minimal in a sense that for any measure $\mathsf{P} \in \boldsymbol{P}$ there exists $\delta > 0$ such that $\mathsf{P} \in \boldsymbol{P}^{\delta}$, i.e. the condition spans the unrestricted set in the limit: $\bigcup_{\delta>0} \boldsymbol{P}^{\delta} = \boldsymbol{P}$. In contrast to conditions found in previous work, for reasonable values of δ, this restriction does not rule out economically relevant *problematic* scenarios, such as solution multiplicity, point-identification and over-identification. Our estimator is shown to be uniformly consistent over $\boldsymbol{P}^{\delta}$ for any fixed $\delta > 0$.

We complement our estimation results by developing a general identification framework in Section 3. In particular, we derive sharp bounds for a broad class of treatment parameters of interest[3] under arbitrary affine inequalities over conditional moments (AICM), potentially augmented with affine almost sure restrictions and missing data conditions. In the simplest case, AICM identifying restrictions have the form:

$$M \ (\mathsf{E}[Y(d)|T=t, Z=z])_{d,t,z} + b \ \geq 0 \quad \text{and} \quad \tilde{M}(Y(d))_d + \tilde{b} \geq 0 \text{ a.s.}, \tag{3}$$

where $(Y(d))_d$ are continuous potential outcomes corresponding to the legs of treatment T and Z are other covariates. Identified matrices M , $\tilde{M}$ and vectors b , $\tilde{b}$ are chosen by the researcher. Our approach accommodates arbitrary combinations of existing 'nonparametric bounds' restrictions, allows to conduct sensitivity analysis, and extends to more complex conditions where sharp bounds were previously unavailable[4]. We show that under (3) the sharp lower (upper) bound on the parameter of interest corresponds to the minimum (maximum) value of a linear program (LP), whose parameters are learned from the data, as in (1). In the context of AICM, θ is a function of observed conditional moments $(\mathsf{E}[Y|T=t, Z=z])_{t,z}$ and the identified joint distribution of T, Z, while x collects relevant unobserved conditional moments.

Finally, we develop an application of our approach to estimating returns to education in Colombia. To that end, in Section 4 we first introduce a family of conditionally monotone

[2]Both Gafarov (2024) (BG) and Cho and Russell (2023) (CR) rely on resampling methods, which require to compute one or multiple LPs at each iteration. Computing a confidence interval for a LP with 32 variables takes 16.81 seconds with the approach of BG and 40.65 seconds with the approach of CR, according to Cho and Russell (2023). Our approach requires computing a LP once, which takes around 0.0022 seconds on average.

[3]Including ATE and CATE, among other typically studied parameters, see Section 3.

[4]For example, cMIV and the mixture of all classical Manski and Pepper (2000) conditions, see below.

instrumental variables assumptions (cMIV), nested in (3), that impose:

$$\mathsf{E}[Y(t)|T \quad A, Z = z] - \text{ monotone in } z,$$

where $(Y(t))_{t \ T}$ are potential log-wage schedules corresponding to education levels T, and Z is a proxy of ability based on Saber test scores. Sets A parametrize different versions of cMIV and are chosen by the researcher[5]. The conditions we consider result in tighter bounds than those obtained under the classical monotone instrumental variables (MIV) assumption of Manski and Pepper (2000), and sharp bounds under cMIV are infeasible with previously developed methods. We argue that cMIV conditions, however, remain unrestrictive in many applications, including ours. While empirical literature has visually examined the monotonicity of the observed conditional moments, i.e. $A = \{t\}$, to justify applying MIV[6], we show that such monotonicity is instead equivalent to a particular form of cMIV given that MIV holds and under a mild regularity condition. The formal test of cMIV is obtained as an extension of Chetverikov (2019). When estimating the returns to education in Colombia using cMIV, we find that the effect of obtaining university education on average wage is at least as large as 5.91%. In contrast, the classical conditions fail to produce an informative bound.

This paper also contributes multiple auxiliary results. The first one is concerned with an important special case of (3) - the combination of all classical Manski and Pepper (2000) conditions. Since such combination possesses the greatest identifying power out of all classical restrictions, empirical work has attempted to use it even in the absence of a theoretical justification, obtaining bounds that were either not sharp, or invalid (Lafférs, 2013). Our method yields sharp bounds and a valid estimation procedure for that setting. Our second auxiliary result is a negative consequence of Le Cam's binary method. We show that if the estimand is a discontinuous functional of the probability measure, there can exist no uniformly consistent estimator. The third auxiliary contribution is relevant to the control theory literature studying ℓ_1-penalized solutions of system of linear inequalities. It is shown that the ℓ_1-deviation $\iota\ (c - Mx)^+$ from a non-empty and bounded polytope $\Theta_I = \{x \quad \mathsf{R}^d : Mx \geq c\}$ is bounded from below by $\frac{d(x,\Theta_I)\kappa(\Theta_I)}{d}$. Here $\kappa(\Theta_I)$ is what we term the *condition number* of a polytope - the smallest positive singular value across its vertices.

We briefly note the limitations of our approach. On the identification side, the absence of restrictions on treatment selection prevents us from studying more granular parameters, such as marginal treatment responses. Furthermore, our identification results are given for discrete treatment and instrument. An extension to the continuous case is feasible, but is outside the scope of this paper[7]. On the estimation side, while our estimator is pointwise $\sqrt{n}-$consistent in general, we only establish $\sqrt{n}/w_n$-uniform consistency for a slowly diverging sequence w_n[8]. We provide further evidence on the uniform rate of consistency in Appendix. A theoretically $\sqrt{n}-$uniformly consistent estimator follows from our analysis, but it depends on an unobserved parameter δ that seems impossible to estimate, so we

[5] We study various possibilities, e.g. all singletons $\{d\}$ and the full support $\mathcal{T}$. See Section 4 for details.

[6] This is true of De Haan (2017), among others.

[7] Even when continuous identification results are available, in practice estimation is still carried out with discretized covariates. This is true for all empirical work referenced below.

[8] Theoretically, w_n can diverge arbitrarily slowly. Practical guidance on its selection is provided below.

do not recommend using it in practice. Finally, while our inference procedure naturally extends to uniform setup under sufficient regularity conditions, exploring this is left for future work.

Relationship to literature

The strand of literature relevant to the estimation of (1) is concerned with statistical inference in the LP estimation framework. Semenova (2023) considers a LP with an estimated constraint vector c, but a fixed matrix M and a coefficient vector p. Mogstad, Santos and Torgovitsky (2018) develop a consistent estimator for problems with fixed constraint sets Θ_I. Methods developed under a fixed M assumption cannot be easily extended to the completely-perturbed setting, as will become evident in Section 2. Syrgkanis, Tamer and Ziani (2021) develop a testing procedure for the failure of LP feasibility. Gafarov (2024) develops uniform inference for a LP described by affine inequalities over unconditional moments, provided uniform Linear Independence Constraint Qualification (LICQ) and Slater's condition hold. This approach is problematic under AICM for three reasons. Firstly, θ is not a linear function of *unconditional* moments. Secondly, Slater's condition fails close to point-identification along some dimension, which may occur for a rich enough AICM. Thirdly, LICQ can be violated in the parameter-on-the-boundary scenario (see Andrews (1999) and Fang and Santos (2018)), which is relevant in the AICM models and corresponds to over-identification at the optimum. Andrews, Roth and Pakes (2023) develop uniform inference in a special case of LP estimation framework, which arises from their model. In their problem, the Slater's condition always holds and θ has a particular structure, making their findings hard to generalize. Cho and Russell (2023) introduce random distortions to the LP and leverage genericity results to establish uniform Hadamard differentiability of the resulting problem. This allows to derive uniformly conservative confidence regions for $B(\mathsf{P})$. We do not find this approach satisfactory in our application for four reasons. Firstly, as the approach of Gafarov (2024), it only applies to unconditional population moments. Secondly, zero local power may be more problematic in settings where the identifying power of restrictions is low to begin with, as may be the case with more robust AICM models. Thirdly, their approach relies on an arbitrarily selected bound for the support of the generated noise with little practical guidance as to its selection, and on a single realization of the noise itself, making it very susceptible to $p-$hacking. Lastly, the approach of Cho and Russell (2023) is, in fact, only practical in situations in which the Slater's condition holds. Otherwise, the procedure is based on another tuning parameter with no guidance on its selection. Our simulations demonstrate that this parameter is far from being innocuous, as the performance of the proposed estimator ranges from very conservative to invalid even in large samples, if that parameter is adjusted[9].

Partially identified models which result in bounds of form (1) are described in the review article by Kline and Tamer (2023). We discuss a subset of these models to motivate our approach. The nonparametric bounds analysis was pioneered by Manski (1997) and Manski and Pepper (2000, 2009). In Blundell et al. (2007) wages are only observed for employed individuals, prompting an extension of the approach to missing data. Boes

[9]In fact, this is exactly what motivates us to work under $w_n \to \infty$ asymptotics, instead of considering a fixed w.

(2009) considered 'parabola-shaped' IV, Siddique (2013) introduced a non-parametric Roy model and a simple cMIV in a $2 \times 2 \times 2$ case with random treatment assignment. Kreider et al. (2012) introduced corrections for under-reporting. De Haan (2017) applied the classical conditions to estimating the effect of providing secondary schools with additional resources for low-ability pupils, while Cygan-Rehm, Kuehnle and Oberfichtner (2017) considered the effect of unemployment on mental health. All of these cases are nested in AICM framework, and our sharp bounds often improve the bounds derived in these papers. To compute the bounds, the above papers use the plug-in estimator $B(\hat{\theta}_n)$ combined with bootstrap described in Imbens and Manski (2004) for inference. Our results on the estimation of (1) show that these procedures are only valid under strong assumptions that do not appear to be justified in these applications. Our estimator $\hat{B}(\hat{\theta}_n; w_n)$ and the developed inference procedure resolve that issue.

AICM approach complements the method of Mogstad, Santos and Torgovitsky (2018), who develop identification theory for generalized IV estimators and obtain bounds in form (1). By virtue of imposing a Heckman and Vytlacil (1999, 2005) treatment selection mechanism and working in the binary treatment case, they accommodate arbitrary a.s. restrictions on the shape of the marginal treatment response functions and produce bounds for a wider family of treatment parameters. Additive separability in treatment selection is equivalent to the Imbens and Angrist (1994) IV conditions under instrument exogeneity (Vytlacil, 2002). Even though our approach nests mean-independence conditions, it appears most useful when an IV is not available. For that reason, a separable selection mechanism is not justified for AICM[10]. AICM is not related to the model in Andrews, Roth and Pakes (2023) other than by virtue of resulting in bounds of form (1). While our inequalities are imposed *over* affine combinations of counterfactual conditional moments, the latter work effectively generalizes the regression framework to moment inequality restrictions on the error term. We are not aware of conditions, similar to those in Imbens and Angrist (1994), that would allow to state linear conditional moment inequalities models in the potential outcomes form.

Notation

All vectors are column vectors, and M' denotes the transpose of $M \in \mathsf{R}^{n \times m}$. If A is a set, A' stands for its complement. A collection $(x(j))_{j \in J}$ is a column vector. 2^A denotes the powerset of set A, and $\overline{m,n}$ is the collection of integers from m to n. $\times$ is a Cartesian product of sets, while $\otimes$ is the Kronecker product. The sign $\sqcup$ denotes a disjoint union. Signs $\wedge$ and $\vee$ stand for logical 'and' and 'or' operators respectively. If M is a $m \times n$ matrix and $A \subseteq \overline{1,n}$, M_A is the $|A| \times n$ submatrix of the rows with indices in A. If $j \in \overline{1,n}$, write $M_j \equiv M'_{\{j\}}$. $\mathcal{R}(M)$ stands for the range of M, and $\sigma_d(M)$ is the $d-$th largest singular value of M. The distance between a point x and a set A is written as $d(x, A) \equiv \inf\{||x - a|| : a \in A\}$, and $d_H(A, B) \equiv \max\{\sup_{b \in B} d(b, A), \sup_{a \in A} d(a, B)\}$ is the Hausdorff distance between sets A, B in a normed space. For a subset A of normed space S, $A^\varepsilon \equiv \{s \in S : d(s, A) < \varepsilon\}$ is its open expansion. $Int(A)$ and $Cl(A)$ are the interior and closure of A, while $Cone(A)$ is its convex cone.

[10]Thus, if one is faced with i) a binary treatment setup, ii) has a valid IV and iii) no outcomes' data is missing, the method of Mogstad, Santos and Torgovitsky (2018) may be used. If any of these conditions fail, our approach is an alternative.

If A is a matrix, $Cone(A) \equiv Cone(\mathcal{R}(A))$. $s(x, A) = \max_{a \in A} x'a$ for a compact $A \subseteq \mathsf{R}^d$ and $x \in \mathsf{R}^d$ is a support function. For $v = (v_j)_{j \in \overline{1,d}}$, define $v^+ \equiv (\max\{v_j, 0\})_{j \in \overline{1,d}}$. For $v, u \in \mathsf{R}^d$ vector inequalities $v > u$ and $v \geq u$ mean $v_i > u_i \ \forall i \in \overline{1,d}$ and $v_i \geq u_i \ \forall i \in \overline{1,d}$ respectively. $\iota_d \in \mathsf{R}^d$ is a vector of ones, $I_d \in \mathsf{R}^{d \times d}$ is the identity matrix, and the subscript is dropped occasionally. Operator E_P is the expectation under a measure $P \in \mathcal{P}$, and the subscript is dropped whenever it does not cause confusion.

2. LP estimation framework

Parameters of interest can be represented as a value of a LP in many applications, most prominently in partial identification (see Mogstad, Santos and Torgovitsky (2018) and Andrews, Roth and Pakes (2023), among others). As shown in Section 2, that is also true for the sharp bounds on treatment parameters under the general class of AICM assumptions. This section develops asymptotic theory for such problems. We derive asymptotic results under the most general conditions, so that our estimation procedures remain useful even outside our identification framework. Throughout this section, we study the parameter of interest $B(\theta)$ of the following type:

$$B(\theta) \equiv \min_{Mx \geq c} p\, x \tag{4}$$

where $\theta \equiv (p\,, c\,, vec(M)\,)$ denotes the vector of all parameters of the initial linear program. Here M is a $q \times d$ matrix, c R^q and p R^d. The true value of these parameters at a fixed true measure will be denoted by θ_0 R^S, where $S = qd + q + d$. The value of interest is therefore $B(\theta_0)$.

Remark. Notice that (4) does not rule out equality constraints. That is because $Ax = b$ $Ax \geq b$ $-Ax \geq -b$.

We denote the constraint set by $\Theta_I(\theta) \equiv \{x$ $\mathsf{R}^d | Mx \geq c\}$ and omit the argument when θ_0 is concerned. In the context of identification results in Section 3 of this paper and in other applications (e.g. Mogstad, Santos and Torgovitsky (2018)), the set Θ_I is the identified set for an unobserved feature x of the underlying distribution.

Assumption A0 is maintained throughout this section, while the rest of the conditions are spilled out explicitly.

Assumption A0 (Pointwise setup). *Suppose that at the fixed true parameter θ_0:*

i) The identified set is non-empty, $\Theta_I(\theta_0) =$, and $\Theta_I(\theta_0)$ X for a known compact X

ii) There is a $\sqrt{n}$-consistent estimator $\hat{\theta}_n \equiv (\hat{p}_n, \hat{c}_n, vec(\hat{M}_n)\,)$:

$$||\hat{\theta}_n - \theta_0|| = O_p(1/\sqrt{n})$$

Our baseline setup therefore assumes that at the fixed true parameter θ_0 the identified set Θ_I for the feature x is non-empty. Intuitively, this means that the underlying model *cannot be rejected*, and does not imply that the identifying restrictions are correctly specified. Existence of a fixed known compact X that contains Θ_I is a mild restriction,

which is usually warranted in applications. Similarly, a $\sqrt{n}-$consistent estimator for θ_0 typically follows from CLT and the Delta-Method[11].

The following two solution sets will prove useful in the further discussion:

$$\mathsf{A}(\theta) \equiv \underset{Mx \geq c}{\arg\min}\ p\ x, \quad \Lambda(\theta) = \underset{M\ \lambda = p}{\arg\max}\ c\ \lambda.$$

Let us first introduce some terminology.

Definition. Slater's condition (SC) is the assertion that $Int(\Theta_I) = \ .$[12]

Intuitively, SC demands that there be no point-identification along any dimension of the identified set. This, of course, also precludes the situation of exact point-identification, in which Θ_I is a singleton. It can be shown that an 'approximate' failure of SC may also be a problem for the existing methods in finite samples, so that a stronger version of SC is usually imposed, see Gafarov (2024).

Definition. Linear independence constraint qualification (LICQ) is the assertion that the submatrix of binding constraints at any optimum is full-rank.

LICQ precludes the existence of overidentifying constraints. It can be problematic in rich AICM models, for example, if more than d constraints coincide at the optimum. Alternatively, one may think that it rules out the parameter-on-the-boundary scenario in Andrews (1999).

Definition. The notion of flat faces refers to the situation where $|\mathsf{A}(\theta_0)| = 1$.

Intuitively, this corresponds to solution multiplicity, namely the situation in which the bound on the parameter of interest is achieved at multiple partially identified features x.

Remark. Assumption A0 does not impose LICQ, nor SC and does not rule out the flat faces. These typically imposed conditions are discussed as a motivation for our approach.

Estimation and inference for a LP value are complicated by the fact that under assumption A0 the value function $B(\cdot)$ may not be continuous at the true parameter value if SC fails. The plug-in estimator is therefore not necessarily pointwise-consistent.

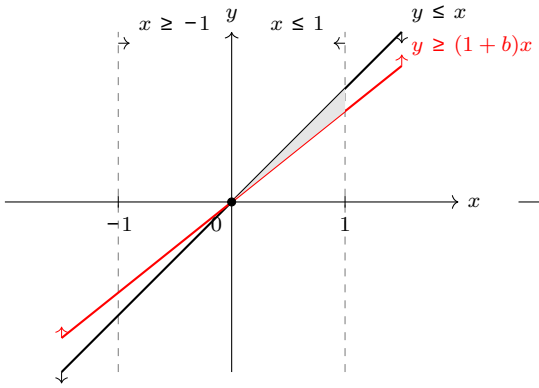
Proposition 1. *If SC fails for $\Theta_I(\theta_0)$, $B(\hat{\theta}_n)$ is not, in general, consistent for $B(\theta_0)$. Moreover, $B(\hat{\theta}_n)$ may fail to exist with non-vanishing probability asymptotically.*

Proof. We provide a simple example. Suppose:

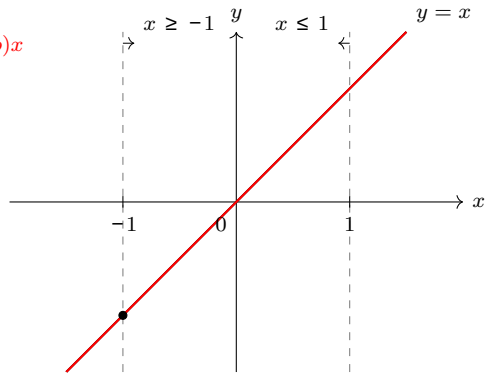
$$B(b) = \min_{x,y}\ x \quad \text{s.t.}: y \geq (1+b)x,\ y \leq x,\ x \quad [-1;1],$$

[11] The $\sqrt{n}$-consistency requirement straightforwardly generalizes to r_n-consistency for some $r_n \to \infty$. We focus on the former case for clarity.

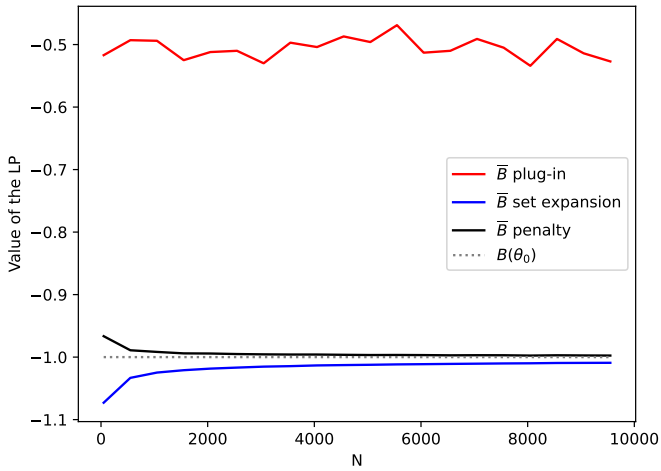
[12] Although our approach nests **equality** constraints, this definition of SC does not account for it. While it is done for simplicity, we note that in the presence of 'true' equalities $Ax = b$, the definition of SC should be stated in terms of *Relint*. In other words, point-identification along 'true' equalities is not problematic. The same disclaimer applies to LICQ.



(a) $b < 0,\ B(b) = 0$



(b) $b = 0,\ B(b) = -1$



(c) Simulations, $N_{sim} = 1000$, $k_n = \ln\ln\ln n$, $w_n = 1.1k_n$.

Figure 2: The first example in Proposition 1, minimizing x s.t. the constraints.

with b estimated via $b_n = \frac{1}{n}\sum_{i=1}^{n} U_i$ and $U_i \quad U[-1;1]$ i.i.d. Here, $\theta = b$ and $\hat{\theta}_n = b_n$. In the population, the optimal value is -1 attained at $x = -1$, $y = -1$. It is, however, straightforward to see that the plug-in estimator collapses to:

$$B(b_n) = -\mathbb{I}\{b_n \geq 0\} \not\to -1 \text{ in probability}$$

For the second part, consider:

$$B(a) = \min_{x,y} x \quad \text{s.t.} : y \geq x + a,\ y \leq x,\ x \quad [-1;1], \tag{5}$$

where a is estimated via $a_n = \frac{1}{n}\sum_{i=1}^{n} U_i$ with U_i as before. For the realisations with $a_n > 0$, the estimator does not exist. That happens with probability $1/2$.

Figure 2 illustrates the Proposition 1. In this section we attempt to address this finding by first developing two novel consistent estimators of the value function of a random

LP under A0. Section 2.1 proposes the penalty function estimator that is shown to be $\sqrt{n}w_n^{-1}$-consistent for any diverging $w_n = o(\sqrt{n})$. We then construct the *debiased* penalty estimator that achieves $\sqrt{n}$-consistency in general. The estimator based on set expansions à la Chernozhukov, Hong and Tamer (2007) is also developed, but it is shown to converge only at $\sqrt{n}\kappa_n$ for $\kappa_n \to 0^+$, $\kappa_n^{-1} = o(\sqrt{n})$.

Turning to inference in Section 2.2, we observe that even if SC holds, it is only under further assumptions of LICQ and solution uniqueness that the plug-in estimator combined with bootstrap, as used in practice, is consistent for the distribution. This is because empirical literature applied the approach of Imbens and Manski (2004) to the estimation of AICM models without justifying Assumption 1.i) in that paper. This assumption actually fails when LICQ fails or in the presence of flat faces, because $B(\theta)$ is only Hadamard directionally differentiable in θ. If $\hat{\theta}_n$ is asymptotically normal, Theorem 3.1 from Fang and Santos (2018) establishes that bootstrap is not consistent for the distribution. We then construct a robust exact inference procedure based on the debiased penalty function estimator using sample splitting. A closed-form expression for the asymptotic variance is provided.

In Section 2.3 we proceed to show that under a uniform version of Assumption A0 there exists no uniformly consistent estimator, because the value function viewed as a functional from the space of probability measures has a discontinuity at some measure. As in certain contexts the Slater's and LICQ conditions typically used to warrant the continuity of $B(\theta)$ may be too strong, we characterize a broader class of probability measures under which a uniformly consistent estimator exists. The proposed $\delta-$condition is *strictly* weaker than either a uniform Slater's condition or a uniform LICQ (e.g. in Gafarov (2024)). In Section 2.4 it is shown that even the biased penalty function approach is $\sqrt{n}w_n^{-1}$-uniformly consistent over the class of DGPs satisfying it for any diverging $w_n = o(\sqrt{n})$, whereas the plug-in still fails to be pointwise consistent. The debiased estimator is shown to enjoy a uniform rate of at least $\sqrt{n}w_n^{-1}$ under a mild regularity condition in Section 2.5. An investigation as to whether the pointwise $\sqrt{n}$-rate is achievable uniformly is provided in the Appendix. As to the selection of w_n, in Appendix we leverage the insights from random matrix theory (Tao and Vu, 2010) to suggest a reasonable candidate that balances finite-sample performance and uniform validity and performs well in our simulations.

2.1. Consistent estimators

2.1.a. Penalty function approach. We now derive a consistent penalty functions-based estimator. The idea is to restate (4) as an unconstrained penalized problem, so let us define:

$$\tilde{B}(\theta; w) \equiv \min_{x \in X} L(x; \theta, w), \quad \tilde{A}(\theta; w) \equiv \arg\min_{x \in X} L(x; \theta, w), \tag{6}$$
$$L(x; \theta, w) \equiv p\, x + w\, (c - Mx)^+.$$

$\tilde{B}(\cdot)$ is our preliminary estimator, which we term *the biased penalty function estimator*. Note that $L(x; \theta, w) = p\, x$ at any x, such that $Mx - c \geq 0$, i.e. the penalized function is equal to the objective function whenever the constraint in (4) holds.

We consider the following preliminary estimator:

$$\tilde{B}_n \equiv \min_{x \in X} \hat{L}_n(x), \quad \hat{L}_n(x) = \hat{p}_n x + w\,(\hat{c}_n - \hat{M}_n x)^+.$$

Assumption A1 (Penalty parameter). *The vector of penalties $w \in \mathbb{R}^q$ is such that at θ_0 in the initial L.P. $\lambda \in \mathbb{R}^Q_+$ - a vector of Lagrange multipliers at the optimum, such that:*

$$w_i > \max_j \{\lambda_j\}, \ i \in \overline{1,q}$$

Remark. Note that it is not necessary for w to be component-wise larger than *all* Lagrange multipliers at the optimum of the initial L.P. Further note that in cases where it is known that i) $B(\theta_0) < K$ for some $K > 0$ and ii) $c > \underline{c} > 0$ for some $\underline{c} > 0$, Assumption A1 is satisfied by $w = \iota K/\underline{c}$.

Condition A1 ensures that, at the true value of θ_0 the unconstrained minimum of $L(\cdot)$ coincides with the value of the original linear program:

Lemma 1. In general, the biased penalty function is conservative in the sense that:

$$\tilde{B}(\theta; w) \leq B(\theta) \quad w \in \mathbb{R}_+, \ \theta \in \mathbb{R}^S. \tag{7}$$

Moreover, for (θ_0, w) satisfying A1: i) (7) holds with an equality, and ii) optimal solutions coincide, $\tilde{A}(\theta_0; w) = A(\theta_0)$.

Proof. In the Appendix.

The deterministic result in Lemma 1, combined with the observation that the value function converges uniformly establish that the biased penalty function estimator is consistent under A1:

Proposition 2. *Under Assumption A1, $\tilde{B}_n$ is consistent, i.e.:*

$$\tilde{B}_n \xrightarrow{p} B(\theta_0) = \min_{Mx \geq c} p\,x$$

Proof. In the Appendix.

It might therefore seem that w should be selected to be as large as possible. However, this yields a generally inconsistent estimator if SC fails. Let us return to the first example in Proposition 1 for illustration. Suppose we set $w > 1$. In light of Lemma 1, the penalty estimator selects an incorrect optimum of 0 when $b_n < 0$ and $w > |b_n^{-1}|$, because the sample Lagrange multiplier is proportional to b_n^{-1} in this case. So, although for any fixed w at a large enough sample size $|1/b_n|$ will exceed it with high probability and the correct minimum of -1 will be selected, it is not the case in finite samples. This observation justifies the need for a careful consideration of $w \to \infty$ asymptotic theory.

We now observe that the penalty parameter can in fact be allowed to diverge at the rate dominated by $\sqrt{n}$. Consider a non-decreasing $w_n > 0$, and substitute it instead of

the fixed w, so that:

$$\tilde{B}_n(\hat{\theta}_n, w_n) \equiv \min_{x \ X} \hat{p}_n x + w_n \iota \ (\hat{c}_n - \hat{M}_n x)^+$$

We obtain the following result:

Theorem 1. For any $w_n \to \infty$ w.p. 1 as. with $\frac{w_n}{\sqrt{n}} \xrightarrow{p} 0$, we have:

$$|\tilde{B}_n(\hat{\theta}_n, w_n) - B(\theta_0)| = O_p\left(\frac{w_n}{\sqrt{n}}\right)$$

Proof. In the Appendix.

Remark. By $w_n \to \infty$ w.p. 1 as. we mean that $M > 0 \ \mathsf{P}[w_n > M] \to 1$.

Remark. Theorem 1 provides a generally pointwise-consistent estimator for $B(\theta_0)$. The only condition for consistency is our baseline Assumption A0.

The discussion of uniform asymptotic theory that follows sheds further light on the issue of selecting the level of w_n. Based on it, we develop a practical approach to its selection in the Appendix.

Careful investigation of the proofs above reveals that the rate at which the penalty function estimator converges is determined by the $O_p(1/\sqrt{n})$ penalty term multiplied by an exploding sequence w_n. Therefore, a reasonable question to ask is whether $\sqrt{n}-$rate could be restored by dropping that term. We show that this can be done. Before we proceed, let us make the following simplification. Without loss of generality, suppose:

$$\hat{p}_n = p \ - \text{non-random.} \tag{8}$$

To see why this is possible, note that one can in general put $p = e_1$ - the first basis vector, adding an auxiliary variable for the value of the problem in the first position of x. See Gafarov (2024) for details.

The following theorem, although its statement appears similar to that of Theorem 1, is perhaps the most mathematically challenging contribution of this paper. Its proof requires a careful examination of the finite-sample behavior of a penalty function estimator and leverages multiple eqiuvalences between the penalty-function estimator and auxiliary piecewise-linear and linear programs.

Theorem 2. Suppose $A(\theta_0) \quad Int(X)$. For any $w_n \to \infty$ w.p. 1 as. with $\frac{w_n}{\sqrt{n}} \xrightarrow{p} 0$:

$$\sup_{x \ A(\hat{\theta}_n, w_n)} |p \ x - B(\theta_0)| = O_p\left(\frac{1}{\sqrt{n}}\right)$$

Proof. In the Appendix.

Intuitively, $\sqrt{n}-$consistency follows because with high probability asymptotically the penalty function estimator manages to select 'a correct face' of the polytope in some sense.

We proceed to show that any 'correct face', which is not necessarily a vertex nor an actual face of the true polytope, converges at the rate of $\sqrt{n}$ in some metric.

Remark. The result in Theorem 2 is uniform over the argmin set, so in the context of lower bound estimation one may use $\sup_{\tilde{A}(\hat{\theta}_n;w_n)} p\ x$ to obtain the tightest bound.

The latter observation motivates us to define *the debiased penalty function estimator* as follows:

$$\hat{B}(\hat{\theta}_n; w_n) \equiv \max_{x \ \tilde{A}(\hat{\theta}_n;w_n)} p\ x.$$

The next subsection explores an alternative approach to constructing a consistent LP estimator. However, the estimator it produces has a conservative rate, and we therefore advocate using $\hat{B}(\cdot)$ for estimation instead.

2.1.b. Set expansions approach. Proposition 1 highlights that the plug-in estimator fails whenever the constraint set has an empty interior. For completeness of our argument, we develop a natural alternative to the penalty function estimator - the *set-expansion approach*. The idea here is to enlarge Θ_I by relaxing each inequality constraint with a sequence κ_n[13]. The resulting estimator has the flavor of the approach in Chernozhukov, Hong and Tamer (2007). Intuitively, it enforces SC at the cost of producing a potentially conservative estimate. We show that, in general, this estimator can indeed have a conservative rate, and thus we do not advocate its use in practice.

The approach in this section is first to prove that the appropriately extended identified set converges to the population identified set in Hausdorff distance, and then use uniform continuity of the criterion function as well as its resemblance to the support function to establish the convergence of the estimator itself.

Consider the following criterion function and its sample analogue:

$$Q(x) \equiv ||(Mx - c)^-||^2, \quad \hat{Q}_n(x) \equiv ||(\hat{M}_n x - \hat{c}_n)^-||^2$$

Denote the identified set as $\Theta_I \equiv \{x \quad X | Q(x) = 0\} = \{x \quad X | Mx - c \geq 0\}$.

Lemma 2. $||\hat{Q}_n(x) - Q(x)||_\infty \xrightarrow{p} 0$, where $||\cdot||_\infty$ is over Θ_I.

Proof. See Appendix.

Analogously to the proof of Lemma 3, one shows that because both $\hat{c}_n$ and $\hat{M}_n$ are $\sqrt{n}$-consistent from A0, we have:

$$\sup_X (Q - \hat{Q}_n)^+ = O_p(1/\sqrt{n}), \quad \sup_{\Theta_I} \hat{Q}_n = O_p(1/n).$$

The plug-in estimator of the identified set, $\{x \quad X | \hat{Q}_n = 0\} = \Theta_I(\hat{\theta}_n)$, may not 'cover' the true asymptotically, as discussed in Chernozhukov, Hong and Tamer (2007) (CHT).

[13] In the presense of 'true equality' constraints $Ax = b$, the corresponding inequalities need not be expanded.

To address that, consider the following class of set estimators:

$$\{x \quad X | n\hat{Q}_n(x) \leq \kappa_n\}$$

Fix κ_n such that $P[\kappa_n \geq \sup_{\Theta_I} n\hat{Q}_n] \to 1$ and $\frac{\kappa_n}{n} \xrightarrow{p} 0$. Let $\hat{\Theta}_n \equiv \{x | \hat{M}_n x - \hat{c}_n \geq -\frac{\sqrt{\kappa_n}}{\sqrt{n}}\iota\}$. It is the set that we want to prove consistent for the population identified set.

The issue is that the set $\hat{\Theta}_n$ is not a criterion-based set, so the results in CHT is not directly applicable. However, we can define $\underline{\Theta}_n \equiv \{x | \hat{Q}_n(x) \leq \frac{\kappa_n}{n}\} \quad \hat{\Theta}_n$ and $\overline{\Theta}_n \equiv \{x | \hat{Q}_n(x) \leq q\frac{\kappa_n}{n}\} \quad \hat{\Theta}_n$.

We then wish to 'sandwich' $\hat{\Theta}_n$ between a smaller set that asymptotically covers Θ_I and a bigger set that is asymptotically covered by Θ_I. The following simple lemma is an analogue of 'sandwich theorem' for sets.

Lemma 3. Consider $\Theta_I \quad X$ and suppose the random set $\hat{\Theta}_n \quad \Theta$ can be sandwiched between two sets: $\underline{\Theta}_n \quad \hat{\Theta}_n \quad \overline{\Theta}_n$, such that:

$$\sup_{x \quad \overline{\Theta}_n} d(x, \Theta_I) = o_p(1)$$
$$\sup_{x \quad \Theta_I} d(x, \underline{\Theta}_n) = o_p(1)$$

Then:

$$d_H(\hat{\Theta}_n, \Theta_I) = o_p(1)$$

Proof. Writing out the definitions and applying CMT yields the result.

The only thing that remains to show consistency of the set-estimator is to prove that the inequalities in Lemma 3 hold in our case. The derivation below follows the usual CHT logic. The first equality is established through:

$$P[\sup_{x \quad \overline{\Theta}_n} d(\theta, \Theta_I) \leq \varepsilon] = P[\overline{\Theta}_n \quad \Theta_I^{\varepsilon}] = \tag{9}$$
$$P[\overline{\Theta}_n \cap X \setminus \Theta_I^{\varepsilon} = \ ] \geq P[\sup_{x \quad \overline{\Theta}_n} Q(\theta) < \inf_{x \quad X \setminus \Theta_I^{\varepsilon}} Q(\theta)]$$

Then, by uniform continuity and by the construction of $\overline{\Theta}_n$:

$$\sup_{x \quad \overline{\Theta}_n} Q(\theta) = \sup_{x \quad \overline{\Theta}_n} \hat{Q}_n(\theta) + o_p(1) = q\frac{\kappa_n}{n} + o_p(1) = o_p(1)$$

By construction of Θ_I and continuity of $Q(\theta)$, $\delta > 0$: $\inf_{x \quad X \setminus \Theta_I^{\varepsilon}} Q(\theta) > \delta$. Thus, the RHS of (9) goes to 1. So, $\sup_{x \quad \overline{\Theta}_n} d(x, \Theta_I) = o_p(1)$.

The other side follows, as by construction $\sup_{x \quad \Theta_I} \hat{Q}_n(x) \leq \frac{\kappa_n}{n} = \Theta_I \quad \underline{\Theta}_n$. So,

$$P[\sup_{x \quad \Theta_I} d(\theta, \underline{\Theta}_n) \leq \varepsilon] \geq P[\sup_{x \quad \Theta_I} \hat{Q}_n(x) \leq \frac{\kappa_n}{n}] \xrightarrow{p} 1$$

Therefore, using Lemma 3, we conclude that:

$$d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$$

The next step is to recall that if we have two convex, compact sets, A, B, the following holds:

$$d_H(A, B) = \max_{||y|| \leq 1} |s(y, A) - s(y, B)|,$$

where $s(y, S) \equiv \max_{t \in S} y\, t$ - the support function.

Using uniform convergence of the value function and combining all the results:

$$|\min_{x \in \hat{\Theta}_n} \hat{p}_n x - \min_{x \in \Theta_I} p\, x| = |\min_{x \in \hat{\Theta}_n} p\, x - \min_{x \in \Theta_I} p\, x| + o_p(1) =$$
$$= |s(-p, \Theta_I) - s(-p, \hat{\Theta}_n)| + o_p(1) \leq ||p|| d_H(\Theta_I, \hat{\Theta}_n) + o_p(1) \xrightarrow{p} 0$$

This establishes the following proposition:

Proposition 3. *Let κ_n : $P[\kappa_n \geq \sup_{\Theta_I} n\hat{Q}_n] \to 1$ and $\frac{\kappa_n}{n} \xrightarrow{p} 0$. Then the following estimator is consistent for the sharp lower bound:*

$$\check{B}_n \equiv \min_{\hat{M}_n x - \hat{c}_n \geq -\sqrt{\frac{\kappa_n}{n}}\iota} \hat{p}_n x \xrightarrow{p} \min_{Mx - c \geq 0} p\, x$$

In practice, Chernozhukov, Hong and Tamer (2007) suggest to select some κ_n that diverges sufficiently slowly with the sample size. Simulations in Figure 2 use $\kappa = \ln \ln \ln n$. Under the Slater's condition the naive estimator is consistent, i.e. one could set $\kappa_n = 0$.

Although it seems intuitive that $\check{B}_n$ should converge at the rate $\sqrt{n\kappa_n^{-1}}$, deriving that result is outside the scope of this paper, because we do not advocate its use. It is immediate to see, however, that $\check{B}_n$ *can converge as slowly as* $\sqrt{n\kappa_n^{-1}}$. For that, consider the example in Proposition 1 without the inequality $y \leq x$ and setting $b_n = 0$. The minimum is attained at $-1 - \sqrt{\frac{\kappa_n}{n}}$.

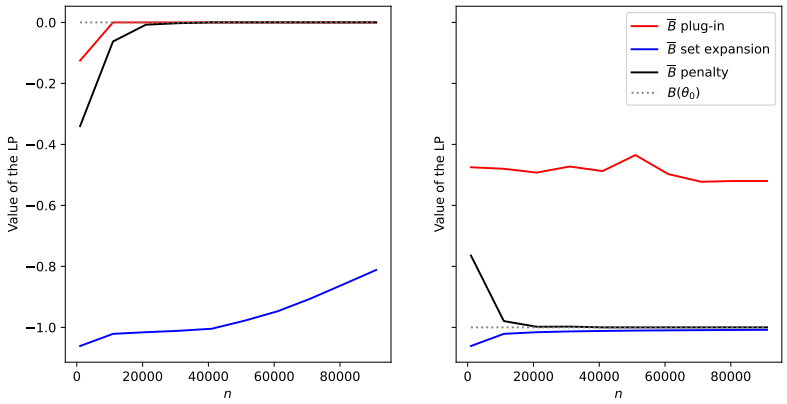


Figure 3: Averages for 400 simulations of the example in Proposition 5 with $b_i \quad U[-1; 1+2b]$. Penalty $w_n = \delta_{0.15}^{-1} \times \ln\ln n / \ln\ln 100$ and set expansion $\kappa_n = (\ln\ln n)^2$ **Left:** $b = -0.02$ ($\alpha = 0.12$), Slater's holds; **Right**: $b = 0$ ($\alpha = 0.75$), Slater's fails. See Appendix for the definition of δ_α and the details on w_n selection.

In practice, the set-expansion estimator can be quite conservative. Figure 3 illustrates that issue. Although its performance at the measure indexed by $b = -0.02$ may be improved by selecting a smaller κ_n parameter, this will harm the performance of the estimator at the measure with $b = 0$. The debiased penalty function estimator with w_n chosen according to the procedure developed in the Appendix appears to perform well across the two measures.

2.2. Inference

This section develops an inference procedure for a general LP estimator, in which all parameters are inferred from the data. This procedure nests special cases in which some parameters remain fixed, as in Semenova (2023) and Mogstad, Santos and Torgovitsky (2018).

2.2.a. Bootstrap fails even under SC. To justify the need for our approach, we first examine the properties of the plug-in estimator combined with bootstrap. Using bootstrap for inference on interval bounds was proposed in Imbens and Manski (2004) and widely used in the empirical literature applying AICM conditions, absent a necessary theoretical justification of the assumptions in the referenced paper. We show that even if SC holds, which may be justified in some applications where an explicit form for $B(\theta)$ exists, this approach fails unless one further rules out both solution multiplicity and LICQ. We begin by defining the Hadamard directional derivative:

Definition. Let D and E be Banach spaces, and $f : \mathsf{D}_f \quad \mathsf{D} \to \mathsf{E}$. The map f is said to be Hadamard directionally differentiable at $\upsilon \quad \mathsf{D}_f$ tangentially to $\mathsf{D}_0 \quad \mathsf{D}$, if there is a continuous map $f_\upsilon : \mathsf{D}_0 \to \mathsf{E}$, s.t.:

$$\lim_{n \to \infty} \left\| \frac{f(\upsilon + t_n h_n) - f(\upsilon)}{t_n} - f_\upsilon(h) \right\|_{\mathsf{E}} = 0,$$

for all sequences $\{h_n \quad D\}$ and $\{t_n\} \quad R_+$ such that $t_n \to 0^+$, $h_n \to h \quad D_0$ as $n \to \infty$ and $v + t_n h_n \quad D_f$ for all n. If, moreover, $f_v(h)$ is linear in h, the map f is said to be fully Hadamard differentiable.

It will be convenient to break the vector of parameters into components: $\theta_0 = (p, c_1, \ldots c_q, M_1 \ldots M_q) \quad R^S$.

Assumption B1 (Asymptotic normality). *There is an estimator $\hat{\theta}_n$, such that:*

$$r_n(\hat{\theta}_n - \theta_0) \xrightarrow{L} G_0 \quad N(0, \Sigma)$$

With $\Sigma < \infty$.

The following result is not novel, but is provided here for reference:

Lemma 4. Under SC, $B(\cdot)$ is Hadamard directionally differentiable at θ_0. The directional derivative is given by:

$$B_{\theta_0}(h) = \inf_{x \in A(\theta_0)\lambda} \sup_{\lambda \in \Lambda(\theta_0)} h_p x + \sum_{i=1}^{Q} \lambda_i (h_{c_i} - h_{M_i} x), \tag{10}$$

where $h = (h_p, h_{c_1}, \ldots, h_{c_q}, h_{M_1}, \ldots, h_{M_q})$ is the direction of the increment corresponding to θ_0.

Proof. Duan et al. (2020), Theorem 4.1 with second-order terms' coefficients set to 0.

Hadamard directional differentiability of $B(\theta)$ is sufficient to guarantee convergence in distribution:

Proposition 4. *Under SC and assumption B1, we have:*

$$r_n(B(\hat{\theta}_n) - B(\theta_0)) \xrightarrow{L} B_{\theta_0}(G_0)$$

Proof. Fang and Santos (2018) Theorem 2.1. combined with Lemma 5.

However, the form of (10) suggests that $B_{\theta_0}(G_0)$ is not normal unless there is full Hadamard differentiability, i.e. $B_{\theta_0}(h)$ is linear in h. This violates the necessary condition for bootstrap consistency given by Assumption 1.i) in Imbens and Manski (2004). Theorem 3.1 from Fang and Santos (2018) also establishes that bootstrap is, in fact, inconsistent when $B_{\theta_0}(h)$ fails to be linear. Given that $\hat{\theta}_n$ is asymptotically Gaussian, the simple bootstrap procedure, employed in Blundell et al. (2007), Kreider et al. (2012), Gundersen, Kreider and Pepper (2012), Siddique (2013), De Haan (2017), and Cygan-Rehm, Kuehnle and Oberfichtner (2017) is *not consistent* for the distribution absent further restrictive assumptions:

Proposition 5. *If Assumption B1 holds and the support of* G_0 *is a vector subspace of* D*, the simple bootstrap is consistent for the distribution of* $B(\hat{\theta}_n)$*, i.e.:*

$$\sup_{f \ BL_1(\mathsf{R})} |\mathsf{E}[f(r_n(B(\theta_n) - B(\hat{\theta}_n)))|\{X_i\}_{i=1}^n] - \mathsf{E}[f(B_{\theta_0}(\mathsf{G}_0))]| = o_p(1),$$

if and only if: i) SC holds, ii) there are no flat faces, iii) LICQ holds.

Proof. In the Appendix.

One way to obtain a consistent estimator for the distribution of a plug in under SC is to combine the Functional Delta Method of Fang and Santos (2018) with the Numerical Delta Method given in Hong and Li (2015). We further show that the biased penalty function estimator, $\tilde{B}(\theta; w)$, is always H.d.d. and thus one can theoretically perform inference on it using the same approach. This method, however, relies on an arbitrarily selected sequence ϵ_n. It is also not practical, because an appropriate fixed w is not known. We thus confine this discussion to the Appendix, where we also point out that conservative inference can be derived for the set-expansion estimator.

The approach we advocate instead is to rely on the asymptotically normal version of the debiased penalty estimator with $w_n \to \infty$, resulting in confidence regions with exact coverage and rate $\sqrt{n}$. The next section discusses this procedure.

2.2.b. Exact inference on a debiased estimator. We now show how to perform exact statistical inference on the basis of our debiased estimator $\hat{B}_n$. In this section, we suppose that we have a dataset $\mathsf{D}_n \equiv \{W_1, W_2, \ldots, W_n\}$, where W_i are i.i.d. and $\hat{\theta}_n = \hat{\theta}_n(\mathsf{D}_n)$.

For $x \quad \mathsf{R}^{qd}$ we define the inverse-vectorization operator:

$$\mathrm{vec}_{q\times d}^{-1}(x) = (\mathrm{vec}(I_d) \quad I_q)\,(I_d \quad x).$$

We further define selector matrices C_c and C_M that select the c and M components of θ respectively, i.e.:

$$C_c\theta = c, \quad C_M\theta = \mathrm{vec}(M).$$

Moreover, for an arbitrary subset $A \quad \{1, 2, \ldots, q\}$, define the selector matrix $C(A)$ that yields:

$$C(A)M = M_A, \quad C(A)c = c_A.$$

We randomly split D_n into two disjoint, collectively exhaustive folds $\mathsf{D}_n^{(f)}$ of size n_f for $f = 1, 2$, with $n_1 = \alpha n$ and $n_2 = n - \alpha n$ for some fixed $\alpha \quad (0; 1)$. For $f = 1, 2$, we also denote the estimator computed over $\mathsf{D}_n^{(f)}$ as $\hat{\theta}^{(f)} = (p\,, \mathrm{vec}(\hat{M}^{(f)})\,, \hat{c}^{(f)}\,)$.

Assumption B2. *B1 holds and there exists an estimator* $\hat{\Sigma}_n$*:*

$$\hat{\Sigma}_n \xrightarrow{p} \Sigma$$

Assumption B2 requires the researcher to possess a consistent estimator for the

asymptotic variance of $\hat{\theta}_n$. If $\hat{\theta}_n = g(\frac{1}{n}\sum_{i=1}^{n} W_i)$ for some appropriately smooth and known $g(\cdot)$, such estimator can be obtained using the traditional Delta-method. In more complicated scenarios, bootstrap on $\hat{\theta}_n$ may be employed. Further discussion of how to obtain $\hat{\Sigma}_n$ is of little theoretical interest and is thus omitted.

To develop our inference procedure, we first work on $D_n^{(1)}$. We begin by finding the set of optimal vertices of the LP formulation of the penalty estimator evaluated at $\hat{\theta}^{(1)}$, $V_x(\hat{\theta}^{(1)})$ (see the proof of Proposition 12)[14]. The value $p\,x$ can vary over x $V_x(\hat{\theta}^{(1)})$. To see that, note that $\tilde{L}(x;\hat{\theta}^{(1)},w_n)$ is constant over x $\tilde{A}(\hat{\theta}^{(1)};w_n)$ $V_x(\hat{\theta}^{(1)})$, but it is not necessarily true that $p\,x + C = \tilde{L}(x;\hat{\theta}^{(1)},w_n)$ for a fixed C R and any x $\tilde{A}(\hat{\theta}^{(1)};w_n)$ if some constraints of the original LP are violated. Still, the result in Proposition 12 applies uniformly over $V_x(\hat{\theta}^{(1)})$ and we can, for example, choose some $\hat{x}$ that satisfies[15]:

$$\hat{x} \quad \underset{x \;\; V_x(\hat{\theta}^{(1)})}{\arg\max} \; p\,x$$

Accordingly, we estimate the set of all constraints binding at that vertex:

$$\hat{A} \equiv \{j \quad \{1,2,\ldots,q\} | \hat{M}_j^{(1)}\hat{x} = \hat{c}_j^{(1)}\}$$

For an arbitrary subset of indices A $\{1,2,\ldots,q\}$, consider two conditions:

$$x \quad A(\theta_0) : M_A x = c_A \tag{11}$$

$$p \quad R(M_A) \tag{12}$$

Further, let:

$$\mathsf{A} \equiv \{A \quad 2^{\overline{1,q}} | A \text{ satisfies (11) and (12)}\}$$

Condition (12) equivalently postulates that there exists a solution to the equation:

$$p = M_A v$$

For a given A, we let $S_A \equiv \{v \quad \mathsf{R}^{|A|} : p = M_A v\}$ denote the set of all such solutions.

From the proof of Proposition 12 it follows that $\hat{A}$ A with probability 1 asymptotically. By definition, $\hat{A}$ A implies that v $S_{\hat{A}}$ such that $||v|| \leq \max\limits_{A \;\; \mathsf{A}} \min\limits_{v \;\; S_A} ||v||$, or:

$$\inf_{v \;\; \mathsf{R}^{|\hat{A}|}: ||v|| \leq \tilde{v}} ||p - M_{\hat{A}} v||^2 = 0$$

for any $\tilde{v} \geq \max\limits_{A \;\; \mathsf{A}} \min\limits_{v \;\; S_A} ||v||$. Consider some globally fixed, large enough $\overline{v}$.

Assumption B3. $\overline{v} \geq \max\limits_{A \;\; \mathsf{A}} \min\limits_{v \;\; S_A} ||v||$

[14] Although we assume this set is estimated precisely, the results do not change if one is only able to estimate a subset of $V_x(\hat{\theta}^{(1)})$. This may occur if numerical errors do not allow the LP-solver to find all of the LP solutions.

[15] In practice, it is very likely that the solution of the sample LP will be unique and one will have $V_x(\hat{\theta}^{(1)}) = \tilde{A}(\hat{\theta}^{(1)};w_n)$.

We can estimate an element of $S_{\hat{A}}$ as:

$$\check{v} \quad \underset{v \;\; \mathsf{R}^{|\hat{A}|}:||v|| \leq \overline{v}}{\arg\min} \; ||p - \hat{M}^{(1)}{}_{\hat{A}} v||^2$$

To avoid dealing with changing dimension, we let $\hat{v}$ R^q be such that $C(\hat{A})\hat{v} = \check{v}$ and $\hat{v}$ is 0 otherwise. By the usual M-estimation argument it then follows that there exists a random sequence $\{\tilde{v}_k\}_{k=1}^{\infty}$ $S_{\hat{A}}$, which is obtained as a measurable function of $\check{v}, \hat{A}$, and:

$$||\tilde{v}_n - \check{v}|| = O_p\left(\frac{1}{\sqrt{n}}\right)$$

Once again, to avoid dealing with the changing dimension of $\tilde{v}_n$, we construct v_n such that $v_n = C(\hat{A})\tilde{v}_n$ and 0 otherwise. We thus also have:

$$||\hat{v} - v_n|| = O_p\left(\frac{1}{\sqrt{n}}\right)$$

Before we proceed to the second fold, let us provide the following simple lemma that justifies our construction.

Lemma 5. Suppose $\hat{A}$ A. Then:

$$\tilde{v}_n c_{\hat{A}} = B(\theta_0) \tag{13}$$
$$\tilde{v}_n M_{\hat{A}} \hat{x} = p\,\hat{x} \tag{14}$$

Proof. If $\hat{A}$ A, condition (11) holds for some x $\mathsf{A}(\theta_0)$ such that $M_{\hat{A}} x = c_{\hat{A}}$. Since such x is a minimizer, it follows that $p\,x = B(\theta_0)$. As $\tilde{v}_n$ $S_{\hat{A}}$, we have $p = M_{\hat{A}}\tilde{v}_n$. Taking transpose and multiplying by $\hat{x}$ yields (14). To show (13), write:

$$p\,x = \tilde{v}_n M_{\hat{A}} x = \tilde{v}_n c_{\hat{A}} \tag{15}$$

Equipped with $\hat{v}$, $\hat{A}$ and $\hat{x}$, we can now move onto the second fold. Consider the following expressions:

$$H_n \equiv \frac{\sqrt{n_2}}{\hat{\sigma}(\hat{A},\hat{v},\hat{x})}\check{v}\left(\hat{c}^{(2)}_{\hat{A}} - c_{\hat{A}} - (\hat{M}^{(2)}_{\hat{A}} - M_{\hat{A}})\hat{x}\right)$$
$$G_n \equiv \frac{\sqrt{n_2}}{\hat{\sigma}(\hat{A},\hat{v},\hat{x})}(\check{v} - \tilde{v}_n)\,(c_{\hat{A}} - M_{\hat{A}}\hat{x})$$

Applying Lemma 5 yields that:

$$\frac{\sqrt{n_2}}{\hat{\sigma}(\hat{A},\hat{v},\hat{x})}\left(\check{v}\,(\hat{c}^{(2)}_{\hat{A}} - \hat{M}^{(2)}_{\hat{A}}\hat{x}) + p\,\hat{x} - B(\theta_0)\right) = H_n - G_n$$

The rest of the construction consists in showing, heuristically speaking, that $G_n = o_p(\frac{1}{\sqrt{n}})$ and thus the confidence intervals constructed based on H_n are also valid for $H_n - G_n$

asymptotically.

Letting $Z_n^{(2)} \equiv \sqrt{n_2}(\hat{\theta}^{(2)} - \theta_0)$, we can then write:

$$\sqrt{n_2}\check{v}\left(\hat{c}_{\hat{A}}^{(2)} - c_{\hat{A}} - (\hat{M}_{\hat{A}}^{(2)} - M_{\hat{A}})\hat{x}\right) = \check{v}\, C(\hat{A})\left(C_c Z_n^{(2)} - \mathrm{vec}_{q\times d}^{-1}(C_M Z_n^{(2)})\hat{x}\right) \tag{16}$$

It is straightforward to observe that, due to bilinearity of the Kronecker product, (16) is linear in $Z_n^{(2)}$ and therefore, for fixed $\hat{A}, \hat{x}, \hat{v}$, converges to:

$$\sqrt{n_2}\check{v}\left(\hat{c}_{\hat{A}}^{(2)} - c_{\hat{A}} - (\hat{M}_{\hat{A}}^{(2)} - M_{\hat{A}})x\right) \overset{d}{\Rightarrow} \mathsf{N}\left(0, \sigma(\hat{A}, \hat{x}, \hat{v}, \Sigma)\right)$$

Lemma 6. At fixed A, x, v:

$$\sigma(A, x, v, \Sigma) = J_1 \Sigma J_1 - 2J_2(I_d \quad C_M \Sigma J_1)x + J_2\,(xx \quad C_M \Sigma C_M)\,J_2 \tag{17}$$

Proof. In the Appendix, along with the definitions of J_1, J_2.

Assumption B4 (Non-degeneracy). *Suppose $\sigma(A, x, v, \Sigma) > 0$ for any optimal triplet A, x, v.*

By assumption B4 we then have, for fixed $\hat{A}, \hat{x}, \hat{v}$:

$$\frac{\sqrt{n_2}}{\sigma(\hat{A}, \hat{v}, \hat{x}, \Sigma)}\check{v}\left(\hat{c}_{\hat{A}}^{(2)} - c_{\hat{A}} - (\hat{M}_{\hat{A}}^{(2)} - M_{\hat{A}})\hat{x}\right) \overset{d}{\Rightarrow} \mathsf{N}(0, 1) \tag{18}$$

Theorem 3. Suppose $w_n \to \infty$ and $w_n = o_p(\sqrt{n})$ and Assumptions B1, B3, B4 hold. Moreover,

$$\hat{\sigma}_n(A, v, x) \overset{p}{\Rightarrow} \sigma(A, v, x, \Sigma)$$

for any fixed A, v, x, which holds for $\hat{\sigma}_n(A, v, x) = \sigma(A, v, x, \hat{\Sigma}_n)$ under Assumption B2. Then, for any $\alpha > 0$:

$$\mathsf{P}\left[\frac{\sqrt{n_2}}{\hat{\sigma}_n(\hat{A}, \hat{v}, \hat{x})}\left(\check{v}\,(\hat{c}_{\hat{A}}^{(2)} - \hat{M}_{\hat{A}}^{(2)}\hat{x}) + p\,\hat{x} - B(\theta_0)\right) \leq z_{1-\alpha}\right] = 1 - \alpha + o(1)$$

Proof. In the Appendix.

2.3. Uniform asymptotic theory

Random LP problems are challenging to study under no further assumptions, since they feature instability with respect to arbitrary parameters' perturbations, as shown in Proposition 5. We now note that this not only leads the plug-in to fail pointwise, but also precludes the existence of uniformly consistent estimators in general.

The following auxiliary result highlights that there exists no uniformly consistent estimator for any functional from a space of probability measures equipped with the total variation norm that is discontinuous at some measure.

Lemma 7. Suppose a functional $V : (P, ||\cdot||_{TV}) \to (\mathbb{R}, |\cdot|)$ is discontinuous at $P_0 \in P$. Then, there exists no uniformly consistent estimator $\hat{V}_n = \hat{V}_n(X)$, which is a sequence of measurable functions of the data $X \sim P^n$. Moreover, if $\delta > 0$ is the jump at P_0, then for all n:

$$\inf_{\hat{V}_n} \sup_{P \in P} \mathbb{E}_P[||V(P) - \hat{V}_n(X(P^n))||] \geq \frac{\delta}{4},$$

where infinum is taken over all measurable functions of the data.

Proof. In the Appendix.

We shall understand the parameter θ_0 as a functional of the underlying probability measure $P \in P$. We then make the following assumption on the pair $\theta_0(\cdot), P$:

Assumption U0 (Uniform setup). *The functional $\theta_0(\cdot)$ and the set of underlying probability measures P are such that:*

i) $\theta_0 : (P, ||\cdot||_{TV}) \to (\mathbb{R}^S, ||\cdot||_2)$ *is a continuous functional*

ii) $\theta_0(P) = \{y \in \mathbb{R}^S$ *s.t.* $\Theta_I(y) \leq , \Theta_I(y) \in X\}$ *for a known and fixed compact* X

Combining U0, Lemma 3 and the results in the previous sections, we establish that no uniformly consistent estimator of $B(\theta_0(P))$ exists absent further restrictions on P.

Theorem 4. Under U0 there exists no uniformly consistent estimator $\hat{B}_n$ of $B(\theta_0)$.

Given this negative result, we may instead ask over which set of measures the penalty function estimator yields uniform consistency. Intuitively, the penalty function estimator is problematic whenever the vertices of the population polygon at the optimum are too sharp, meaning that the associated Lagrange multipliers are too large. The condition that ensures uniform consistency of our estimator thus controls that sharpness. Consider:

$$(P) : B(\theta) = \min_x \ p\,x \quad \text{s.t.} : Mx \geq c$$

Form Lagrangean:

$$L \equiv p\,x + \lambda\,(c - Mx)$$

Because X is a compact, whenever the problem has a solution, it must be that there is also a solution $\lambda \geq , x \geq$ at which $J \subseteq \{1, 2, \ldots, q\}$ with $|J| = k$:

$$M_J x \geq c_J,$$

where $M_J \in \mathbb{R}^{k \times d}$ is a matrix of full column rank, i.e. $rk(M^J) = d$. Define the set of inactive constraints $I \equiv \{1, 2, \ldots, q\} \setminus J$ where:

$$M_I x > c_I$$

It follows that $\lambda_I = 0$.

Theorem 5. In the problem (P) there exists a solution x and the associated vector of KKT multipliers λ such that for some index subset J $\{1, \ldots, q\}$ with $|J| = d$, M_J is invertible and:

$$\begin{aligned} x &= M_J^{-1} c_J \\ \lambda_J &= M_J^{-1} p \\ \lambda_i &= 0, \ i \ / \ J \end{aligned}$$

Proof. Consider the solution x of the original problem at the vertix where the binding constraints are defined by the set J with $rk(M_J) = d$. Notice that the KKT condition that:

$$p = M_J \lambda_J$$

for some $\lambda_J \geq 0$ means that p $Cone(M_J)$. By the conical hull version of Caratheodory's Theorem, it follows that J J such that $|J| = r \leq d$ and p $Cone(M_J)$ and, moreover, the vectors M_J are linearly independent. If the Caratheodory number r is strictly smaller than the dimension of x, i.e. $r < d$, then we shall complement J with $d - r$ vectors from M_J such that we obtain $rk(M_J) = d$, setting the appropriate λ_i to 0. By necessity and sufficiency of KKT for LP problems, this constitues a solution.

Assumption U1 (δ-condition). *The set P of probability measures is such that P P we have $\theta(P) = \theta$ such that:*

$$\max_J \sigma_d(M_J(\theta)) > \delta, \tag{19}$$

where sets J with $|J| = d$ are those defined in Theorem 4. Moreover, the objective function vector is bounded on P, i.e. for some $\overline{p}$ $\mathbb{R}_+$:

$$\sup_{P \ P} ||p(P)|| < \overline{p}$$

Remark. The boundedness condition on p may always be imposed through redefining the variables in the objective linear program.

The δ-condition is weaker than the conditions usually imposed to establish uniform consistency of LP estimators. To formalise this notion, let us introduce three families of measures. Firstly, we shall denote the family of measures satisfying U1 by P^{δ}. We say that the measure satisfies a uniform ε-Slater's condition if P $P^{Slater;\varepsilon}$ where:

$$P^{Slater;\varepsilon} \equiv \{P \quad P | Volume(\Theta_I(\theta(P))) > \varepsilon\}$$

Similarly, a measure satisfies a uniform ε-LICQ condition as in Gafarov (2024) if P $P^{LICQ;\varepsilon}$, where:

$$P^{LICQ;\varepsilon} \equiv \{P \quad P \,|\, M(v) \quad \mathbb{R}^{d \times d}, \sigma_d(M(v)) > \varepsilon \quad v \quad V(P)\},$$

where the set $V(\mathsf{P})$ consists of all vertices of the polygon Θ_I evaluated at $\theta(\mathsf{P})$, while the operator $M(v)$ for a vertex v constructs the matrix of all binding constraints.

Proposition 6. *The following hold:*

1. $\lim_{n\to\infty} P^{Slater;1/n} \quad P^{LICQ;1/n} \quad P = \lim_{n\to\infty} P^{1/n}$, *where the inclusion is strict*
2. $P^{LICQ;\varepsilon} \quad P^{\delta}$ *for any* $\delta \leq \varepsilon$, *where the inclusion is strict*
3. *If, in addition to U0,* $\mathsf{P} \quad P$ *we have* $\sigma_1(M_J(\mathsf{P})) < \overline{\sigma}$ *for some* $\overline{\sigma} > 0$, *which is the case if the matrix* $M_0(\mathsf{P})$ *is normalized by row, then:* $\varepsilon > 0, \quad \delta$ *such that* $P^{Slater;\varepsilon} \quad P^{\delta}$ *and the inclusion is strict*

Intuitively, the $\delta > 0$ in assumption U1 merely parametrizes the degree of irregularity that the researcher is willing to allow for the constraint set over the considered set of measures. The resulting family of measure sets 'covers' the whole set of measures asymptotically as δ is allowed to decrease to 0. Other typically used conditions, however, restrict the set of measures even asymptotically. Moreover, for any there always exists a set of measures from the δ-condition family that strictly contains both of the above formulated conditions at fixed ε. In this sense, assumption U1 appears minimal for achieving uniform consistency.

2.4. Uniform consistency of penalty function estimator

Demanding uniform consistency for an arbitrarily small value of δ is not feasible, since this requires an arbitrarily large value of w_n, which will lead to poor finite-sample performance at some of the considered measures. We investigate this further in the Appendix.

Theorem 6. Suppose i) the set of measures P satisfies the $\delta-$ condition for some $\delta > 0$, and ii) $\hat{\theta}_n$ converges to $\theta(\mathsf{P})$ a.s. uniformly over P, i.e. for any $\varepsilon > 0$:

$$\lim_{n\to\infty} \sup_{\mathsf{P} \quad P} \mathsf{P}[\sup_{m\geq n} ||\hat{\theta}_m - \theta(\mathsf{P})|| \geq \varepsilon] = 0 \tag{20}$$

Then, the penalty function estimator $\tilde{B}(\hat{\theta}_n, w_n)$ with $w_n = ||\hat{p}_n||\delta^{-1} + \zeta$ for any globally fixed $\zeta > 0$ is uniformly consistent in the sense of a.s. convergence, i.e., for any $\varepsilon > 0$:

$$\lim_{n\to\infty} \sup_{\mathsf{P} \quad P} \mathsf{P}[\sup_{m\geq n} |\tilde{B}(\hat{\theta}_m, w_m) - B(\theta(P))| \geq \varepsilon] = 0 \tag{21}$$

Moreover if convergence in (20) is at rate r_n, (21) holds for any $w_n \to \infty$ with $\frac{w_n}{r_n} \to 0$.

Proof. In the Appendix.

In our context, θ_0 is linear in population moments of interactions of $Y(t)$ with treatment indicators and linear or hyperbolic in joint probabilities of $T = t, Z = z$ (moments of indicators). Thus, condition ii) in Theorem 6 is established by, firstly, imposing that:

$$\lim_{C\to\infty} \sup_{\mathsf{P} \quad P} \mathsf{E}_{\mathsf{P}}||\mathsf{Y}||\mathsf{I}\{||\mathsf{Y}|| > C\} = 0,$$

which, for example, holds whenever bounded outcomes are assumed. We shall also strengthen the full-support assumption to:

$$\mathsf{P}[T = t, Z = z] > C$$

for some $C > 0$, for all t, z $T \times Z$ and P P. Under these conditions, $\theta(\cdot)$ is a uniformly continuous function of population moments. Combining this and the LLN uniform in probability measure (see Proposition A.5.1 on p. 456 of Van Der Vaart et al. (1996)) yields condition (ii). Moreover, if one additionally assumes that:

$$\lim_{C \to \infty} \sup_{\mathsf{P}} \mathsf{E}_{\mathsf{P}} ||\mathsf{Y}||^2 \mathsf{I}\{||\mathsf{Y}|| > C\} = 0,$$

the rate in Theorem 6 is $r_n = \sqrt{n}$ (see Proposition A.5.2 on p. 457 of Van Der Vaart et al. (1996)).

Remark. Note that the existence of a uniformly consistent estimator over P implies that $B(\cdot)$ is continuous over $\theta_0(P)$. However, as we have seen in Proposition 5, it is not necessarilly continuous over the whole support of $\hat{\theta}_n$ and, in fact, $\theta_0(P) = Supp(\hat{\theta}_n)$ in general (where, of course, the support may vary over P). It is straightforward to see that the example in Proposition 5 satisfies the δ-condition for a relatively large δ, but the plug-in estimator is still pointwise inconsistent at such measure.

In light of the findings in this section, we develop an approach to selecting the penalty parameter w_n. It leverages random matrix theory and is given in Appendix.

2.5. Uniform consistency of the debiased penalty function estimator

2.5.a. Geometry of polytope projections. We need to introduce the following two objects, which we call the condition numbers.

Definition (The face condition number)**.** For a k-face of a polytope, f, which is described by binding constraints A $\overline{1, q}$ with $|A| \geq d - k$ such that $\mathrm{rk}(M_A) = d - k$, define the face condition number to be:

$$\tilde{\kappa}(f) \equiv \min_{B \quad A:\ \mathrm{rk}(M_B) = d-k} \sigma_{d-k}(M_B) \tag{22}$$

Definition (Polytope condition number)**.** For a polytope Θ, define the polytope condition number as:

$$\kappa(\Theta) = \min_{f - \text{face of } \Theta} \tilde{\kappa}(f) = \min_{f - \text{vertex of } \Theta} \tilde{\kappa}(f) \tag{23}$$

Remark. Any full-rank matrix at a k-face f for $k > 0$ can be obtained by removing k vectors from a full-rank matrix at some vertex ($0-$face) f , to which f belongs, so the condition number of f is greater or equal than that of f by the submatrix inequality for singular values.

Assumption U2 (Polytope δ-condition). *The class of measures P satisfies the polytope*

$\delta-$*condition, if, for some* $\delta > 0$*:*

$$\inf_{\mathsf{P}\ P} \kappa(\Theta_I(\mathsf{P})) \geq \delta \tag{24}$$

Remark. The polytope $\delta-$condition has the same flavor as the δ-condition, but essentially imposes it for all vertices of the polytope and all full-rank matrices at these vertices. It similarly just parametrizes the whole unconstrained set of probability measures, since at any fixed P we always have $\kappa(\Theta_I(\mathsf{P})) > 0$ by definition.

The following theorem appears to be mathematically novel.

Theorem 7. For any non-empty and bounded polytope $\Theta = \{x \quad \mathsf{R}^d | Mx \geq c\}$:

$$\iota \ (c - Mx)^+ \geq \frac{d(x,\Theta)\kappa(\Theta)}{d} \tag{25}$$

Proof. If $x \quad \Theta$, the inequality holds trivially. Consider x such that $d(x,\Theta) = \varepsilon > 0$. We construct a projection of x onto the polytope. It must be a solution of the following program:

$$\min_{y \ \Theta} \frac{1}{2}(y - x) \ (y - x) \tag{26}$$

Construct Lagrangean:

$$\mathsf{L} = (y - x) \ (y - x) + \lambda \ (c - My) \tag{27}$$

FOCs:

$$y - x - M \ \lambda = 0 \tag{28}$$

$$\lambda_j(M_j y - c_j) = 0 \tag{29}$$

$$My \geq c \tag{30}$$

This problem is convex and thus has a global minimum characterized by the KKT conditions. Let that minimum be y . Denote the subset of binding equalities:

$$J \equiv \{j \quad \overline{1,q} | M_j y \ = c_j\} \tag{31}$$

Suppose y belongs to at least k -face f , meaning that face f is given by:

$$f \ = \bigcap_{f-\text{face of } \Theta_I:\ y \ \ f} f, \tag{32}$$

with the associated set of binding equalities J such that $|J| \geq d - k$ and $\text{rk}(M_J) = d - k$. By construction:

$$y - x \quad \text{Cone}(M_J), \tag{33}$$

Therefore, by Caratheodory's Conical Hull theorem, there exists a subset $J \quad J$ such

that $|J| = r \leq d - k$ and a corresponding $\lambda_J > 0$:

$$y - x = M_J \lambda_J \tag{34}$$

Forming λ as $(\lambda)_J \equiv \lambda_J$ and setting $\lambda_j = 0$ for $j \notin J$, one can observe that y, λ solve the above of KKT conditions. Moreover, if $r < d - k$, we can complement J with $d - k - r$ linearly independent constraints from $J \setminus J$ to obtain $J \supset J$, such that: $|J| = \mathrm{rk}(M_J) = d - k$. Finally, setting $\lambda \equiv \lambda_J$, we get:

$$y - x = M_J \lambda \tag{35}$$

From where it follows that:

$$\lambda = (M_J M_J)^{-1} M_J (y - x) \tag{36}$$

Recall that $\|y - x\| = \varepsilon > 0$, and note that, because $(M_J M_J)^{-1} M_J$ is the left inverse of M_J:

$$\|(M_J M_J)^{-1} M_J\| \leq \sigma_{d-k}^{-1}(M_J) \leq \kappa^{-1}(\Theta_I) \tag{37}$$

By Cauchy-Schwarz, we then obtain:

$$\|\lambda\| \leq \varepsilon \kappa^{-1}(\Theta_I) \tag{38}$$

Since $\|\lambda\|_\infty \leq \|\lambda\|$, it also follows that:

$$\|\lambda\|_\infty \leq \varepsilon \kappa^{-1}(\Theta_I) \tag{39}$$

Further, since $M_J y = c_J$ by construction, multiplying both sides of (35) by M_J yields:

$$c_J - M_J x = M_J M_J \lambda \tag{40}$$

And plugging (35) into the value function, one gets:

$$\varepsilon^2 = \lambda_J M_J M_J \lambda = \lambda_J (c_J - M_J x) \tag{41}$$

Combining (41), the fact that at least one of the components of λ_J is positive from $y - x = 0$ and (35), as well as $0 \leq \lambda_J \leq \iota \varepsilon \kappa^{-1}(\Theta_I) \lambda_J$ from the definition and the bound on $\|\lambda\|_\infty$, one observes that:

$$j \notin J : (c - M_J x)_j \geq \frac{\kappa(\Theta_I)\varepsilon}{d - k} \tag{42}$$

It then follows that:

$$\iota (c - Mx)^+ \geq \frac{\kappa(\Theta_I)\varepsilon}{d - k} \tag{43}$$

Taking the minimum over k yields the claim of the proposition.

2.5.b. The debiased estimator is at least $\frac{\sqrt{n}}{w_n}$ uniformly consistent.. Recall that the original penalty is uniformly consistent at rate $\frac{\sqrt{n}}{w_n}$ under the δ-condition.

Theorem 8. If P satisfies Assumption U2 for some $\delta > 0$, the debiased estimator converges uniformly at the rate of at least $\frac{\sqrt{n}}{w_n}$.

Proof. Note that:

$$p\, x_n - B(\theta_0) \geq \min_{x \ \Theta_I^{d(x_n,\Theta_I)}} p\, x - \min_{x \ \Theta_I} p\, x = \tag{44}$$

$$\frac{1}{||p||}\left(s\left(\frac{p}{||p||}, \Theta_I^{d(x_n,\Theta_I)}\right) - s\left(\frac{p}{||p||}, \Theta_I\right)\right) \geq \tag{45}$$

$$-\frac{1}{||p||}\max_{||y||\leq 1}\left|s\left(y, \Theta_I^{d(x_n,\Theta_I)}\right) - s\left(y, \Theta_I\right)\right| = \tag{46}$$

$$-\frac{d_H\left(\Theta_I^{d(x_n,\Theta_I)}, \Theta_I\right)}{||p||} \geq -\frac{d\iota\ (c - Mx_n)^+}{\kappa(\Theta_I)||p||} \tag{47}$$

Thus:

$$O_p(1) = \frac{\sqrt{n}}{w_n}\left(p\, x_n - B(\theta_0) + w_n\iota\ (\hat{c}_n - \hat{M}_n x_n)\right) = \tag{48}$$

$$\frac{\sqrt{n}}{w_n}\left(p\, x_n - B(\theta_0) + w_n\iota\ (c - Mx_n)\right) + O_p(1) \geq \tag{49}$$

$$\sqrt{n}(1 - \frac{1}{w_n}\frac{d}{\kappa(\Theta_I)||p||})\iota\ (c - Mx_n)^+ + O_p(1) \tag{50}$$

From where it follows that:

$$\iota\ (c - Mx_n)^+ = O_p\left(\frac{1}{\sqrt{n}}\right) \tag{51}$$

Using:

$$\frac{\sqrt{n}}{w_n}\left(p\, x_n - B(\theta_0) + w_n\iota\ (\hat{c}_n - \hat{M}_n x_n)^+\right) \geq \frac{\sqrt{n}}{w_n}(p\, x_n - B(\theta_0)) \geq \tag{52}$$

$$\frac{-1}{w_n}\frac{d\sqrt{n}\iota\ (c - Mx_n)^+}{\kappa(\Theta_I)||p||} \tag{53}$$

One deduces that $p\, x_n - B(\theta_0)$ is $O_p(\frac{w_n}{\sqrt{n}})$. All arguments above are uniform if the convergence of $\hat{\theta}_n$ is uniform and as $\kappa(\Theta_I) \geq \delta$ for some $\delta > 0$.

3. Identification results

Let $Y \subseteq \mathsf{R}$ denote the outcome of interest[16], $T \subseteq \mathsf{R}$ stand for the treatment, and $Z \subseteq \mathsf{R}^{d_Z}$ be the candidate instrument. We denote the supports of these variables as $\mathcal{Y}$, $\mathcal{T}$

[16] Univariate case is considered for simplicity of exposition, but the extension to multivariate outcomes is immediate.

and Z respectively. The reader seeking an economic intuition may interpret this in line with the classical example from Manski and Pepper (2000), treating Y as the wage, T as an indicator of educational degree and Z as the level of ability.

In the considered context, treatment is defined to be any variable which effect on Y we attempt to infer, whereas the term *instrument* merely refers to an auxiliary variable that allows us to (partially) identify the treatment effect of interest. This latter distinction is important: although our approach nests the usual IV condition on Z, it is more general and allows the researcher to impose arbitrary linear inequality restrictions on conditional moments of potential outcomes.

Throughout this paper we consider the case of continuous outcomes and discrete treatment and instrument, namely Y is uncountable, while $N_T \equiv |T| < \infty$ and $N_Z \equiv |Z| < \infty$. In non-parametric bounds literature it is rather conventional to employ a discrete instrument at the estimation stage (see Manski and Pepper (2009)). While our main identification result could be extended to continuous Z, we make the discreteness assumption early on to avoid unnecessary technical complications.

Our setup accommodates missing observations of the dependent variable Y. Namely, we split the set of treatments into two disjoint subsets $T = O \cup U$. Whenever $T \in O$, the researcher observes Y, T, Z, whereas if $T \in U$, only the covariates T, Z are observed. For example, in Blundell et al. (2007) the wage is observed only if an individual is employed. Corresponding to the legs of the treatment are the potential outcomes $Y(t),\ t \in T$:

$$Y = \sum_{t \in O} \mathbb{I}\{T = t\} Y(t) + \sum_{t \in U} \mathbb{I}\{T = t\} Y(t)$$

Continuing the wages and education example, the value of $Y(t)$ for a fixed $t \in T$ may then correspond to the potential wage that an individual with the associated random characteristics would get, had she obtained education t.

Let us collect the potential outcomes in the vector $\mathbf{Y} \equiv (Y(t))_{t \in T} \in \mathbb{R}^{N_T}$. Variables $(\mathbf{Y}, T, Z, Y)$ are jointly defined on the true probability space $(\mathbb{P}, \Omega, \mathcal{S})$ and we let $\mathcal{P}$ denote the considered collection of probability measures on $(\Omega, \mathcal{S})$, such that $\mathbb{P} \in \mathcal{P}$. We impose the following conditions on the space of considered measures throughout the paper:

Assumption (Conditions on $\mathcal{P}$). *Set of probability measures $\mathcal{P}$ is such that $P \in \mathcal{P}$ if:*

i) Identification: P generates $F_{T,Z}(\cdot)$ and $\{F_{Y|T=t,Z}(\cdot)\}_{t \in O}$

ii) Full support of T, Z: $P[T = d, Z = z] > 0 \quad \forall d, z \in T \times Z$

iii) Finite conditional moments: $|\mathbb{E}_P[Y(t)|T = d, Z = z]| < \infty$ for all $z \in Z$ and $t, d \in T$

Part i) of the Conditions formalizes the assumed identification pattern. It says that the joint distribution of T, Z is always identified and the researcher also observes the joint distribution of Y, T, Z whenever $T \in O$. Parts ii) and iii) of the Conditions are the technical assumptions that ensure that all conditional expectations and probabilities are well-defined and finite[17]. In particular, this assumption implies that all moments of form $\mathbb{E}[Y(t)|T \in A, Z \in B]$ for some $A \in 2^T \setminus \{\emptyset\}$, $B \in 2^Z \setminus \{\emptyset\}$ and $t \in T$ are well-defined and finite.

[17] Similar identification results can still be obtained if one relaxes the full-support condition for some known pairs from $Z \times T$. Note that it can also be verified in the data.

Remark. Under no missing data, i.e. $T = O$, condition i) is equivalent to P generating the identified joint distribution $F_{Y,T,Z}(\cdot)$.

Let m $\mathsf{R}^{N_T^2 N_Z}$ be a vector collecting all elementary conditional moments of potential outcomes. For P P:

$$m(P) \equiv (\mathsf{E}_P[Y|T = d, Z = z])_{d\ T,z\ Z}$$

In general, $m = m(P)$ is a functional of the probability measure, however, we omit this dependence whenever it does not cause confusion. We suppose that the researcher is interested in the following target parameter:

$$\beta = \mu\ m(P) \tag{54}$$

Where μ $\mathsf{R}^{N_T^2 N_Z}$ is an identified vector *chosen* by the researcher. It parametrizes the choice of the outcome of interest, as the following remark clarifies.

Remark. Average treatment effect $ATE_{td} = \mathsf{E}[Y(t) - Y(d)]$ for some t, d T; conditional average treatment effect $CATE_{td,A,B} = \mathsf{E}[Y(t) - Y(d)|T$ A, Z $B]$ for given t, d T, A T, B Z, as well as average potential outcomes $\mathsf{E}[Y(t)]$ can all be represented in the form (54).

Some clarifications regarding the relationship of our approach to the literature are in order. Firstly, our setup corresponds to fully non-parametric potential outcomes as we do not assume any form of treatment selection mechanism, unlike the literature following Heckman and Vytlacil (1999, 2005). Whenever Z is a usual IV, additive separability of treatment selection is meaningful as it is equivalent to the LATE independence and monotonicity assumptions (see Vytlacil (2002)). The present paper instead focuses on the *instruments* that are merely auxiliary variables and are not necessarily required to be independent or mean-independent of the potential outcomes. It is therefore not clear why a generalized version of the Heckman-Vytlacil treatment selection should hold in our scenario[18]. Secondly, unlike Mogstad, Santos and Torgovitsky (2018), we do not restrict ourselves to the case of binary outcomes[19] The greater generality that we are willing to allow on the modelling side naturally limits the set of options for the target parameter, β. While accommodating ATE, CATE, and the means of potential outcomes, we are not able to consider, for example, marginal treatment effects.

3.1. Affine inequalities over conditional moments

We now introduce the general class of restrictions under study in this paper. One of our main contributions is an observation that many commonly employed identifying assumptions take the form of shape restrictions that restrict the set of admissible measures to P :

$$P \equiv \{P \quad P|M\ m(P) + b \geq 0\} \tag{55}$$

[18]The nonparametric Roy model is, however, nested in our approach as a potential identifying restriction.

[19]The setup of Mogstad, Santos and Torgovitsky (2018) allows to accommodate some monotonicity conditions by virtue of selecting the appropriate model set M. Notably, it can accommodate MTR ($Y(t)$ - increasing in t), but not monotone instruments.

Where a $R \times 1$ identified vector b and a $R \times N_T^2 N_Z$ identified matrix M are *chosen* by the researcher. These parametrize the choice of R N identifying inequalities on conditional moments of potential outcomes.

Section 2.2 demonstrates that identification results for the family $\mathcal{P}$ may be extended to a bigger family of restrictions $\mathcal{P}$. We may also allow for almost sure linear restrictions on potential outcomes:

$$\mathcal{P} \equiv \{P \quad P|M\ m(P) + b \geq 0 \quad \tilde{M}\mathsf{Y} + \tilde{b} \geq 0 \text{ a.s.}\} \tag{56}$$

Where a $\tilde{R} \times 1$ vector $\tilde{b}$ and a $\tilde{R} \times N_T$ identified matrix $\tilde{M}$ are chosen by the researcher and parameterize the choice of $\tilde{R}$ N almost sure inequalities on the potential outcomes.

The family of models that can be written in the form (56) is very rich, as illustrated by the following examples.

Example 3.1. MIV with Z R (Manski and Pepper, 2000) imposes that for each t $\mathcal{T}$ and z, z $\mathcal{Z}$, $z \geq z =$ $\mathsf{E}[Y(t)|Z = z\,] \geq \mathsf{E}[Y(t)|Z = z]$. It is nested for an appropriate choice of matrix $M = M_{MIV}^{Z}$ and $b = 0$. MTS from (Manski and Pepper, 2000) obtains when $Z = T$.

Example 3.2. IV with Z R imposes that for each t $\mathcal{T}$ and z, z $\mathcal{Z}$, $\mathsf{E}[Y(t)|Z = z\,] = \mathsf{E}[Y(t)|Z = z]$. It is nested for an appropriate choice of matrix M_{IV} that can, for example, be constructed as $M = M_{IV} = \begin{pmatrix} M_{MIV}^{Z} \\ M_{MIV}^{-Z} \end{pmatrix}$ and $b = 0$.

Example 3.3. MTR (Manski and Pepper, 2000) imposes that for each t, t $\mathcal{T}: t > t$, $Y(t\,) \geq Y(t)$ a.s. It is nested for an appropriate choice of matrix $\tilde{M} = \tilde{M}_{MTS}$ with $\tilde{b} = 0$.

Example 3.4. Roy model (Lafférs, 2019) imposes that for each t $\mathcal{T}$, the individual's choice is, on average, optimal $\mathsf{E}[Y(t)|T = t, Z = z] = \max_{d \in \mathcal{T}} \mathsf{E}[Y(d)|T = t, Z = z]$. It is nested for an appropriate choice of matrix $M = M_{ROY}$ and $b = 0$.

Example 3.5. Missing data. Blundell et al. (2007) derives bounds on $F(w|x)$ - the cdf of wages evaluated at some w, with the wages observed if the individual is employed, $E = 1$, and unobserved otherwise, if $E = 0$. Introduce $\mathcal{O} = \{1\}$ and $\mathcal{U} = \{0\}$. Let $Y(t) \equiv \mathsf{I}\{W \leq w\}$, so that $\mathsf{E}[Y(t)|X = x] = F(w|x)$. Our approach allows to accommodate all identifying conditions in the original paper by appropriately choosing M , b and $\tilde{M}, \tilde{b}$.

We omit the construction of the respective matrices here, but we briefly note that assumptions of form (56) can accommodate the commonly imposed almost sure bounds on potential outcomes of form $Y(t)$ $[K_0; K_1]$ a.s. for t $\mathcal{T}$ and known K_0, K_1 R with $K_0 < K_1$.

Combinations of various assumptions can be straightforwardly obtained by stacking the respective matrices, as in Example 2.3. Furthermore, formulations (55) and (56) allow to perform sensitivity analysis through an appropriate choice of b . For example, $\mathsf{E}[Y(t)|Z = z\,] - \mathsf{E}[Y(t)|Z = z] \geq -\delta(z, z\,)$ for all z, z $\mathcal{Z}$ with $z > z$ and t $\mathcal{T}$, for a collection of non-negative $\{\delta(z, z\,)\}_{z,z}$ chosen by the researcher, yields a relaxation of MIV. In applications one could focus on the range of potentially problematic instrument values. For instance, in De Haan (2017) the shape of observed moments may suggest

a potential failure of instrument monotonicity near the boundaries. Selecting positive $\delta(z, z\ )$ for values of $z, z\ $ close to the boundaries can therefore constitute a meaningful robustness check.

3.2. Sharp bounds

The identification procedure that we propose applies to general linear inequality restrictions on conditional moments of potential outcomes, as formulated in (55). To that end, let us construct the vector that collects counterfactual pointwise conditional moments and the vector of those moments that are identified:

$$x \equiv (\mathsf{E}[Y(t)|T = d, Z = z])_{z\ \ Z\ \ (t,d\ \ T:t=d\ \ t,d\ \ U:t=d)}, \quad \overline{x} \equiv (\mathsf{E}[Y(t)|T = t, Z = z])_{z\ \ Z,t\ \ O}$$

Denote the dimension of x as N_x and note that for known selector matrices $P_m, \overline{P}_m$, one can decompose m as:

$$m = P_m x + \overline{P}_m \overline{x}$$

Define the identified set for $\beta\ $ under $P\ $ as:

$$\Theta\ \equiv \{\beta\ \ \mathsf{R}|\ P\ \ P : \beta = \mu\ \ m(P)\ \ b\ + M\ m(P) \geq 0\}$$

We are now ready to state our main identification result.

Theorem 9. Under non-empty $P\ $, the sharp identified set for $\beta\ $ is given by:

$$\Theta\ = \{\beta\ \ \mathsf{R}|\overline{p}\ \overline{x} + \inf_{x:Mx\geq b} p\ x \leq \beta \leq \overline{p}\ \overline{x} + \sup_{x:Mx\geq b} p\ x\} \tag{57}$$

Where:

$$\overline{p} \equiv \overline{P}_m \mu\ , \quad p \equiv P_m \mu \tag{58}$$

$$M \equiv M\ P_m \quad b \equiv -b\ - M\ \overline{P}_m \overline{x} \tag{59}$$

Proof. In the Appendix.

Theorem 1 establishes that under any restriction which can be restated in the form (55), the sharp bounds on the parameter of interest $\beta\ $ can be obtained as the value of a simple linear program. We now wish to obtain a similar representation for the bounds under the more general restrictions of form (56). To that end, let us define the identified set under $P\ $ as:

$$\Theta_I\ \equiv \{\beta\ \ \mathsf{R}|\ P\ \ P : \beta = \mu\ \ m(P)\ \ b\ + M\ m(P) \geq 0\ \ \tilde{M}\mathsf{Y} \geq \tilde{b}\ \text{a.s.}\}$$

Notice that for any pair of $d\ \ T, z\ \ Z$, $\tilde{M}\mathsf{Y} + \tilde{b} \geq 0$ a.s. implies:

$$\tilde{M}\mathsf{E}[\mathsf{Y}|T = d, Z = z] + \tilde{b} \geq 0$$

Which can be equivalently restated as:

$$(I_{N_T N_Z} \quad \tilde{M}) m(P) + \iota_{N_T N_Z} \quad \tilde{b} \geq 0, \tag{60}$$

where is the Kronecker product. Let the matrix M and the vector b combine the conditional restrictions and the implications of the almost sure restrictions. Formally:

$$M \equiv \begin{pmatrix} I_{N_T N_Z} & \tilde{M} \\ & M \end{pmatrix}, \quad b \equiv \begin{pmatrix} \iota_{N_T N_Z} & \tilde{b} \\ & b \end{pmatrix} \tag{61}$$

It then directly follows that:

$$\Theta_I \quad \{\beta \quad \mathsf{R}| \ P \quad P : \beta = \mu \quad m(P) \quad b \quad + M \quad m(P) \geq 0\}, \tag{62}$$

The inverse inclusion does not hold in general[20]. It does, however, hold in three scenarios that appear to be the most relevant in practice. Whenever applied work using non-parametric bounds augments the conditional inequalities by almost sure ones, it usually does so by imposing outcome boundedness, the Monotone Treatment Response condition or both (Blundell et al. (2007), Kreider et al. (2012), Gundersen, Kreider and Pepper (2012), Siddique (2013)). These restrictions are then transformed into a restriction on conditional moments[21], as in (60). Since the general approach has not been available, sharp bounds have only been obtained for some combinations of these almost sure restrictions and the conditional inequalities. For those combinations of almost sure and conditional restrictions that have been studied and for which sharp bounds are available, our approach also attains sharpness.

A notable setup in which, to the best of our knowledge, no such result exists in the continuous outcomes case is the combination of MIV, MTR and MTS conditions[22]. This combination of assumptions is practically relevant because it allows for the tightest bounds under the classical monotonicity restrictions. Consequently, it has been applied even when the theoretical foundation was missing, resulting in the bounds that were not sharp or not theoretically valid.

[20] See the Appendix for a counter-example.
[21] Which is also the approach of Manski and Pepper (2000) in the MIV + MTR and MTR + MTS cases.
[22] See Lafférs (2013) for a review of the methods that have been used to obtain bounds under this combination, and the related fallacies.

Theorem 10. For arbitrary M , b , if one assumes that either:

1. MTR holds, $Y(t_1) \geq Y(t_0)$ a.s. t_1, t_0 T s.t. $t_1 > t_0$:

$$\tilde{M} = \tilde{M}_{MTR} \equiv \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}, \quad \tilde{b} = 0_{N_T - 1} \tag{63}$$

2. Outcomes are bounded, $Y(t)$ $[K_0; K_1]$, t T a.s. for known $K_1 > K_0$:

$$\tilde{M} = \tilde{M}_b \equiv \begin{pmatrix} I_{N_T} \\ -I_{N_T} \end{pmatrix}, \quad \tilde{b} = \tilde{b}_b \equiv \begin{pmatrix} -K_0 \cdot \iota_{N_T} \\ K_1 \cdot \iota_{N_T} \end{pmatrix} \tag{64}$$

3. MTR holds, outcomes are bounded and (Ω, S) can support a $U[0;1]$ r.v.:

$$\tilde{M} = \begin{pmatrix} \tilde{M}_{MTR} \\ \tilde{M}_b \end{pmatrix}, \quad \tilde{b} = \begin{pmatrix} \tilde{b}_{MTR} \\ 0_{N_T - 1} \end{pmatrix} \tag{65}$$

Then the converse of (62) holds and under non-empty P the sharp identified set for β is given by:

$$\Theta = \{\beta \quad R | \overline{p}\,\overline{x} + \inf_{x: Mx \geq b} p\,x \leq \beta \leq \overline{p}\,\overline{x} + \sup_{x: Mx \geq b} p\,x\}, \tag{66}$$

where $M \equiv M$ P_m and $b \equiv -b$ $- M$ $\overline{P}_m$ with M , b defined in (61).

Proof. In the Appendix.

The additional requirement of a $U[0;1]$ random variable existence if MTR and bounded outcomes are assumed is a technical condition needed to establish Lemma 7 in the Appendix. Intuitively, it requires the underlying space of elementary outcomes and the sigma-algebra to be rich enough. If it does not hold, the space may simply not support certain candidate $Y(t)$, which may result in narrower bounds, but which is clearly not a suitable assumption in practice.

Theorems 1 and 2 show that our approach results in sharp bounds under arbitrary conditional moment inequalities, possibly augmented with almost sure outcome boundedness and/or the MTR condition.

Remark. Theorem 2 yields sharp bounds for the MIV + MTS + MTR combination.

Since the emphasis of this paper is on the use of conditional moment inequalities, as opposed to almost sure inequalities on potential outcomes, we do not seek to obtain sharp bounds under general almost sure constraints. This is because it does not seem feasible to study the inverse inclusion in (62) in the continuous case for general $\tilde{M}$ and $\tilde{b}$ without imposing any further assumptions. For example, Mogstad, Santos and Torgovitsky (2018) accommodate almost sure constraints on the potential outcomes by virtue of both

restricting the analysis to the binary case and imposing the Heckman and Vytlacil (2005) selection mechanism, which allows to define the marginal treatment response functions. Our approach is highly complementary, because, unlike Mogstad, Santos and Torgovitsky (2018), we accommodate arbitrary linear inequalities on conditional moments.

Since the bounds in Theorems 1 and 2 are sharp, they are equivalent to sharp bounds obtained in some special cases by Manski and Pepper (2000), Blundell et al. (2007), Boes (2009), Siddique (2013) and others. The analysis in this paper, however, also extends to the cases in which such bounds cannot be derived analytically. For example, it allows the researcher to combine various identifying assumptions, to introduce arbitrary relaxations of the usually employed conditions and to consider more complicated setups. Theorems 1 and 2 thus present a unifying theory of identification under shape restrictions over linear combinations of conditional moments and commonly used almost sure linear inequalities.

Notably, our formulation solves the identification problem in the space of conditional moments instead of searching over all possible joint probability distributions. The latter technique is not applicable whenever potential outcomes are continuous, because the corresponding program becomes infinite-dimensional[23].

An important special case of Theorem 2 obtains under the *conditional monotonicity* family of assumptions. The more complicated structure of cMIV conditions has precluded previous research from obtaining the analytical sharp bounds. As Section 4 explains, under some versions of cMIV these bounds can only be obtained in the LP form, which additionally motivates the more general approach of Theorems 1 and 2.

4. cMIV assumptions

A particular family of identifying conditions that can be written in the form (56) is the *conditional monotonicity* class of assumptions. These impose that potential outcomes are mean-monotone in the instrument even within some treatment subgroups. While more restrictive than the conventional MIV, conditionally monotone instrumental variables (cMIV) allow to sharpen the bounds on the outcomes of interest. Throughout this section, we assume that outcomes are bounded $Y(t) \quad [K_0; K_1]$ *a.s.* for $K_0, K_1 \quad \mathsf{R}$, $K_0 < K_1$ and the bounds are known. We also suppose that there are no missing data[24], i.e. $T = O$.

We argue that cMIV assumptions are reasonable in classical applications, provide examples to illustrate the difference between MIV and cMIV and develop a formal testing strategy for a particular version of cMIV. This testing procedure relies on the observed outcomes' monotonicity, which has been typically used in applied work to justify applying MIV. Our results show that if such monotonicity is observed and the researcher is comfortable with MIV, the cMIV assumption is *inexpensive*, and can be applied to sharpen the bounds on the outcomes of interest. In some applications, as is the case in Section 5, cMIV allows to obtain informative bounds on the parameters of interest even though the MIV assumption does not.

[23] Making the problem infinite-dimensional complicates estimation and inference up to a point where no 'good' methods seem to be available; albeit an identification result may be obtained in the continuous case, the estimation stage will likely need to be discrete.

[24] Although it is hopefully clear from our general approach how cMIV conditions extend to the missing data case.

We now present three versions of conditional monotonicity assumptions. While we only consider these three variations, we should note that the class of such assumptions is richer[25] and Theorem 2 applies in any such conditional and unconditional monotonicity frameworks.

Assumption cMIV-s. *Suppose that for any* $t \quad T$, $A \quad T : A = \{t\}$ *and* $z, z \quad Z$ *s.t.* $z > z$ *we have:*

$$\mathsf{E}[Y(t)|T \quad A, Z = z\,] \geq \mathsf{E}[Y(t)|T \quad A, Z = z] \tag{67}$$

i.e. the potential outcomes are, on average, non-decreasing in Z *for any treatment subgroup.*

The strong conditional monotonicity assumption possesses the greatest identifying power across all cMIV conditions. The sharp bounds for it, however, can only be obtained as value functions of linear programs in Theorem 2. To see that Assumption cMIV-s implies MIV, set $A = T$ in the above definition.

Assumption cMIV-w. *Suppose MIV holds and for any* $t \quad T$ *and* $z, z \quad Z$ *s.t.* $z > z$ *we have:*

$$\mathsf{E}[Y(t)|T = t, Z = z\,] \geq \mathsf{E}[Y(t)|T = t, Z = z] \tag{68}$$

i.e. the potential outcomes are, on average, non-decreasing in Z *for the non-treated subgroup and for the whole population.*

The weak conditional monotoncity assumption allows for closed-form expressions for sharp bounds that are easy to compute and perform inference on.

Assumption cMIV-p. *Suppose MIV holds and for any* $t \quad T, d \quad T \setminus \{t\}$ *and* $z, z \quad Z$ *s.t.* $z > z$ *we have:*

$$\mathsf{E}[Y(t)|T = d, Z = z\,] \geq \mathsf{E}[Y(t)|T = d, Z = z] \tag{69}$$

i.e. the potential outcomes are, on average, non-decreasing in Z *conditional on any counterfactual level of treatment.*

The pointwise conditional monotonicity assumption, as we shall see in Section 3.2, allows for the cleanest mathematical intuition. The test of this form of monotonicity is also derived in Section 3.3.

We call Z a *strong (weak) conditionally monotone instrument* (for brevity, strong (weak) cMIV) if it satisfies Assumption cMIV-s and cMIV-w respectively. If Z satisfies cMIV-p, we call it a *pointwise conditionally monotone instrument*.

Conditional monotonicity restrictions differ in the collection of treatment subsets over which monotonicity in the instrument is assumed. The strong conditionally monotone instruments are such that, among individuals from any given counterfactual treatment subgroup, higher values of Z are, on average, associated with higher potential outcomes. The weak conditional monotonicity restriction only imposes the same mean-monotonicity

[25] One can consider the class of conditional restrictions $\mathsf{E}[Y(t)|T \quad A, Z = z'] \geq \mathsf{E}[Y(t)|T \quad A, Z = z], \quad A \quad F_t$ for all $t \quad T$ where subcollections $F_t \quad T$ are chosen by the researcher.

on the whole population and on the untreated, whereas the pointwise form assumes it over the entire population as well as conditional on each counterfactual level of treatment.

Remark. All cMIV assumptions imply MIV. Moreover, cMIV-w, cMIV-p are implied by cMIV-s. If treatment is binary, cMIV-s, cMIV-w and cMIV-p are equivalent.

While this is possible for the general apporach of form (55), cMIV conditions avoid assuming monotonicity over the observed treatment subset $\{T = t\}$. This is because such monotonicity is identified. If it holds, it should not add any identifying power to our conditions in theory. In practice, imposing such monotonicity may render the LP infeasible, because under noisy data it may fail in finite samples, even if it holds in the population. On the other hand, large violations of the observed outcomes' monotonicity will lead the test developed in Section 3.3 to reject cMIV-p and cMIV-s.

The following observation motivates the use of cMIV assumptions.

Proposition 7. *Manski and Pepper (2000) MIV bounds are not sharp under either cMIV-s, cMIV-w or cMIV-p.*

Proof. Consider a binary treatment T and three levels of the instrument $Z \quad \{z_0, z_1, z_2\}$ with $z_0 < z_1 < z_2$. Suppose for a fixed $t \quad \{0, 1\}$, we have $\mathsf{E}[Y(t)|T = t, Z = z_i] = 0$, with $P[T = t|Z = z_0] = 0.125$, $P[T = t|Z = z_1] = 0.5$, $P[T = t|Z = z_2] = 0.25$. Further impose $-K_0 = K_1 = 1$. The no-assumptions bounds on $\mathsf{E}[Y(t)|Z = z_i]$ are:

$$[-0.125; 0.125],\ [-0.5; 0.5],\ [-0.25; 0.25]$$

The MIV lower bounds 'iron' the no-assumptions bounds:

$$-0.125,\ -0.125,\ -0.125$$

Which also implies the following lower bounds on $\mathsf{E}[Y(t)|T = t, Z = z_i]$:

$$-1,\ -0.25,\ -0.5$$

Under cMIV assumptions, one can 'iron' the above array to improve the lower bound for z_2 up to -0.25, so that the lower bound on $\mathsf{E}[Y(t)|Z = z_2]$ becomes $-1/16 > -1/8$.

Sharp bounds for all versions for cMIV follow from Theorem 2. We also show that under cMIV-w the bounds can be characterized explicitly, which is especially convenient if the treatment is binary, so that all cMIV assumptions coincide. For didactic purposes, we further provide the detailed construction of the triplet M, c, p from Theorem 2 under cMIV-s and cMIV-p. All details on the identification under cMIV are provided in the Appendix 7.5.

4.1. Discussion of cMIV

This section illustrates the difference between MIV and cMIV assumptions by considering two parametric examples with classical applications. Any restrictions imposed for this purpose do not apply to the rest of the paper.

4.1.a. Education selection. To better illustrate the distinction between MIV and cMIV, consider the following empirical setup. Suppose T is an indicator of whether or not an individual has a university degree, $Y(t)$ are potential log wages and Z is an observed indicator of ability.

The usual MIV assumption on Z implies that more able individuals can do better both with and without a college degree on average: $\mathsf{E}[Y(t)|Z=z]$ - monotone in z. cMIV additionally imposes that: i) among those who have a college degree, a *smarter* individual could have done relatively better on average than their counterpart if both did not have it: $\mathsf{E}[Y(0)|Z=z,T=1]$ - monotone in z; and ii) among those who do not have a college degree, a *smarter* individual could have done relatively better on average than their counterpart if both had it: $\mathsf{E}[Y(1)|Z=z,T=0]$ - monotone in z.

We now consider a parametric example. Suppose that η measures how *diligent* one is from birth and is ex-ante mean-independent of Z. While Z is observed by both the employers and the econometrician (e.g. observing an IQ score), the employer additionally observes a noisy signal of diligence by virtue of knowing the employee's effort level: $\eta+\varepsilon$ and $\varepsilon \perp\!\!\!\perp (Z,T,\eta)$. Suppose $Var(Z)=Var(\eta)=1$ and $\mathsf{E}[Z]=\mathsf{E}[\eta]=\mathsf{E}[\varepsilon]=0$.

Consider a stylized Roy selection model with:

$$Y(t)=\beta_0(t)+\beta_1(t)Z+\beta_2(t)\eta+\varepsilon(t), \quad T=\mathsf{I}\{\mathsf{E}[Y(1)-Y(0)|Z,\eta]+\nu\geq 0\},$$

where $\nu \perp\!\!\!\perp (Z,\eta,\varepsilon)$ and we let $\varepsilon(t)\equiv\beta_2(t)\varepsilon$. In this case MIV collapses to:

$$(\text{MIV}): \beta_1(t)\geq 0$$

MIV postulates that the direct effect of ability on potential earnings is positive. It seems reasonable to suppose that $\beta_i(t)\geq 0,\ i=1,2,\ t=0,1$, i.e. both diligence and ability increase potential wages.

Letting $\delta_z\equiv\beta_1(1)-\beta_1(0)$ and $\delta_\eta\equiv\beta_2(1)-\beta_2(0)$ denote the differential in the effects of ability and effort respectively on the potential wage schedule, the additional requirement of cMIV is that:

$$\underbrace{\beta_1(0)z}_{\text{direct effect}}+\underbrace{\beta_2(0)\mathsf{E}[\eta|\delta_z z+\delta_\eta\eta+\tilde{\nu}\geq 0]}_{\text{selection given } T=1} - \text{increasing} \tag{70}$$

$$\underbrace{\beta_1(1)z}_{\text{direct effect}}+\underbrace{\beta_2(1)\mathsf{E}[\eta|\delta_z z+\delta_\eta\eta+\tilde{\nu}\leq 0]}_{\text{selection given } T=0} - \text{increasing}, \tag{71}$$

where $\tilde{\nu}\equiv\beta_0(1)-\beta_0(0)+\nu$.

Notice that if δ_z and δ_η are of different signs, for example because the jobs that one may apply for with a college degree are more ability-intensive ($\delta_z>0$), whereas those which are available otherwise are more skill-intensive ($\delta_\eta<0$), the additional conditional monotonicity requirements (70)-(71) are less strict than MIV. This is because, *conditional* on both having a degree and not having it, ability and effort are *positively* associated.

Intuitively, among those who do not have a degree ($T=0$), people of higher ability must have had stronger incentives to forgo college. This should have been because a higher level of diligence gives them a comparative advantage in effort-intensive jobs. Among those with a degree, higher ability implies a comparative advantage in ability-intensive

occupations, which explains their willingness to select into this option ($T = 1$). It does not, therefore, signal as low an effort level as it would for a less capable individual.

Now consider the same setup with[26]:

$$T = \mathsf{I}\{\eta + Z \geq 0\}$$

This selection mechanism can be explained by the fact that to get a degree one needs to be either hard-working or of high ability. The requirement of MIV is unchanged, and cMIV necessitates that:

$$\beta_1(0)z + \beta_2(0)\mathsf{E}[\eta|\eta \geq -z] - \text{increasing} \tag{72}$$

$$\beta_1(1)z + \beta_2(1)\mathsf{E}[\eta|\eta \leq -z] - \text{increasing} \tag{73}$$

In this case, conditional on each level of education, effort level η and ability Z are negatively associated, so the conditional selection terms in (72)-(73) make cMIV a stricter assumption than MIV.

Intuitively, a more able individual with a college degree did not need to work as hard to get it relative to her counterpart with a lower ability. Similarly, if an individual is capable, but does not have a degree, she has to be of rather low effort as otherwise she would have selected into education.

Even if MIV holds, cMIV can fail if employer prefers effort over ability to the extent that the negative association between the two conditional on having or not having a degree outweighs the direct impact of ability on wages as well as any ex-ante positive correlation between the employer-observed signal of diligence and the ability.

An examination of equations (70) and (71) suggests that cMIV is more likely to hold whenever δ_z is small relative to δ_η, while $\beta_1(\cdot)$ is large relative to $\beta_2(\cdot)$. This means that Z should be *relatively weak* in the parlance of the classical IV models, and *strongly monotone*. cMIV is also more likely to hold the more noisy is the selection mechanism. This is because a greater variance of the idiosyncratic component[27] $\tilde{\nu}$ makes the conditional link between Z and η more loose.

Overall, it seems reasonable to use a proxy for the level of ability as a conditionally monotone instrument in the estimation of returns to schooling. One would be inclined to think that while Z does enter selection, it affects the potential outcomes directly and strongly enough, so that there are no subgroups by schooling for which a higher value of ability would correspond to lower potential wages on average.

4.1.b. Simultaneous equations. As some aspects of mathematical intuition may be muted in discrete models, we also consider a simple continuous setup to confirm the insights derived from the previous analysis. For illustrative purposes, drop the boundedness and discreteness assumptions and consider the demand and supply simultaneous equations:

$$q^k(p) = \alpha^k(p) + \beta^k(p)Z + \gamma^k(p)\eta + \kappa^k(p)\varepsilon^k, \;\; k \quad \{s, d\}$$

[26] Setting $\delta_z = \delta_\eta > 0$ and $\tilde{\nu} = 0$.

[27] This holds formally in jointly normal case.

The observed log-price P clears the market:

$$P \quad \{p \quad \mathsf{R}|\mathsf{E}[q^s(p)|Z,\eta] = \mathsf{E}[q^d(p)|Z,\eta]\}, \tag{74}$$

where η, Z are continuous unobserved and observed random variables respectively, with $\mathsf{E}[\eta|Z=z]=0$[28] and $\mathsf{E}[\varepsilon^k]=0$ with $\varepsilon^k \perp\!\!\!\perp (\eta, Z, \varepsilon^{-k})$ for $k \quad \{s,d\}$. Further assume that all functions of p are continuous.

Potential price p indexes the potential outcomes, giving rise to the demand and supply *schedules*. Suppose we aim to identify the supply elasticity, so that the relevant potential outcomes are $q^s(p)$. Z is a considered monotone instrument, while P can be interpreted as treatment. η is unobserved heterogeneity and ε^k are random violations from the market clearing condition or measurement errors independent of the rest of the model. For an individual realization of market clearing an econometrician observes $\{P, \{q^k(P)\}_k, Z\}$, but does not observe the schedules at other prices $\{q^k(p)\}_k$ for $p = P$, nor disturbances $\{\eta, \{\varepsilon^k\}_k\}$.

Define $\delta_z(p) \equiv \beta^s(p) - \beta^d(p)$ and similarly for η, with $\delta_p(p) \equiv \alpha^s(p) - \alpha^d(p)$. As stated, the model is potentially *incomplete* or *incoherent*, as for a given vector (Z,η) equation (74) may have multiple or no solutions. To avoid that, so long as that the support of Z, η, ε^k is full, it is necessary that $\delta_z(p)$, $\delta_\eta(p)$ be constant. We shall assume that for simplicity. Provided that $\delta_p(p)$, which determines the *excess supply* at fixed (Z,η), is strictly increasing and has full image, the model is *complete* and *coherent* and:

$$P = \delta_p^{-1}(-\delta_z Z - \delta_\eta \eta) \tag{75}$$

Equation (75) introduces a deterministic linear relationship between Z and η conditional on each given value of P. As we saw in the previous example, this constitutes the worst-case scenario for cMIV, if δ_z and δ_η have the same sign. A noisier selection mechanism would relax the conditional link between Z and η, and would thus weaken the conditional selection channel.

Note that the reduced-form error is $u \equiv \gamma^s(P)\eta + \kappa^s(P)\varepsilon^s$ and there is a simultaneity bias:

$$\mathsf{E}[Pu] = \mathsf{E}[P\gamma^s(P) \quad \underbrace{\mathsf{E}[\eta|\delta_z Z + \delta_\eta \eta = P]}_{\text{simultaneity/ommited variable}} \quad] = 0$$

In this setup, MIV requires:

$$(\text{MIV}): \ \beta^s(p) \geq 0, \quad p \quad \mathsf{R}$$

Whereas cMIV additionally imposes that:

$$\beta^s(p)z + \gamma^s(p)\mathsf{E}[\eta|\delta_z z + \delta_\eta \eta = -\delta_p(d)] \geq 0 \text{ – increasing in } z, \quad p,d \quad \mathsf{R}: d = p \tag{76}$$

[28] Note that, once again, mean independence is not restrictive, as otherwise we could always redefine the data generating process in an observationally equivalent way.

Suppose that $\delta_z, \delta_\eta = 0$ to rule out uninteresting cases. (76) rewrites as:

$$\beta^s(p) \geq \gamma^s(p) \frac{\beta^s(p) - \beta^d(p)}{\gamma^s(p) - \gamma^d(p)} \tag{77}$$

For concreteness, consider two positive supply shocks, i.e. $\beta^s(p), \gamma^s(p) > 0$. Equation (77) then says that either η and Z affect the reduced-form equilibrium price in different directions (recall the comparative advantage example), or the effect of Z on the equilibrium price relative to its effect on the supply schedule is smaller than that of η:

$$sgn(\delta_\eta) = sgn(\delta_z) \quad \text{or} \quad \left| \frac{\beta^s(p) - \beta^d(p)}{\beta^s(p)} \right| \leq \left| \frac{\gamma^s(p) - \gamma^d(p)}{\gamma^s(p)} \right| \tag{78}$$

Under $sgn(\delta_\eta) = sgn(\delta_z)$, equation (78) once again requires that Z be *strongly monotone* and *relatively weak*. The logic we described may help the researcher navigate the potential economic forces in a given application to decide whether cMIV is a suitable assumption.

For example, consider estimating the supply elasticity in the market for plane tickets in the early days of Covid-19 pandemic. Suppose Z is an inverse Covid-stringency index for the economy, while η may be interpreted as residual cost shocks, defined to be mean-independent of Z. It is likely that $\delta_\eta \approx \gamma^s$, i.e. residual cost shocks affect mainly the supply in that sector, and not the demand. It is also likely that either supply is less responsive to Z than demand (so that cMIV is implied by MIV), or the effects are of the same order of magnitude. Z is therefore likely to be a conditionally monotone instrument.

4.2. Testing cMIV

One could argue that cMIV-s or cMIV-p can be rejected whenever $\mathsf{E}[Y(t)|T = t, Z = z]$ fails to be monotone in the data. In general, the power of that test is not immediately clear, for example if the outcomes fail to be conditionally mean-monotone on some other subset of the support. There is, however, a special case when cMIV can be tested directly, given that the researcher believes in MIV. In some applications one may conjecture that the potential outcomes' functions $Y(t)$, either in the reduced or in the structural form, are such that the relative effects of Z and the unobserved variable(s) η, potentially correlated with Z, are unchanged across outcome indices t.

In practice, researchers often impose even stricter versions of this homogeneity assumption. For example, Manski and Pepper (2009) discuss MIV-driven identification under HLR assumption, which amounts to imposing $Y(t) = \beta t + \eta$. Conditions in Proposition 4 relax HLR to an arbitrary shape of response of a potential outcome to treatment and allow for a generally heteroscedastic/treatment-specific response to unobserved variables and instrument, so long as the relative effects are unchanged across potential outcomes.

Recalling that $\mathsf{E}[Y(t)|T = t, Z = z]$ is identified, testing the counterfactual part of Assumption cMIV-p becomes equivalent to testing the monotonicity of $\mathsf{E}[Y(t)|T = t, Z = z]$ given that MIV holds:

Proposition 8. *Suppose that a): i)* $Y(t) = g(t,\xi) + h(t)\psi(Z,\eta)$, $h(t) = 0$ *with* $\xi \perp\!\!\!\perp (T,Z,\eta)$ *and ii) MIV holds, strictly for some* z, z Z *with* $z > z$*; or b): i)* $Y(t) = g(t,\xi,T) + h(t)\psi(Z,\eta)$ *with* $\xi \perp\!\!\!\perp (T,Z,\eta)$*, ii)* $\frac{h(t)}{h(d)} > 0$ t, d T*; and iii) MIV holds. Then Assumption cMIV-p holds i* $\mathsf{E}[Y(t)|T = t, Z = z]$ *are all monotone.*

Proof. In the Appendix.

Clearly, in case treatment is binary, the same holds for cMIV-s and cMIV-w. Further note that whether or not $h(t) = 0$ is observable in the data for case (a) and whether or not $h(t)/h(d) > 0$ is also identified for (b).

cMIV is testable in the Example 3.2.2, because the reduced form expression has the form $b) : i)$. It also becomes testable in the Example 3.2.1 if instead of separately observing η, Z, employers on average observe a mixed signal of ability and effort, $s \equiv aZ + b\eta$ for some a, b R.

It is important to note that in practice the monotonicity of *observed* outcomes has been routinely used to motivate the use of MIV (e.g. De Haan (2017)). Here we have shown that under a homogeneity condition the same *observed monotonicity* establishes cMIV-p under MIV.

Under the conditions in Proposition 4, a test of cMIV-p is thus the test of all $f_t(z) \equiv \mathsf{E}[Y(t)|T = t, Z = z]$ being monotone. Formally, we test the following null hypothesis:

$$H_0 : f_t(z) - \text{increasing in } z, \quad t \quad T \tag{79}$$

For this purpose, we may extend the testing procedure developed in Chetverikov (2019) to accommodate the test of joint monotonicity of the conditional moments[29]. Denote the set of all observations with treatment level t as $I_t \equiv \{i \quad \overline{1,n} : T_i = t\}$ with $n_t \equiv |I_t|$. Suppose $\phi^t_{n_t}$ is the corresponding Chetverikov's regression monotonicity test (or a corresponding parametric test for discrete Z) with the confidence level $\alpha_t \quad (0; 0.5)$. We define the joint test as:

$$\phi_n \equiv \max_{t \ T} \ \phi^t_{n_t} \tag{80}$$

Denote P^C to be the set of probability measures, such that for all $P \quad P^C$ and all $t \quad T$ the conditional probability measure given $T = t$ that P generates satisfies the regularity conditions in Theorem 3.1 in Chetverikov (2019). Similarly, let P^C_t be the set of all the conditional probability measures given $T = t$ that measures from P^C generate. In the proposition below, we interpret H_0 as a set of such measures that satisfy (79).

[29] This test is developed for continuous Z, which is used in our application. Although the instrument is discretized at the estimation stage, the monotonicity of $\mathsf{E}[Y(t)|T = t, Z = z]$ for continuous Z clearly implies the monotonicity of the discretized moments. The procedure we describe straightforwardly accommodates testing discrete instruments. As noted in Chetverikov (2019), for discrete conditioning variable the test is a simple parametric problem, since the conditional moment function can be $\sqrt{n}$-consistently estimated at each point from the support.

Proposition 9. *If* $\Pi_t \ T\,(1-\alpha_t) \geq 1-\alpha$, *then:*

$$\inf_{P \ P^C \cap H_0} P[\phi_n = 0] \geq 1-\alpha+o(1) \tag{81}$$

as $n \to \infty$.

Proof. Notice that each $\phi^t_{n_t}$ is a function of the observations from I_t only. Since I_t are mutually exclusive by construction and because the data are i.i.d., we have $P[\phi_n = 0] = \Pi_t \ T\, P[\phi^t_{n_t} = 0]$.

By standard optimization arguments:

$$\Pi_t \ T \inf_{P \ P^C_t} P[\phi^t_{n_t} = 0] \leq \inf_{P \ P^C} \Pi_t \ T\, P[\phi^t_{n_t} = 0] \tag{82}$$

Theorem 3.1 from Chetverikov (2019) and $\Pi_t \ T\,(1-\alpha_t) \geq 1-\alpha$ then yield the result.

Remark. One may set $\alpha_t = 1-(1-\alpha)^{1/N_T}$ as a baseline. If the domain knowledge suggests that for some treatments monotonicity is more likely to hold, one can set a higher α_t for them, so long as $\Pi_t \ T\,(1-\alpha_t) \geq 1-\alpha$. This may improve the power of the test.

We implement this test when studying the returns to education in Colombia, see Section 5.

5. Returns to education in Colombia

Our data is comprised of 861492 observations from Colombian labor force. The sample represents a snapshot of those individuals who could be matched across the educational, formal employment and census datasets in 2021[30]. For 664633 individuals from this dataset we observe their average lifetime wages, education level and Saber 5 or Saber 11 scores for Mathematics and Spanish language tests[31].

The outcome variable we consider (Y_i) is a log-wage, and T_i is the education level. We distinguish four education levels: primary, secondary and high school as well as 'university'[32]. Our measure of ability is constructed as a CES aggregator, which is then split into deciles:

$$Z_i \equiv (MATH_i^{1/2} + SPANISH_i^{1/2})^2$$

[30] Educational dataset was assembled by the testing authority Instituto Colombiano para la Evaluación de la Educación (ICFES), formal employment dataset comes from social security data based on Planilla Integrada de Liquidación de Aportes (PILA), whereas census data is handled by Departamento Administrativo Nacional de Estadística (DANE). The data was merged and anonymized by ICFES.

[31] Saber 5 and 11 tests are taken at different ages, but designed to be comparable between each other, which justifies merging them.

[32] T_i is based on the number of years of schooling, S_i. If $S_i < 9$, set $T_i \equiv 0$ meaning the individual only graduated from primary school. $S_i \quad [9;11)$ and $T_i \equiv 1$ correspond to completing compulsory education (secondary school), $S_i = 11$ and $T_i \equiv 2$ means that the individual is a high-school graduate, whereas $S_i > 11$ with $T_i \equiv 3$ means university education. Unfortunately, S_i is capped at 17 years in our sample, making it impossible to distinguish between those who continued to graduate education and those who just finished the 6−years degree.

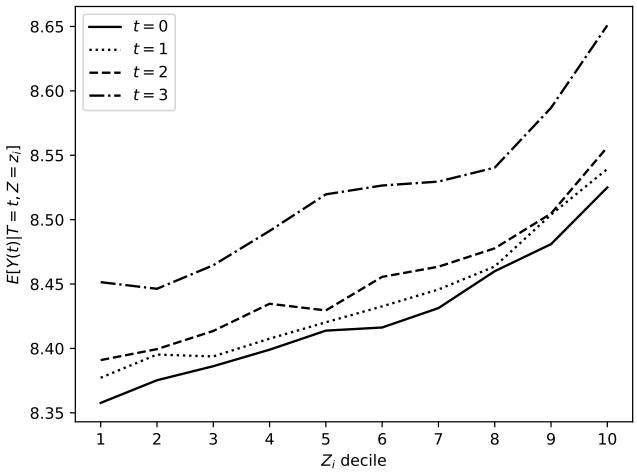


Figure 4: Estimated conditional moments of log-wages given ability and education level.

We first test whether cMIV is a reasonable assumption in our setup by implementing the test discussed in Section 3.3. To that end, we use the parameters and kernel functions recommended by Chetverikov (2019) and focus on the theoretically most powerful procedure, the step-down approach. The estimated p-value of the test is 0.29, see Table 1. We thus conclude that cMIV-p is a credible assumption provided that MIV holds.

t	R_t^{st}	$R_{t;0.1}^{crit}$	p-value	n_t
0	0.98	2.33	0.34	274295
1	-1.17	2.17	0.95	143299
2	-1.51	2.30	1.00	216336
3	1.86	2.38	0.08	30703

Table 1: Results of the monotonicity test, see Section 3.2. Second column gives the estimated Chetverikov (2019) test-statistic, third column contains the $\alpha = 0.1$ critical values, corresponding to $\alpha_t = 1 - (1 - 0.1)^{1/4} \approx 0.026$ individual critical value. The last column gives a p-value against the individual null for each t. The overall p-value is 0.29.

The data we study is rather noisy. One would expect a considerable measurement error in the construction of both treatment levels and the outcome variable[33]. In line with that, the strongest form of cMIV is not sufficient to provide identification in the absence of further assumptions. While the resulting bounds are tighter than that under MIV, they remain uninformative.

To achieve identification, we augment our assumptions with the MTR condition. While MIV and cMIV-w remain uninformative, both cMIV-p and cMIV-s result in positive lower bounds on the ATEs. Under cMIV-p the effect of obtaining a 'university education' is estimated to be at least as large as 3.62%, and 5.91% under cMIV-s. This is consistent with previous evidence. Causal estimates for the US (Card (1993), Brand and Xie (2010)

[33] In particular, age is self-reported when filling an online questionnaire and appears to be of low quality, so we are forced to merge multiple cohorts.

and Angrist and Chen (2011)) report the return of at least 10% for a 4−year college degree. Recent evidence suggests that this number may be substantially lower for Colombia (Gomez, 2022).

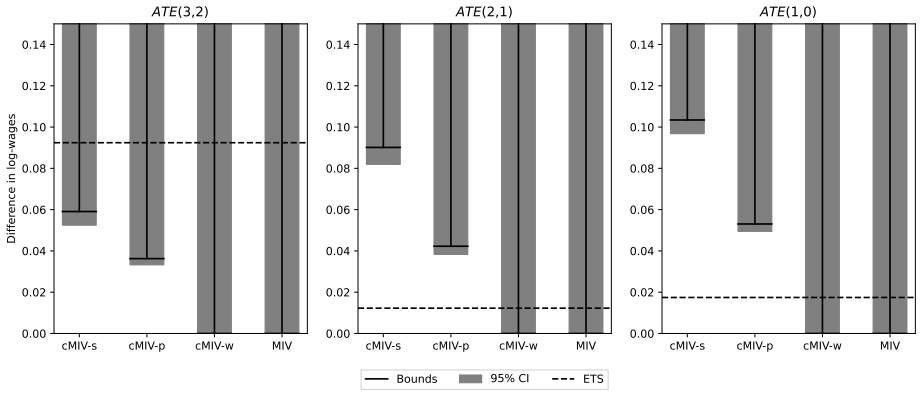


Figure 5: Estimation results for the monotonicity assumptions augmented with MTR. CI constructed according to Proposition 11. The exogenous treatment selection estimates (ETS) are $ATE^{ETS}_{t,d} \equiv \mathsf{E}[Y(t)|T=t] - \mathsf{E}[Y(d)|T=d]$

We also find significantly positive effects at other education stages, see Figure 5. Further details on data construction and estimation as well as robustness checks are available in the Appendix.

References

Aliprantis, C.D. and K.C. Border. 2007. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer.
URL: *https://books.google.com/books?id=4hIq6ExH7NoC*

Andrews, Donald WK. 1999. "Estimation when a parameter is on a boundary." *Econometrica* 67(6):1341–1383.

Andrews, Isaiah, Jonathan Roth and Ariel Pakes. 2023. "Inference for Linear Conditional Moment Inequalities." *The Review of Economic Studies* 90(6):2763–2791.
URL: *https://doi.org/10.1093/restud/rdad004*

Angrist, Joshua D, Guido W Imbens and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91(434):444–455.

Angrist, Joshua D and Stacey H Chen. 2011. "Schooling and the Vietnam-era GI Bill: Evidence from the draft lottery." *American Economic Journal: Applied Economics* 3(2):96–118.

Beresteanu, Arie and Francesca Molinari. 2008. "Asymptotic properties for a class of partially identified models." *Econometrica* 76(4):763–814.

Bertsekas, Dimitri P. 1975. "Necessary and sufficient conditions for a penalty method to be exact." *Mathematical programming* 9(1):87–99.

Blundell, Richard, Amanda Gosling, Hidehiko Ichimura and Costas Meghir. 2007. "Changes in the distribution of male and female wages accounting for employment composition using bounds." *Econometrica* 75(2):323–363.

Boes, Stefan. 2009. "Bounds on counterfactual distributions under semi-monotonicity constraints.".

Brand, Jennie E. and Yu Xie. 2010. "Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." *American Sociological Review* 75(2):273–302. PMID: 20454549.
URL: *https://doi.org/10.1177/0003122410363567*

Card, David. 1993. "Using geographic variation in college proximity to estimate the return to schooling.".

Chernozhukov, Victor, Han Hong and Elie Tamer. 2007. "Estimation and Confidence Regions for Parameter Sets in Econometric Models." *Econometrica* 75(5):1243–1284.
URL: *http://www.jstor.org/stable/4502031*

Chernozhukov, Victor, Sokbae Lee and Adam M. Rosen. 2013. "Intersection Bounds: Estimation and Inference." *Econometrica* 81(2):667–737.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA8718*

Chetverikov, Denis. 2019. "TESTING REGRESSION MONOTONICITY IN ECONOMETRIC MODELS." *Econometric Theory* 35(4):729–776.

Cho, JoonHwan and Thomas M. Russell. 2023. "Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments." *Journal of Business & Economic Statistics* 0(0):1–16.
URL: *https://doi.org/10.1080/07350015.2023.2203768*

Cygan-Rehm, Kamila, Daniel Kuehnle and Michael Oberfichtner. 2017. "Bounding the causal effect of unemployment on mental health: Nonparametric evidence from four countries." *Health Economics* 26(12):1844–1861.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.3510*

De Haan, Monique. 2017. "The Effect of Additional Funds for Low-ability Pupils: A Non-parametric Bounds Analysis." *The Economic Journal* 127(599):177–198.

Demyanov, Vladimir F. 2009. *Minimax: directional di erentiabilityMinimax: Directional Di erentiability*. Boston, MA: Springer US pp. 2075–2079.
URL: *https://doi.org/10.1007/978-0-387-74759-0*$_3$69

Duan, Qingsong, Mengwei Xu, Liwei Zhang and Sainan Zhang. 2020. "Hadamard directional

differentiability of the optimal value of a linear second-order conic programming problem." *Journal of Industrial and Management Optimization* 17(6):3085–3098.

Fang, Zheng and Andres Santos. 2018. "Inference on Directionally Differentiable Functions." *The Review of Economic Studies* 86(1):377–412.
URL: *https://doi.org/10.1093/restud/rdy049*

Gafarov, Bulat. 2024. "Simple subvector inference on sharp identified set in affine models." *arXiv e-prints* pp. arXiv–1904. Conditionally Accepted at Journal of Econometrics, 2024.

Gomez, Norma. 2022. "Returns to college education in Colombia." *Higher Education Policy* 35(3):692–708.

Gundersen, Craig, Brent Kreider and John Pepper. 2012. "The impact of the National School Lunch Program on child health: A nonparametric bounds analysis." *Journal of Econometrics* 166(1):79–91. Annals Issue on "Identification and Decisions", in Honor of Chuck Manski's 60th Birthday.
URL: *https://www.sciencedirect.com/science/article/pii/S0304407611001205*

Hansen, Bruce E. 2017. "Regression kink with an unknown threshold." *Journal of Business & Economic Statistics* 35(2):228–240.

Heckman, James J and Edward J Vytlacil. 1999. "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the national Academy of Sciences* 96(8):4730–4734.

Heckman, James J and Edward Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation 1." *Econometrica* 73(3):669–738.

Hong, Han and Jessie Li. 2015. The numerical delta method and bootstrap. Technical report Working paper.

Imbens, Guido W. and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72(6):1845–1857.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2004.00555.x*

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2):467–475.
URL: *http://www.jstor.org/stable/2951620*

Kline, Brendan and Elie Tamer. 2023. "Recent Developments in Partial Identification." *Annual Review of Economics* 15(Volume 15, 2023):125–150.
URL: *https://www.annualreviews.org/content/journals/10.1146/annurev-economics-051520-021124*

Kreider, Brent, John V Pepper, Craig Gundersen and Dean Jolliffe. 2012. "Identifying the effects of SNAP (food stamps) on child health outcomes when participation is endogenous and misreported." *Journal of the American Statistical Association* 107(499):958–975.

Laffers, Lukáš. 2019. "Bounding average treatment effects using linear programming." *Empirical economics* 57:727–767.

Laffers, Lukáš. 2013. "A note on bounding average treatment effects." *Economics Letters* 120(3):424–428.
URL: *https://www.sciencedirect.com/science/article/pii/S0165176513002668*

Li, Wu. 1993. "The sharp Lipschitz constants for feasible and optimal solutions of a perturbed linear program." *Linear algebra and its applications* 187:15–40.

Manski, Charles F. 1997. "Monotone Treatment Response." *Econometrica* 65(6):1311–1334.
URL: *http://www.jstor.org/stable/2171738*

Manski, Charles F. and John V. Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68(4):997–1010.
URL: *http://www.jstor.org/stable/2999533*

Manski, Charles F. and John V. Pepper. 2009. "More on monotone instrumental variables." *The*

Econometrics Journal 12(suppl$_1$) : $S200 - -S216$.
URL:*https://doi.org/10.1111/j.1368-423X.2008.00262.x*

Masten, Matthew A. and Alexandre Poirier. 2018. "IDENTIFICATION OF TREATMENT EFFECTS UNDER CONDITIONAL PARTIAL INDEPENDENCE." *Econometrica* 86(1):317–351.
URL: *http://www.jstor.org/stable/44955202*

Meyer, Robert. 1979. Continuity properties of linear programs. Technical report University of Wisconsin-Madison Department of Computer Sciences.

Mogstad, Magne, Andres Santos and Alexander Torgovitsky. 2018. "Using instrumental variables for inference about policy relevant treatment parameters." *Econometrica* 86(5):1589–1619.

Richey, Jeremiah. 2016. "An odd couple: Monotone instrumental variables and binary treatments." *Econometric Reviews* 35(6):1099–1110.

Rockafellar, Ralph Tyrell. 1970. *Convex Analysis*. Princeton: Princeton University Press.
URL: *https://doi.org/10.1515/9781400873173*

Semenova, Vira. 2023. "Adaptive Estimation of Intersection Bounds: a Classification Approach.".

Shapiro, Alexander. 1990. "On concepts of directional differentiability." *Journal of optimization theory and applications* 66:477–487.

Siddique, Zahra. 2013. "Partially Identified Treatment Effects Under Imperfect Compliance: The Case of Domestic Violence." *Journal of the American Statistical Association* 108(502):504–513.
URL: *https://doi.org/10.1080/01621459.2013.779836*

Syrgkanis, Vasilis, Elie Tamer and Juba Ziani. 2021. "Inference on auctions with weak assumptions on information." *arXiv preprint arXiv:1710.03830* .

Tao, Terence and Van Vu. 2010. "Random matrices: The distribution of the smallest singular values." *Geometric And Functional Analysis* 20:260–297.

Van Der Vaart, Aad W, Jon A Wellner, Aad W van der Vaart and Jon A Wellner. 1996. *Weak convergence*. Springer.

Vytlacil, Edward. 2002. "Independence, monotonicity, and latent index models: An equivalence result." *Econometrica* 70(1):331–341.

Wainwright, Martin J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48 Cambridge university press.

Wright, Stephen J. 1997. *Primal-dual interior-point methods*. SIAM.

Yakusheva, Olga. 2010. "Return to college education revisited: Is relevance relevant?" *Economics of Education Review* 29(6):1125–1142.

6. Appendix

6.1. Proof of Lemma 1

Proof. To see why (7) holds, note that any minimum of the LP is feasible for the penalized unconstrained problem and it yields the same value $B(\theta)$.

Now consider the second part. It is a well-known result in the theory of penalty functions that if w in a linear penalty function is component-wise larger than the vector of Lagrange multipliers λ at a local optimum (subject to restrictions on the initial problem), then the local minimum corresponding to that λ is also a local minimum of the penalized unconstrained problem (e.g. Bertsekas (1975)). i) then follows from the fact that any local minimum of a convex program is also global.

Now consider the claim ii). Suppose that $(\overline{\lambda}, w)$ are the KKT vector and the penalty vector that satisfy Assumption A1 and $\overline{x}$ is the associated optimum of (4) and $\overline{B} \equiv p\,\overline{x}$. Note that one direction of ii) is trivial, since any $\tilde{x}$ that is optimal in the initial problem yields the same value in the penalized problem. For another direction, suppose x is a local (global) minimum of the penalized problem (6). If x is feasible, it is also an optimum of the initial problem. Suppose it is not feasible. By the assumption on $(w, \overline{\lambda})$:

$$p\,x + w\,(c - M\,x)^{+} > p\,x + \overline{\lambda}\,(c - M\,x) \tag{83}$$

The definition of a KKT vector in Rockafellar (1970) also requires that:

$$\overline{B} = \inf_{x \in \mathsf{R}^{N(S-1)}} p\,x + \overline{\lambda}\,(c - M\,x) \leq p\,x + \overline{\lambda}\,(c - M\,x) \tag{84}$$

Therefore,

$$\overline{B} = p\,x + w\,(c - M\,x) > p\,x + \overline{\lambda}\,(c - M\,x) \geq \overline{B} \tag{85}$$

Which yields a contradiction, so there can be no such x . Thus, the sets of optimal solutions coincide.

6.2. Proof of Theorem 1

Proof. The following Lemma is a well-known result, provided here for reference:

Lemma 8. For $\Theta \subseteq \mathsf{R}^n$ - compact, and for random sequences $f_n(\cdot), f(\cdot) : \Theta \to \mathsf{R}$, the following holds:

$$\sup_{\theta \in \Theta} |f_n(\theta) - f(\theta)| \xrightarrow{p} 0 \;=\; \sup_{\theta \in \Theta} |f_n^{+}(\theta) - f^{+}(\theta)| \xrightarrow{p} 0$$
$$\sup_{\theta \in \Theta} |f_n(\theta) - f(\theta)| \xrightarrow{p} 0 \;=\; \sup_{\theta \in \Theta} |f_n^{-}(\theta) - f^{-}(\theta)| \xrightarrow{p} 0$$

Proof. $|f_n^{+}(\theta) - f^{+}(\theta)| \leq |f_n(\theta) - f(\theta)|$

From Lemma 1 we have $B(\theta_0) = \tilde{B}(\theta_0; w)$, so proving consistency is equivalent to proving:

$$\tilde{B}_n \xrightarrow{p} \tilde{B}(\theta_0; w) \tag{86}$$

Now, it is straightforward to see that $L_n(x), L(x) \quad C[X]$. Thus, $|\hat{L}_n(x) - L(x)| \xrightarrow{p} 0$ $x \quad X$ by A0 and CMT, so there is pointwise convergence. To establish uniform convergence, first apply Cauchy-Schwarz and triangle inequality:

$$|L(x) - \hat{L}_n(x)| \leq ||\hat{p}_n - p|| \cdot ||x|| + ||w|| \cdot ||(\hat{c}_n - \hat{M}_n x)^+ - (c - Mx)^+|| \tag{87}$$

For the second term, dropping the positive value functions, Cauchy-Schwarz and triangle inequality yield:

$$||\hat{c}_n - \hat{M}_n x - (c - Mx)|| \leq ||\hat{c}_n - c|| + ||M - \hat{M}_n|| \cdot ||x|| \tag{88}$$

So that $||\hat{c}_n - \hat{M}_n x - (c - Mx)||_\infty \xrightarrow{p} 0$ by A0 and CMT. Applying Lemma 8, we get:

$$||(\hat{c}_n - \hat{M}_n x)^+ - (c - Mx)^+||_\infty \xrightarrow{p} 0 \tag{89}$$

Combining with (87) and noting that the first term converges uniformly:

$$||\hat{L}_n - L||_\infty \xrightarrow{p} 0 \tag{90}$$

It then immediately follows that:

$$|\tilde{B}_n - \tilde{B}(\theta_0; w)| = |\min_{x \ X}\{\hat{L}_n(x)\} - \min_{x \ X}\{L(x)\}| \xrightarrow{p} 0 \tag{91}$$

6.3. Proof of Lemma 2

Proof.

$$|\hat{Q}_n(x) - Q(x)| = \left| \sum_j \left([\hat{M}_n x - \hat{c}_n]_j^- \right)^2 - \left([Mx - c]_j^- \right)^2 \right| = \tag{92}$$

$$= \left| \sum_j ([\hat{M}_n x - \hat{c}_n]_j^- - [Mx - c]_j^-)([\hat{M}_n x - \hat{c}_n]_j^- + [Mx - c]_j^-) \right| \leq \tag{93}$$

$$\leq \sum_j |[\hat{M}_n x - \hat{c}_n]_j^- - [Mx - c]_j^-| \cdot |[\hat{M}_n x - \hat{c}_n]_j^- + [Mx - c]_j^-| \leq \tag{94}$$

$$\leq \left(\max_j [\hat{M}_n x - \hat{c}_n]_j^- + [Mx - c]_j^- \right) \sum_j \left| \left[(\hat{M}_n - M)x + c - \hat{c}_n \right]_j \right| \tag{95}$$

Where (95) uses the fact that $|(y_0)^- - (y_1)^-| = |\max\{0, -y_0\} - \max\{0, -y_1\}| \leq |y_0 - y_1| \quad y_0, y_1 \quad \mathbb{R}$. We now show that the last line converges to 0 is supremum over $x \quad X$. Note that, since $\hat{M}_n \xrightarrow{p} M$, $\hat{c}_n \xrightarrow{p} c$, the estimator asymptotically lies in any δ-vicinity of the true population parameter. In other words, $\delta > 0$, we have $(\hat{c}_n, vec(\hat{M}_n)) \quad B_\delta((c, vec(M)))$ w.p. 1 asymptotically.

Since X is a compact and because of the former result, both $\hat{M}_n x - \hat{c}_n$ and $Mx - c$

are bounded w.p. 1 asymptotically, so there exists $K > 0$ - large enough:[34]:

$$\sup_{x \in X} \max_j [\hat{M}_n x - \hat{c}_n]_j^- + [Mx - c]_j^- \leq K + o_p(1) \tag{96}$$

Note that by Cauchy-Schwarz, $\sum_j \left| \left[(\hat{M}_n - M)x + c - \hat{c}_n \right]_j \right| \leq |F_t| \cdot ||(\hat{M}_n - M)x + c - \hat{c}_n||$. Further using (95), (96) and noting that for nonnegative f, g one has $\sup_A fg \leq \sup_A f \cdot \sup_A g$, we get:

$$||\hat{Q}_n(x) - Q(x)||_\infty \leq (K + o_p) \cdot |F_t| \cdot \sup_{x \in X} ||(\hat{M}_n - M)x + c - \hat{c}_n|| \leq \tag{97}$$

$$\leq (\tilde{K} + o_p) \cdot \left(||\hat{M}_n - M|| \cdot ||x||_\infty + ||c - c_n|| \right) = o_p(1) \tag{98}$$

The proof is complete.

6.4. Proof of Proposition 5

First, notice that the iff result for the following conditions: i) SC holds and ii) $A(\theta_0)$ and $\Lambda(\theta_0)$ are both singletons follows from Theorem 3.1 in Fang and Santos (2018) combined with Lemma 4 and Proposition 4. Then, observe that $A(\theta_0)$ and $\Lambda(\theta_0)$ are both singletons if and only if both LICQ hold and there are no flat faces.

6.5. Inference for LP under SC and the biased penalty function estimator

6.5.a. LP under Slater's condition. We now consider inference for the original LP estimator under Slater's condition. Proposition 5 showed that, in the absence of this condition, the plug-in may fail to be consistent, because the value function is not continuous in the parameter θ_0[35].

Assumption B2 (Slater's condition). $Int(\Theta_I) =$

The following lemma establishes Hadamard directional differentiability of a linear program under Assumption B2.

There is no apparent reason to suppose that ii) should hold in practice, and therefore we do not endorse applying Proposition 9. Instead, it is intended to illustrate the difficulty with inference on general LP value functions even under the Slater's condition.

Corollary 1. Under assumption B2, Proposition 6 holds with $\kappa_n = 0$, i.e. the naive estimator is consistent.

One way to obtain a consistent estimator is to employ the procedure developed in Hong and Li (2015). Let:

$$\tilde{B}_n(Z_n) \equiv \frac{B(\hat{\theta}_n + \epsilon_n Z_n) - B(\hat{\theta}_n)}{\epsilon_n} \tag{99}$$

[34] In the cMIV setup all terms of $\hat{M}_n$, $\hat{c}_n$ are known to be bounded, so asymptotic arguments are not necessary. We consider a more general case here.

[35] Although Slater's condition is not necessary for duality in the case of LP, its failure allows for unboundedness of the dual solution set at θ_0, see Wright (1997). As shown in Meyer (1979), this will imply that the optimal value function is not continuous with respect to perturbations in c.

For $\epsilon_n \to 0$ with $r_n \epsilon_n \to \infty$, we have the following proposition:

Proposition 10. *If Assumptions B1, B2 hold, and the bootstrapped Z_n satisfies the measurability conditions in Hong and Li (2015):*

$$\sup_{f \ BL_1(\mathsf{R})} |\mathsf{E}[f(\tilde{B}_n(\mathsf{Z}_n))|\{X_i\}_{i=1}^n] - \mathsf{E}[f(B_{\theta_0}(\mathsf{G}_0))]| = o_p(1) \tag{100}$$

Assumption B2 is rather strong, and one may not be comfortable imposing it directly. This is especially true in cases where many inequality restrictions are involved, such as under cMIV-s, because one would be concerned that the defined system may be close to point-identification. An even more serious problem in practice is that, even if an open ball is contained in Θ_I at θ_0, the radius of that ball is not inconsequential in finite samples. A thinner identified set leads the bootstrap iterations of the N.D.M. to fail more often, as the constraint set turns empty at perturbed parameter values. Dropping the failed iterations introduces an unknown bias to the estimates, and so is not advised.

One potential solution would be to use the set-expansion estimator as in Section 4.2. Indeed, as long as the true system is feasible, expanding the set from the RHS renders the Slater's condition true, and the procedure described in this section becomes applicable. The bias of such expansion would be controlled as follows:

$$\min_{\Theta_I} p\,x - ||p|| d_H(\Theta_I, \tilde{\Theta}_I) \leq \min_{\tilde{\Theta}_I} p\,x \leq \min_{\Theta_I} p\,x \tag{101}$$

Moreover, by Lipschitz continuity of systems of linear inequalities, $d_H(\Theta_I, \tilde{\Theta}_I) \leq C|\kappa|$ for some $C > 0$ depending on θ_0, where the vector $\kappa > 0$ is the RHS-expansion.

This estimator, however, would still be problematic both because it is conservative even in terms of the convergence rate, and because it relies on an arbitrarily selected set expansion. Since a larger expansion leads to a more conservative lower bound, in applied work the researcher would be tempted to select the minimal value that ensures the bootstrap iterations do not fail. The statistical properties of that approach are unclear.

6.5.b. Inference for the biased penalty. We now consider the penalty function estimator $\tilde{B}(\cdot)$ defined in Section 4.1. The main difficulty when conducting inference for it consists of proving its Hadamard directional differentiability.

Observe that we can write $\tilde{B} \equiv \phi \circ \tilde{L}$, where $\tilde{L}(\theta) \equiv L(\cdot;\theta)$ is a functional $\tilde{L} : \mathsf{R}^S \to \ell^\infty(\mathcal{X})$, and $\phi : \ell^\infty(\mathcal{X}) \to \mathsf{R}$ is given by:

$$\phi(q) \equiv \inf_{x \ \mathcal{X}} q(x),$$

and where we equip $\ell^\infty(\mathcal{X})$ with the sup norm. By Lemma S.4.9 in the Online Appendix of Fang and Santos (2018), ϕ is Hadamard directionally differentiable. It is therefore tempting to apply the chain rule to find the derivative of $\tilde{B}$, which only requires that $\tilde{L}$ is H.d.d. However, in the spirit of the example from Hansen (2017), this is not the case. The following remark illustrates that issue.

Remark. $g(y)(x) \equiv (x+y)^+$ viewed as a map $g : \mathsf{R} \to \ell^\infty(A)$ for $x \quad A \equiv [-C; C]$ for

some $C > 0$ is **not** Hadamard directionally differentiable for any fixed $y \quad [-C/2; C/2]$:

$$\lim_{t_n \to 0^+, h_n \to h} \| \frac{(y + x + t_n h_n)^+ - (y + x)^+}{t_n} - f(h)(x) \|_\infty = 0$$

for any continuous $f(h)(x)$. To see that, note that the first term converges pointwise to $\mathbb{I}\{y + x = 0\} h^+ + \mathbb{I}\{y + x > 0\} h$. Suppose that $h < 0$ and consider: $x_n = -y - \frac{t_n}{2} h_n$, we have:

$$\left| \frac{(y + x_n + t_n h_n)^+ - (y + x_n)^+}{t_n} - \mathbb{I}\{y + x_n = 0\} h^+ + \mathbb{I}\{y + x_n > 0\} h \right| =$$
$$= o(1) - \frac{h}{2} = o(1)$$

In light of this finding, it should be almost surprising that $\tilde{B}(\cdot)$ is still Hadamard directionally differentiable, as we now demonstrate. Instead of using the chain rule, which is of course only a sufficient condition for differentiability, we notice that $\tilde{B}$ can be rewritten as a new linear program that has a non-empty interior of the constraint set[36].

Proposition 11. *The penalty function estimator, $\tilde{B}(\theta; w)$ is Hadamard directionally di erentiable in θ at θ_0 if either i) X is a polyhedron with $Int(X) =$, or ii) x $\tilde{A}(\theta_0; w) \cap Int(X)$. The H.d.d. is given by:*

$$\tilde{B}_{\theta_0}(h; w) = \inf_{x \ \tilde{A}(\theta_0; w)_\lambda} \sup_{\tilde{\lambda}(\theta_0; w)} h_p x + \sum_{j=1}^{2^q} \lambda_j \sum_{i \ \Pi_j} w_i (h_{c_i} - h_{M_i} x) \tag{102}$$

where $h = (h_p, h_{c_1}, \ldots, h_{c_q}, h_{M_1}, \ldots, h_{M_q})$ is the direction and an upper-hemicontinuous correspondence $\tilde{\lambda} : \mathbb{R}^S \to 2^{\overline{1,2^q}}$ is as defined in the proof.

Proof. Throughout this proof w is taken to be fixed, therefore some dependencies on it are omitted in notation for brevity. We proceed in four steps:

1. Notice that $L(x; \theta, w)$ is a convex piecewise-linear function and it has the following representation:

$$L(x; \theta, w) = \max_{j \ \overline{1,2^q}} \left\{ p\, x + \sum_{i \ \Pi_j} w_i (c_i - M_i x) \right\}, \tag{103}$$

where $\{\Pi_j\}_{j=1}^{2^q} = 2^{\overline{1,q}}$, so that Π_j for different j contain indices of all possible combinations of positive penalty term. At a given x these can be interpreted as the sets of violated constraints. Let $g_j(x, \theta) \equiv p\, x + \sum_{i \ \Pi_j} w_i (c_i - M_i x)$ for $j \quad \overline{1, 2^q}$.

The initial estimator can then be represented as:

$$\tilde{B}(\theta; w) = \min_{x \ X_j} \max_{\overline{1,2^q}} g_j(x, \theta) \tag{104}$$

[36] Clearly, this new LP is not equivalent to the original one point-by-point, as that would mean that the plug-in, $B(\cdot)$, is always H.d.d., contradicting Proposition 5.

2. Assumptions i) or ii) allow us to impose w.l.g. that the known compact set X is a fixed, non-empty and bounded polyhedron. To see that for ii), note that the program is convex and therefore the sets of local and global minima coincide. If there exists an interior local minimum, it means that expanding the constraint set does not change the value, and therefore we can set X to be some compact and non-empty polyhedron that contains the original set. Then, another representation of the considered problem follows:

$$\tilde{B}(\theta; w) = \min_{t,x} t \quad \text{s.t.:} \quad \begin{cases} t & [\underline{t}; \overline{t}] \\ x & X \\ g_j(x, \theta) \leq t, j & \overline{1, 2^q} \end{cases} \tag{105}$$

For some sufficiently wide $[\underline{t}, \overline{t}]$, given θ is close to θ_0 and such that $\tilde{B}(\theta_0; w)$ $(\underline{t}, \overline{t})$. This is justified because $\tilde{B}(\theta; w)$ is continuous in θ, as shown in the proof of Proposition 5.

3. Note that the constraint set of (105) is compact, non-empty at $\theta = \theta_0$ and, moreover, it contains an open set. To see that, consider some pair $x(\theta_0), t(\theta_0)$ from the argmin of the problem, where $x(\theta_0)$ $\tilde{A}(\theta_0; w)$ X and $t(\theta_0) \equiv \tilde{B}(\theta_0; w)$. Consider $\varepsilon \equiv \overline{t} - t(\theta_0)$ and take $t \equiv t(\theta_0) + \frac{\varepsilon}{2}$. Note that by definition $t(\theta_0) \geq \max_j g_j(x(\theta_0))$. By continuity of $g_j(x, \theta_0)$ in x for all j $\overline{1, 2^q}$, $\delta > 0$ such that $t \geq \max_j g_j(x)$ t $(t - \frac{\varepsilon}{4}; t + \frac{\varepsilon}{4})$, x $\mathsf{B}_\delta(x(\theta_0))$. By either i) or ii) $Int(X) =$ and as $x(\theta_0)$ X it follows that $Int(X) \cap \mathsf{B}_\delta(x(\theta_0))$ is non-empty. It is also open as an intersection of two open sets. Therefore, the open set $O \equiv (t - \frac{\varepsilon}{4}; t + \frac{\varepsilon}{4}) \times (\mathsf{B}_\delta(x(\theta_0)) \cap Int(X))$ is contained in the constraint set of the induced LP at θ_0. That is, the problem at θ_0 satisfies the Slater's condition and Lemma 6 applies.

4. Suppose $\check{\Lambda}(\theta_0)$ is the set of Lagrange multipliers of (105) at $\theta = \theta_0$, and $\tilde{\Lambda}(\theta_0)$ is its projection on the coordinates corresponding to the constraints of form $g_j(x; \theta_0) \leq t$ for all j $\overline{1, 2^q}$. A typical element of $\tilde{\Lambda}(\theta_0)$ will be written as $\lambda = (\lambda_j)_{j=1}^{2^q}$. Recall that for θ in some small open neighbourhood of θ_0 the value function of (105) is equal to $\tilde{B}(\theta; w)$ and, moreover, the problems are equivalent, so if $\check{A}(\theta)$ is the arg min of (105), then $\check{A}(\theta) = \{\tilde{B}(\theta; w)\} \times \tilde{A}(\theta; w)$. Using the conclusion of Step 3, direct application of Lemma 6 to (105) yields:

$$\tilde{B}_{\theta_0}(h; w) = \inf_{x \ A(\tilde{\theta_0}; w)\lambda} \sup_{\tilde{\Lambda}(\theta_0)} \sum_{j=1}^{2^q} \lambda_j \left(h_p x + \sum_{i \ \Pi_j} w_i (h_{c_i} - h_{M_i} x) \right), \tag{106}$$

where note that there are no terms corresponding to the objective function and the constraints t $[\underline{t}, \overline{t}]$ and x X, because there are no corresponding increments. Moreover, differentiating the Lagrangean of (105) and recalling that $t(\theta_0)$ $(\underline{t}; \overline{t})$, so the constraints t $[\underline{t}, \overline{t}]$ do not bind and the corresponding multipliers are 0, one gets that λ $\tilde{\Lambda}(\theta_0)$, we have $\sum_{j=1}^{2^q} \lambda_j = 1$, establishing (102).

Remark. By Lemma 2, Assumption A1 ensures ii) in Proposition 11 if Θ_I $Int(X)$.

Assuming A1 holds, exact pointwise inference is then obtained via Proposition 10. It is also straightforward to show that if A1 does not hold, but conditions i) or ii) in Proposition 11 are otherwise satisfied, this inference is asymptotically conservative.

Computational considerations may be important in practice, especially as bootstrap is involved. In Appendix we further show that the penalty function estimator may be computed as a value of a simple LP. If there are k constraints defining X and q constraints for Θ_I, with d variables, the penalty-induced LP will feature $d+q$ variables and $2q+k$ constraints, which makes it almost as simple computationally as the usual plug-in estimator with d variables and $q+k$ constraints.

6.6. Proof of Theorem 2

Proof. Fix $\theta = (p\,, vec(M)\,, c\,) = \theta_0$. We proceed in six steps, first proving the following lemma:

Lemma 9. Consider $B \equiv \arg\min_{x \in A} f(x)$ and $c \equiv f(x\ )$ for any $x \in B$, where $f(\cdot)$ is continuous and A is a non-empty compact. Then, for any measurable random sequence $\{x_n\} \in A$ such that $f(x_n) \xrightarrow{p} c$, there exists a measurable random sequence $\{x_n\} \in B$ such that $||x_n - x_n|| \xrightarrow{p} 0$.

Proof. Under the assumptions of the Lemma, Berge's maximum theorem implies that B is a non-empty compact. Because the distance is continuous, the projection x_n of x_n onto B is always well-defined for each n. If it is not unique, we select one of the values that yield the minimum distance. Measurability of at least one such selection is established by reference to Theorem 18.19 in Aliprantis and Border (2007). We then proceed by contradiction. Suppose that $\varepsilon > 0$:

$$\mathsf{P}[||x_n - x_n|| > \varepsilon] \to 0 \tag{107}$$

Then, there exists a $\delta > 0$ and a subsequence $\{n_k\}_{k=1}^{\infty}$ such that, for all $k \in \mathsf{N}$:

$$\mathsf{P}[||x_{n_k} - x_{n_k}|| > \varepsilon] > \delta \tag{108}$$

Consider the following problem:

$$\min_{x \in A,\ d(x,B) \geq \varepsilon} f(x) \tag{109}$$

Notice that the constraint set is compact. It is also non-empty, as for any k some of the realisations of x_{n_k} are in it by (108). Therefore the minimum is attained at some $\tilde{x}$. Suppose that the minimum is equal to $f(\tilde{x}) = \tilde{c}$. If $\tilde{c} = c$, it follows that $\tilde{x} \in B$, which is not possible as $d(\tilde{x}, B) \geq \varepsilon$. Clearly, $\tilde{c} < c$ is also infeasible as the constraint set of that problem is smaller than that of the original one. Therefore, $\tilde{c} - c = K > 0$. Then, note that for any $k \in \mathsf{N}$:

$$||x_{n_k} - x_{n_k}|| > \varepsilon \ \Rightarrow \ f(x_{n_k}) \geq f(\tilde{x}) = c + K > c \tag{110}$$

So,

$$\mathsf{P}[f(x_{n_k}) - f(x\ ) \geq K] \geq \mathsf{P}[||x_{n_k} - x_{n_k}|| > \varepsilon] > \delta > 0, \tag{111}$$

where the LHS goes to 0 as $k \to \infty$, since $f(x_{n_k}) \xrightarrow{p} f(x\ )$ by assumption of the Lemma. This yields a contradiction. Therefore, $||x_n - x_n|| \xrightarrow{p} 0$.

1. We first prove that $\{\delta_n\} \to 0^+$ such that $A(\hat{\theta}_n, w_n) \quad A(\theta_0)^{\delta_n}$ w.p. 1 asymptotically. For this purpose, recall that by Theorem 3 for any sequence $x_n \quad A(\hat{\theta}_n, w_n)$ for all n and for any $x \quad A(\theta_0)$, we have:

$$p\ x_n + w_n \iota\ (\hat{c}_n - \hat{M}_n x_n)^+ - p\ x\ \ = o_p(1) \tag{112}$$

Furthermore, since $w_n = o_p(\sqrt{n})$, we have:

$$w_n ||\hat{c}_n - \hat{M}_n x - c + Mx||_\infty = o_p(1) \tag{113}$$

Because the argmin is contained in a compact, $A(\hat{\theta}_n, w_n) \quad X$, the first term in (112) is bounded in probability: $p\ x_n = O_p(1)$, thus, from (112), it also follows that $w_n \iota (\hat{c}_n - \hat{M}_n x_n)^+ = O_p(1)$. By triangle inequality and using with (113), we therefore conclude:

$$w_n \iota\ (c - M x_n)^+ = O_p(1) \tag{114}$$

As $w_n \to \infty$, it further follows that:

$$(c - M x_n)^+ = o_p(1) \tag{115}$$

We shall now consider $\tilde{x}_n$ - a projection of x_n onto $\{x \quad \mathsf{R}^d | Mx \geq c\}$. Note that it exists, because distance is a continuous function and the set is a non-empty compact. Note that (115) implies that, for some random $\kappa_n \geq 0$ for all n:

$$c - M x_n \leq \iota \kappa_n \tag{116}$$

where $\kappa_n = o_p(1)$. We get:

$$||x_n - \tilde{x}_n|| = d(x_n, \{x \quad \mathsf{R}^d | Mx \geq c\}) \leq \tag{117}$$

$$\leq d_H(\{x \quad \mathsf{R}^d | Mx \geq c - \kappa_n\}, \{x \quad \mathsf{R}^d | Mx \geq c\}) \leq C\kappa_n, \tag{118}$$

where $C > 0$ is some fixed constant. The first equality is by definition of projection, the second inequality follows from the definition of the Hausdorff distance and (116) as well as:

$$d(x_n, \{x \quad \mathsf{R}^d | Mx \geq c\}) \leq \sup_{x \quad \{x \quad \mathsf{R}^d | Mx \geq c - \kappa_n\}} d(x, \{x \quad \mathsf{R}^d | Mx \geq c\})) \tag{119}$$

The final inequality is implied by Lipschitz-continuity of polyhedra in Hausdorff

distance with respect to RHS expansions (see Li (1993)). Therefore:

$$\tilde{x}_n - x_n \xrightarrow{p} 0 \tag{120}$$

We now wish to show that $p\ x_n \xrightarrow{p} p\ x$, where x is some value from $A(\theta_0)$. For arbitrary $\varepsilon > 0$ note that:

$$\mathsf{P}[|p\ x_n + w_n\iota\ (\hat{c}_n - \hat{M}_n x_n) - p\ x\ | > \varepsilon] \geq \tag{121}$$

$$\mathsf{P}[p\ x_n > p\ x\ + \varepsilon - w_n\iota\ (\hat{c}_n - \hat{M}_n x_n)] \geq \tag{122}$$

$$\mathsf{P}[p\ x_n > p\ x\ + \varepsilon] \tag{123}$$

As the LHS goes to 0 by (112), we have:

$$\mathsf{P}[p\ x_n > p\ x\ + \varepsilon] \to 0 \tag{124}$$

To prove the other side, note that, as $\tilde{x}_n \in \Theta_I(\theta_0)$, by definition of x , it must be that $p\ \tilde{x}_n \geq p\ x$. Therefore,

$$\mathsf{P}[p\ x_n < p\ x\ - \varepsilon] \leq \mathsf{P}[p\ x_n < p\ \tilde{x}_n - \varepsilon] \to 0, \tag{125}$$

where the RHS converges to 0 by (120) and CMT. We thus conclude that $p\ x_n \xrightarrow{p} p\ x$ and, moreover, $p\ \tilde{x}_n \xrightarrow{p} p\ x$.

Notice that by Lemma 2, for a fixed, large enough w satisfying Assumption A1 Lemma 9 applies, where one sets $f(x) = L(x; \theta_0, w)$, $B = A(\theta_0)$ with $f(x^*) = p\ x$ for any $x^* \in A(\theta_0)$. Thus, $\forall x_n \in A(\theta_0)$ such that $||x_n - x_n|| \xrightarrow{p} 0$. Therefore, $\forall \delta_n \to 0^+$ such that:

$$\mathsf{P}[||x_n - x_n|| < \delta_n] \to 1 \tag{126}$$

Recall that the sequence x_n was arbitrarily selected from $A(\hat{\theta}_n, w_n)$, and we can, for example, select a measurable $\{x_n\}_{n=1}^{\infty}$ (by Theorem 18.19 in Aliprantis and Border (2007)):

$$x_n \in \underset{x \in A(\hat{\theta}_n, w_n)}{\arg\max}\ d(x, A(\theta_0)) \tag{127}$$

For such x_n, we get:

$$||x_n - x_n|| < \delta_n \ = \ d(x, A(\theta_0)) < \delta_n \quad \forall x \in A(\hat{\theta}_n, w_n) \tag{128}$$

So:

$$\mathsf{P}[A(\hat{\theta}_n, w_n) \subseteq A(\theta_0)^{\delta_n}] \geq \mathsf{P}[||x_n - x_n|| < \delta_n] \to 1 \tag{129}$$

This establishes the existence of a deterministic $\delta_n \to 0^+$ such that $A(\hat{\theta}_n, w_n) \subseteq A(\theta_0)^{\delta_n}$ w.p. 1 as.

2. By (129) and using the representation found in Proposition 11 we have that:

$$\inf_{x \in X} L(x; \hat{\theta}_n, w_n) = \inf_{x \in A(\theta_0)^{\delta_n}} L(x; \hat{\theta}_n, w_n) + o_p(1) \tag{130}$$

$$= \min_{x \in A(\theta_0)^{\delta_n}} p\, x + w_n \max_{j \in \overline{1,2^q}} \left\{ \sum_{i \in \Pi_j} (\hat{c}_{ni} - \hat{M}_{ni} x) \right\} + o_p(1), \tag{131}$$

where $o_p(1)$ encompasses realizations at which $A(\hat{\theta}_n, w_n) \not\subset A(\theta_0)^{\delta_n}$ or where $\hat{\theta}_n$ is not in a fixed open vicinity of θ_0 that was argued to exist in Proposition 11. Suppose that at θ_0 the constraints that do not bind at any $x \in A(\theta_0)$ are given by $I \subseteq \{1, 2, \ldots, q\}$. By continuity, it follows that $\exists\, \delta > 0$ and $\varepsilon > 0$ such that:

$$c_i - M_i x < -\varepsilon, \quad i \in I \tag{132}$$

for any $x \in A(\theta_0)^{\delta}$. From (129) it then also follows that:

$$\inf_{x \in X} L(x; \hat{\theta}_n, w_n) = \min_{x \in A(\theta_0)^{\delta_n}} p\, x + w_n \max_{\Pi \in 2^{\overline{1,q}} \setminus I} \left\{ \sum_{i \in \Pi_j} (\hat{c}_{ni} - \hat{M}_{ni} x) \right\} + o_p(1) \tag{133}$$

3. Consider the problem in the linear programming representation found in Proposition 11, which it admits w. p. 1 as.:

$$\inf_{x \in X} L(x; \hat{\theta}_n; w_n) = \min_{t,x} t \quad \text{s.t.:} \quad \begin{cases} t \in [\underline{t}; \overline{t}] \\ x \in X \\ p\, x + \sum_{i \in \Pi_j} w_n (\hat{c}_{ni} - \hat{M}_{ni} x) \leq t, j \in \overline{1,2^q} \end{cases} \tag{134}$$

The Lagrangian reads as:

$$L = t + \sum_{\Pi \in 2^{\overline{1,q}}} \lambda_\Pi \left(p\, x - t + w_n \sum_{j \in \Pi} \hat{c}_{nj} - \hat{M}_{nj} x \right), \tag{135}$$

Where the constraints $x \in X$ and $t \in [\underline{t}; \overline{t}]$ are omitted, as they are not binding with probability 1 as. This holds, as $A(\theta_0) \subseteq Int(X)$ and $B(\theta_0) \subseteq Int([\underline{t}; \overline{t}])$ by assumption. Because $A(\theta_0)$ is compact, there further exists[37] a $\overline{\delta} > 0$: $A(\theta_0)^{\overline{\delta}} \subseteq Int(X)$ and as $\tilde{A}(\hat{\theta}_n; w_n) \subseteq A(\theta_0)^{\delta_n}$ w.p. 1 as. for some $\delta_n \to 0^+$, it follows that w.p. 1 as. $A(\hat{\theta}_n; w_n) \subseteq Int(X)$. Similar argument establishes that $t_n \in Int([\underline{t}; \overline{t}])$ w.p. 1 as. In what follows, we will simply call such optimal pairs *interior*.

[37] To see that, consider $A, B \subseteq \mathbb{R}^d$ such that A is compact, B is open and $A \subseteq B$. Since B is open, for any $b \in B$ $\exists\, \varepsilon > 0 : B_\varepsilon(b) \subseteq B$. This defines an open cover of A, as $A \subseteq \bigcup_{b \in B} B_{\varepsilon_b/2}(b)$. Since A is compact, for any cover there exists a finite subcover, i.e. $(b_k, \varepsilon_{b_k}/2)_{k=1}^K$ such that $b_k \in B$ and $A \subseteq \bigcup_{k=1}^K B_{\varepsilon_{b_k}/2}(b_k)$. Take $\delta = \min_k \varepsilon_{b_k}/2$. Then, pick any $x \in A^\delta$. It follows that $\forall\, y \in A$: $||x - y|| < \delta$. Because $y \in A$, there further $\exists\, k$: $||y - b_k|| \leq \varepsilon_{b_k}/2$. Thus, $||x - b_k|| \leq ||y - b_k|| + ||x - y|| < \varepsilon_{b_k}/2 + \delta \leq \varepsilon_{b_k}$, and so $x \in B_{\varepsilon_{b_k}}(b_k) \subseteq B$.

Differentiating with respect to t, one notes that:

$$\sum_{\Pi} \lambda_{\Pi} = 1 \tag{136}$$

Next, at any *interior* optimal t, x:

$$t = p\ x + w_n \max_{\Pi} \sum_{j\ \Pi} (\hat{c}_{nj} - \hat{M}_{nj} x) \tag{137}$$

To see that, note that by contradiction, if:

$$t > p\ x + w_n \max_{\Pi} \sum_{j\ \Pi} (\hat{c}_{nj} - \hat{M}_{nj} x) \tag{138}$$

Then, as we assumed that the pair (t, x) is *interior*, there exists $\tilde{t} < t$ such that the pair $(\tilde{t}, x)$ satisfies all the constraints. Therefore, (t, x) is not optimal. The other direction of the inequality is infeasible, and so the equality must hold. Moreover, since Π may be empty, we also have at any optimal x:

$$t \geq p\ x \tag{139}$$

Furthermore, the problem has a solution w.p. 1 as., and therefore it has a vertex-solution, i.e. a solution that is pinned down by a matrix of binding constraints of full column-rank. Because w.p. 1 as. any solution is *interior*, any such matrix w.p. 1 as. does not feature constraints $x \quad X, t \quad [\underline{t}, \overline{t}]$. The only constraints that can be satisfied at such vertex-solution with an equality are of the following type:

$$p\ x - t = w_n \sum_{j\ \Pi_k} \hat{c}_{nj} - \hat{M}_{nj} x,\ k \quad \tilde{J} \tag{140}$$

for some $\tilde{J} \quad 2^{\overline{1,q}} : |\tilde{J}| \geq d+1$, where the latter inequality holds by definition of a vertex of a linear program[38]. One can write the complete set of the binding constraints (140) as:

$$\hat{R}_{\tilde{J}n} \begin{pmatrix} t \\ x \end{pmatrix} = \hat{r}_n^{\tilde{J}}, \tag{141}$$

where the $|\tilde{J}| \times (d+1)$ matrix $\hat{R}_{\tilde{J}n}$ is of full column rank and the system yields a unique solution t_n, x_n.

4. Denote the set of all vertices $(t\ , x\ )$ that satisfy (140) with $|\tilde{J}| \geq d+1$ at a given $\hat{\theta}_n$ by $V\ (\hat{\theta}_n)$. From the previous arguments it follows that $V\ (\cdot)$ is non-empty w.p. 1 as. and finite, because any finite-dimensional polygon has finitely many vertices and therefore the corresponding LP has finitely many optimal vertices. We will write $V_x(\hat{\theta}_n)$ for the projection of that set on the x-coordinates. For any vertex-solution $(t\ , x\ ) \quad V\ (\hat{\theta}_n)$, suppose constraints $V \quad \{1, \dots q\}$ are violated at it, meaning

[38] Any finite feasible LP has a vertex-solution, at which the matrix of binding constraints has full rank, so that its dimension is at least that of $(t\ x')'$.

that:

$$V\ (\hat{\theta}_n, x\ ) \equiv \{j \quad \overline{1,q}|\hat{c}_{nj} - \hat{M}_{nj}x\ > 0\} \tag{142}$$

For brevity, we will write $V_n \equiv V\ (\hat{\theta}_n, x_n)$ where $t_n, x_n \quad V\ (\hat{\theta}_n)$ is some (measurable) sequence of optimal vertices. Note that:

$$t_n = p\ x_n + w_n \max_{\Pi} \sum_{j\ \Pi} (\hat{c}_{nj} - \hat{M}_{nj}x_n) = p\ x_n + w_n \sum_{j\ V_n} (\hat{c}_{nj} - \hat{M}_{nj}x_n) \tag{143}$$

Consider (140) and suppose $\tilde{J}_n = \tilde{J}(t_n, x_n)$ with $|\tilde{J}_n| \geq d+1$ is the set of the corresponding subsets, i.e.:

$$t_n = p\ x_n + w_n \sum_{j\ \Pi_i} (\hat{c}_{nj} - \hat{M}_{nj}x_n) \quad i \quad \overline{1,k} \tag{144}$$

It must be that $V_n \quad \Pi_i \quad i \quad \tilde{J}_n$, because $j\ /\ V_n\ = \quad (\hat{c}_{nj} - \hat{M}_{nj}x_n) \leq 0$, and so we have:

$$\sum_{j\ V_n} (\hat{c}_{nj} - \hat{M}_{nj}x_n) = \sum_{j\ \Pi_i} (\hat{c}_{nj} - \hat{M}_{nj}x_n) \leq \sum_{j\ \Pi_i \cap V_n} (\hat{c}_{nj} - \hat{M}_{nj}x_n), \tag{145}$$

where the first equality follows from (144) and (143). We now proceed by contradiction. Suppose that $j : j \quad V_n \cap \Pi_i$ (where the complement is taken with respect to $\overline{1,q}$), then:

$$\sum_{j\ \Pi_i \cap V_n} (\hat{c}_{nj} - \hat{M}_{nj}x_n) < \sum_{j\ \Pi_i \cap V_n} (\hat{c}_{nj} - \hat{M}_{nj}x_n) + \sum_{j\ \Pi_i \cap V_n} (\hat{c}_{nj} - \hat{M}_{nj}x_n) = \\ = \sum_{j\ V_n} (\hat{c}_{nj} - \hat{M}_{nj}x_n),$$

which yields a contradiction with (145), so there can be no such j. In light of (145) it then also follows that $i \quad \tilde{J}_n$ and $j \quad \Pi_i \cap V_n$ it must be that:

$$\hat{c}_{nj} - \hat{M}_{nj}x_n = 0 \quad j \quad \Pi_k \setminus V_n \tag{146}$$

Therefore, the complete system described by equation (144), is equivalent to:

$$\begin{cases} \hat{c}_{nj} - \hat{M}_{nj}x_n = 0 \quad i \quad \tilde{J}_n : \Pi_i = V_n\,, \quad j \quad \Pi_i \setminus V_n \\ t_n = p\ x_n + w_n \sum_{j\ V_n} \hat{c}_{nj} - \hat{M}_{nj}x_n \end{cases} \tag{147}$$

From the representation (141), we know that the matrix corresponding to system (147) must be of full column rank, $d+1$. Dropping the equation defining t_n, it implies that there exists at least d linearly independent equations of form:

$$\hat{c}_{nj} - \hat{M}_{nj}x_n = 0$$

We denote the set of all binding constraints by $\Pi\ (\hat{\theta}_n, x_n) \equiv \{j \quad \overline{1,q}|\hat{c}_{nj} - \hat{M}_{nj}x_n =$

$0\}$, which we shall occasionally write as Π_n for brevity. We thus have:

$$|\Pi_n| \geq d, \quad \mathrm{rk}(\hat{M}_{\Pi_n}) = d \tag{148}$$

5. Consider two collections of sets:

$$E \equiv \{A \quad 2^{[q]} : M_A x = c_A \quad x \quad A(\theta_0)\} \tag{149}$$

$$F \equiv \{A \quad 2^{[q]} : p \ / \ R(M_A)\} \tag{150}$$

We shall now consider two events E_n and F_n:

$$E_n \equiv \{\Pi_n \quad E\}, \quad F_n \equiv \{\Pi_n \quad F\} \tag{151}$$

$$\tag{152}$$

We wish to show that $\mathsf{P}[E_n] \to 0$ and $\mathsf{P}[F_n] \to 0$ and therefore $\mathsf{P}[E_n \cap F_n] \to 1$.

a) Let us consider E_n first. Since $A(\theta_0)$ is compact, for a fixed set $A \quad E$, the condition $M_A x = c_A \quad x \quad A(\theta_0)$ implies that there exists $\varepsilon(A) > 0$:

$$\inf_{x \ A(\theta_0)} ||M_A x - c_A|| > \varepsilon(A) \tag{153}$$

Because E is a finite collection of sets, we can pick $\varepsilon = \min_{A \ E} \varepsilon(A)$, so that:

$$\min_{A \ E} \inf_{x \ A(\theta_0)} ||M_A x - c_A|| > \varepsilon \tag{154}$$

By continuity of the objective function in x, there further $\kappa > 0$, such that:

$$\min_{A \ E} \inf_{x \ A^{\kappa}(\theta_0)} ||M_A x - c_A|| > \frac{\varepsilon}{2} \tag{155}$$

We now consider:

$$\mathsf{P}[E_n] \leq \mathsf{P}\left[||\hat{M}_{\Pi_n} x_n - \hat{c}_{\Pi_n}|| = 0, \inf_{x \ A^{\kappa}(\theta_0)} ||M_{\Pi_n} x - c_{\Pi_n}|| > \frac{\varepsilon}{2}\right] \tag{156}$$

Observe that for any non-empty $A \quad [q]$, by Cauchy-Schwartz and triangle inequalities:

$$\begin{aligned} ||(\hat{M}_{nA} x_n - \hat{c}_{nA})|| = \\ \left|\left|(M_A x_n - c_A) - \left((\hat{c}_{nA} - c_A) + (M_A - \hat{M}_{nA}) x_n\right)\right|\right| \geq \\ ||M_A x_n - c_A|| - \left|\left|\hat{M}_{nA} - M_A\right|\right| ||x||_{\infty} - ||\hat{c}_{nA} - c_A|| \end{aligned}$$

We can thus further rewrite:

$$\mathsf{P}\left[\|\hat{M}_{\Pi_n}x_n - \hat{c}_{\Pi_n}\| \leq 0, \inf_{x \ \mathcal{A}^\kappa(\theta_0)} \|M_{\Pi_n}x - c_{\Pi_n}\| > \frac{\varepsilon}{2}\right] \leq$$
$$\mathsf{P}\left[\left|\left|M_{\Pi_n}x_n - c_{\Pi_n}\right|\right| \leq \eta_n, \inf_{x \ \mathcal{A}^\kappa(\theta_0)} \|M_{\Pi_n}x - c_{\Pi_n}\| > \frac{\varepsilon}{2}\right],$$

where $\eta_n \equiv \left|\left|\hat{M}_{\Pi_n} - M_{\Pi_n}\right|\right| \|x\|_\infty + \left|\left|\hat{c}_{\Pi_n} - c_{\Pi_n}\right|\right| = o_p(1)$. Finally, using $\mathsf{P}[A \cap B\,] + \mathsf{P}[A \cap B] = \mathsf{P}[A]$:

$$\mathsf{P}\left[\left|\left|M_{\Pi_n}x_n - c_{\Pi_n}\right|\right| \leq \eta_n, \inf_{x \ \mathcal{A}^\kappa(\theta_0)} \|M_{\Pi_n}x - c_{\Pi_n}\| > \frac{\varepsilon}{2}\right] =$$
$$\mathsf{P}\left[\left|\left|M_{\Pi_n}x_n - c_{\Pi_n}\right|\right| \leq \eta_n, \inf_{x \ \mathcal{A}^\kappa(\theta_0)} \|M_{\Pi_n}x - c_{\Pi_n}\| > \frac{\varepsilon}{2}, x_n \quad \mathcal{A}^\kappa(\theta_0)\right] +$$
$$+\mathsf{P}\left[\left|\left|M_{\Pi_n}x_n - c_{\Pi_n}\right|\right| \leq \eta_n, \inf_{x \ \mathcal{A}^\kappa(\theta_0)} \|M_{\Pi_n}x - c_{\Pi_n}\| > \frac{\varepsilon}{2}, x_n \ / \ \mathcal{A}^\kappa(\theta_0)\right],$$

where the second term is $o(1)$ by Step 1 of the proof. Finally, we note that x_n $\mathcal{A}^\kappa(\theta_0) = \left|\left|M_{\Pi_n}x_n - c_{\Pi_n}\right|\right| > \varepsilon/2$, which, combined with $\left|\left|M_{\Pi_n}x_n - c_{\Pi_n}\right|\right| \leq \eta_n$, further implies that $\eta_n > \varepsilon/2 > 0$, so that:

$$\mathsf{P}\left[\left|\left|M_{\Pi_n}x_n - c_{\Pi_n}\right|\right| \leq \eta_n, \inf_{x \ \mathcal{A}^\kappa(\theta_0)} \|M_{\Pi_n}x - c_{\Pi_n}\| > \frac{\varepsilon}{2}, x_n \quad \mathcal{A}^\kappa(\theta_0)\right] \leq$$
$$\mathsf{P}\left[\frac{\varepsilon}{2} \leq \eta_n\right] = o(1)$$

This concludes the proof.

b) We now consider F_n. To do so, it is convenient to observe that the penalty function estimator and problem (134) are equivalent to yet another LP:

$$B(\hat{\theta}_n) + o_p(1/\sqrt{n}) = \min_{x,a} p\,x + w_n\iota\,a \quad \text{s.t.} : \begin{cases} a \geq 0 \\ a \geq \hat{c}_n - \hat{M}_n x \end{cases} \tag{157}$$

Note that we drop the constraints corresponding to $x \quad \mathcal{X}$ in (157), and $o_p(1/\sqrt{n})$ accommodates the potential non-existence of the interior solution. Write Lagrangian:

$$\mathcal{L} = p\,x + w_n\iota\,a + \mu\,(\hat{c}_n - \hat{M}_n x - a) - \omega\,a$$

The KKT conditions at an interior optimum are:

$$p = \hat{M}_n \mu \tag{158}$$
$$w_n = \omega + \mu \tag{159}$$
$$\omega\ a = 0 \tag{160}$$
$$\mu\ (\hat{c}_n - \hat{M}_n x - a) = 0 \tag{161}$$
$$a \geq \hat{c}_n - \hat{M}_n x \tag{162}$$
$$a \geq 0, \omega \geq 0, \mu \geq 0 \tag{163}$$

Analyzing the above system, one observes that if at x_n $V_x(\theta_n)$ a constraint is violated, j V_n, then $a_j > 0$, and so $\omega_j = 0$, which implies $\mu_j = w_n$. If $\hat{M}_{nj} x_n - \hat{c}_{nj} > 0$, then $\hat{c}_{nj} - \hat{M}_{nj} x_n - a_j < 0$, and so $\mu_j = 0$. Finally, if j Π_n, then μ_j $[0; w_n]$. Therefore, (158) rewrites as:

$$p = w_n \sum_{j\ V_n} \hat{M}_{nj} + \sum_{j\ \Pi_n} \hat{M}_{nj} \mu_j \tag{164}$$

Since $\mu_j \leq w_n$ and as $\hat{M}_n - M = O_p(1/\sqrt{n})$, we have:

$$p = w_n \sum_{j\ V_n} M_j + \sum_{j\ \Pi_n} M_j \mu_j + O_p(\frac{w_n}{\sqrt{n}}) \tag{165}$$

Consider a projection P_{Π_n} from R^d onto $R(M_{\Pi_n})$. For example, one can construct it as $M_{\Pi_n}(M_{\Pi_n})^\dagger$, where $\dagger$ denotes a Moore-Penrose pseudoinverse. We can write:

$$p - O_p(\frac{w_n}{\sqrt{n}}) = w_n(I - P_{\Pi_n}) \sum_{j\ V} M_j + \underbrace{w_n P_{\Pi_n} \sum_{j\ V} M_j + \sum_{j\ \Pi_n} M_j \mu_j}_{T_n \quad R(M_{\Pi_n})} \tag{166}$$

Notice that, if $\sum_{j\ V} M_j \not\in R(M_\Pi)$, then the RHS of (166) has unbounded norm:

$$\left|\left| w_n(I - P_{\Pi_n}) \sum_{j\ V_n} M_j + T_n \right|\right|^2 = \tag{167}$$
$$= w_n^2 ||(I - P_{\Pi_n}) \sum_{j\ V_n} M_j||^2 + ||T_n||^2$$

Since the square norm of the LHS of (166) is bounded from above by $||p||^2 + O_p(\frac{w_n^2}{n}) = ||p||^2 + o_p(1)$, (167) will contradict the equality in (166) w.p. 1 as. Suppose, alternatively, that $v : \sum_{j\ V_n} M_j = M_{\Pi_n} v$. Equation (166) rewrites:

$$p - O_p(\frac{w_n}{\sqrt{n}}) = M_{\Pi_n}(\mu_{\Pi_n} + w_n v),$$

which implies, for example, that:

$$(I - P_{\Pi_n})p + P_{\Pi_n}p - M_{\Pi_n}(\mu_{\Pi_n} + w_n v) = O_p(\frac{w_n}{\sqrt{n}}) \tag{168}$$

The norm of the LHS of (168) must go to 0, however, if $p \notin \mathcal{R}(M_{\Pi_n})$, we have, by orthogonality:

$$\left|\left|(I - P_{\Pi_n})p\right|\right|^2 + \left|\left|P_{\Pi_n}p - M_{\Pi_n}(\mu_{\Pi_n} + w_n v)\right|\right|^2 \geq \left|\left|(I - P_{\Pi_n})p\right|\right|^2 > 0,$$

which will also yield a contradiction w.p. 1 as. To complete the proof, one applies the same probabilistic arguments as used in step 5.a above, which we omit here. Thus, $\mathsf{P}[F_n] \to 0$.

6. We define the *correct set of vertices*, $\mathcal{G}$, as follows:

$$\mathcal{G} \equiv \{A \in [q] : x \in \mathcal{A}(\theta_0) \text{ s.t. } M_A x = c_A,\ p \in \mathcal{R}(M_A)\}$$

In line with previous notation, let $G_n \equiv \{\Pi_n \in \mathcal{G}\}$. The results of point 5 imply that $\mathsf{P}[E_n \cap F_n] = \mathsf{P}[G_n] \to 1$.

Consider any $A \in \mathcal{G}$. Suppose $p = M_A v$ for some $v \in \mathsf{R}^{|A|}$. Further, fix any $x \in \mathcal{A}(\theta_0) : M_A x = c_A$, then:

$$B(\theta_0) = p\ x = v\ M_A x = v\ c_A \tag{169}$$

The conclusion then follows from the following chain of equalities:

$$\begin{aligned} G_n \ &= \ p\ x_n - B(\theta_0) = v\ M_{\Pi_n} x_n - v\ c_{\Pi_n} = & (170)\\ &= v\ \hat{M}_{\Pi_n} x_n - v\ c_{\Pi_n} + v\ (M_{\Pi_n} - \hat{M}_{\Pi_n}) x_n = & (171)\\ &= v\ (\hat{c}_{\Pi_n} - c_{\Pi_n}) + v\ (M_{\Pi_n} - \hat{M}_{\Pi_n}) x_n & (172) \end{aligned}$$

Finally, from (172), applying the triangle and Cauchy-Shwartz inequalities as well as noting that over the event G_n one has $\Pi_n \in \mathcal{G}$ by definition, it follows that:

$$\begin{aligned} G_n \ &= \ |p\ x_n - B(\theta_0)| \leq \varpi_n \equiv & (173)\\ \max_{A \in \mathcal{G}} &\left\{ \left(||\hat{c}_A - c_A|| + ||x||_\infty ||M_A - \hat{M}_A|| \right) \cdot \min_{v \in \mathsf{R}^{|A|}:\ M_A v = p} ||v|| \right\} \end{aligned}$$

One concludes by noting that the RHS is clearly $O_p(1/\sqrt{n})$, as $\mathcal{G}$ is finite and $\hat{\theta}_n - \theta_0 = O_p(1/\sqrt{n})$ by assumption. Formally, for any $\varepsilon > 0$:

$$\begin{aligned} \mathsf{P}[r_n|p\ x_n - B(\theta_0)| > \varepsilon] &= \mathsf{P}[r_n|p\ x_n - B(\theta_0)| > \varepsilon, G_n] + o(1) \leq & (174)\\ &\mathsf{P}[r_n \varpi_n > \varepsilon, G_n] + o(1) \leq \mathsf{P}[r_n \varpi_n > \varepsilon] + o(1) & (175) \end{aligned}$$

and $r_n \varpi_n = O_p(\frac{y_n}{\sqrt{n}})$ for any $r_n \to \infty$, where we used the fact that $\mathsf{P}[G_n \cap O_n] \leq \mathsf{P}[G_n] = o(1)$ for any measurable O_n. Recalling that the choice of $x_n \in \mathcal{V}_x(\hat{\theta}_n)$ was

arbitrary and that neither ϖ_n, nor the $o(1)$ depend on x_n, one gets:

$$\sup_{x \in V_x(\hat{\theta}_n)} |p\ x - B(\theta_0)| = O_p(1/\sqrt{n}) \tag{176}$$

But because any $x \in A(\hat{\theta}_n; w_n)$ can be represented as a convex combination of vertices, $\{x_j\}_{j=1}^K \in V_x(\hat{\theta}_n)$, as: $x = \sum_j \omega_j x_j$, where $\omega_j \in [0;1]$ and $\sum_j \omega_j = 1$. Using that, applying the triangle inequality and taking maximum, one gets, for any $x \in A(\hat{\theta}_n; w_n)$:

$$|p\ x - B(\theta_0)| = \left| \sum_j \omega_j (p\ x_j - B(\theta_0) \right| \leq$$
$$\max_j |p\ x_j - B(\theta_0)| \leq \sup_{x \in V_x(\hat{\theta}_n)} |p\ x - B(\theta_0)| = O_p\left(1/\sqrt{n}\right)$$

taking supremum on the left hand side establishes the claim of the theorem.

6.7. Proof of Lemma 6

Proof. With σ computed as follows:

$$\text{Var}\left(\check{v}\ C(\hat{A})\left(C_c Z - \text{vec}^{-1}_{q\times d}(C_M Z)\hat{x}\right)\right) = \tag{177}$$
$$= \text{Var}\left(\check{v}\ C(\hat{A})C_c Z\right) - 2\text{Cov}\left(\check{v}\ C(\hat{A})C_c Z, \check{v}\ C(\hat{A})\text{vec}^{-1}_{q\times d}(C_M Z)\hat{x}\right) + \tag{178}$$
$$+\text{Var}\left(\check{v}\ C(\hat{A})\text{vec}^{-1}_{q\times d}(C_M Z)\hat{x}\right) \tag{179}$$

where $Z \sim N(0, \Sigma)$ has the asymptotic distribution of $Z_n^{(2)}$. Let:

$$J_1 \equiv \check{v}\ C(\hat{A})C_c \tag{180}$$
$$J_2 \equiv \check{v}\ C(\hat{A})(\text{vec}(I_d) \otimes I_q) \tag{181}$$

The first term clearly rewrites as:

$$\text{Var}\left(\check{v}\ C(\hat{A})C_c Z\right) = J_1 \Sigma J_1 \tag{182}$$

To deal with the last term, rewrite:

$$\text{Var}\left(\check{v}\ C(\hat{A})\text{vec}^{-1}_{q\times d}(C_M Z)\hat{x}\right) = J_2 \text{Var}\left((I_d \otimes C_M Z)\ \hat{x}\right) J_2 \tag{183}$$

Direct computation yields:

$$(I_d \otimes C_M Z)\hat{x} = \begin{pmatrix} C_M Z \hat{x}_1 \\ C_M Z \hat{x}_2 \\ \cdots \\ C_M Z \hat{x}_d \end{pmatrix} \tag{184}$$

So:

$$\text{Var}\left((I_d \quad C_M Z)\,\hat{x}\right) = \hat{x}\hat{x} \quad C_M \Sigma C_M \tag{185}$$

Consider:

$$\text{Cov}\left(\check{v}\; C(\hat{A}) C_c Z, \check{v}\; C(\hat{A}) \text{vec}_{q\times d}^{-1}(C_M Z)\hat{x}\right) = \mathsf{E}[J_1 Z J_2 (I_d \quad C_M Z)\hat{x}] = \tag{186}$$

$$= J_2 \mathsf{E}[(I_d \quad C_M Z Z\; J_1)]\hat{x} = J_2 (I_d \quad C_M \Sigma J_1)\hat{x} \tag{187}$$

Combining everything, we get:

$$\sigma(\hat{A}, \hat{x}, \hat{v}, \Sigma) = J_1 \Sigma J_1 \; - 2 J_2 (I_d \quad C_M \Sigma J_1)\hat{x} + J_2 \left(\hat{x}\hat{x} \quad C_M \Sigma C_M\right) J_2 \tag{188}$$

We thus have, for fixed $\hat{A}, \hat{v}, \hat{x}$ with $\hat{v} = 0$.

6.8. Proof of Theorem 3

Proof. We begin by taking the infeasible $\hat{\sigma}_n(A, v, x) = \sigma(A, v, x, \Sigma)$. Note that:

$$\mathsf{P}[H_n \leq z_{1-\alpha} | \mathcal{D}_n^{(1)}] = \mathsf{P}[H_n \leq z_{1-\alpha} | \hat{A}, \hat{v}, \hat{x}] \tag{189}$$

Because the data in $\mathcal{D}_n^{(1)}$ is independent from $\mathcal{D}_n^{(2)}$ and all dependencies of H_n on $\mathcal{D}_n^{(1)}$ can be described as measurable functions of $\hat{A}, \hat{v}, \hat{x}$. Consider the set:

$$(\underline{v}, \underline{\sigma}) \equiv \{(A, v, x) \quad 2^{\overline{1,q}} \setminus \{\;\} \times \mathsf{R}^q \times \mathcal{X} : \underline{v} \leq ||v_n|| \leq \overline{v}, \sigma(A, v, x, \Sigma) \geq \underline{\sigma}\} \tag{190}$$

We now fix an aribtrary deterministic sequence $(A_n, v_n, x_n) \quad (\underline{v}, \underline{\sigma})$ for some small $\underline{v} > 0$ and $\underline{\sigma} > 0$ for all $n \quad \mathsf{N}$. Consider the limit (integration is with respect to $\mathcal{D}_n^2$ only):

$$\lim_{n\to\infty} \mathsf{P}[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}] \tag{191}$$

The space $2^{\overline{1,q}} \setminus \{\;\}$, to which A_n belongs, is endowed with a discrete metric, and we consider the space $(\underline{v}, \underline{\sigma})$ as endowed with the induced product metric ρ_p. It is straightforward to notice that $\sigma(\cdot)$ is continuous in its first three arguments with respect to ρ_p even on the unrestricted space $2^{\overline{1,q}} \setminus \{\;\} \times \mathsf{R}^q \times \mathcal{X}$, and thus $(\underline{v}, \underline{\sigma})$ is a compact space for any $\underline{v} > 0, \underline{\sigma} > 0$. It is also non-empty for some small enough $\underline{v} > 0, \underline{\sigma} > 0$ by Assumption B4. Suppose $\underline{v} > 0, \underline{\sigma} > 0$ are small enough and pick any convergent subsequence $(A_{n_k}, v_{n_k}, x_{n_k}) \to (A, v, x)$. Recall that:

$$H_n(A_n, v_n, x_n) = g(\sqrt{n_2}(\hat{\theta}^{(2)} - \theta_0), A_n, v_n, x_n) \tag{192}$$

for a continuous function g and:

$$\begin{pmatrix} \sqrt{(n_2)_k}(\hat{\theta}^{(2)} - \theta_0) \\ A_{n_k} \\ v_{n_k} \\ x_{n_k} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{N}(0, \Sigma) \\ A \\ v \\ x \end{pmatrix} \tag{193}$$

we conclude that, by continuous mapping theorem, as $k \to \infty$:

$$g(\sqrt{(n_2)_k}(\hat{\theta}^{(2)}_{n_k} - \theta_0), A_{n_k}, v_{n_k}, x_{n_k}) = H_{n_k}(A_{n_k}, v_{n_k}, x_{n_k}) \xrightarrow{d} g(Z, A, v, x), \tag{194}$$

where $Z \quad N(0, \Sigma)$. By (18), this implies:

$$\lim_{k \to \infty} P[H_{n_k}(A_{n_k}, v_{n_k}, x_{n_k}) \leq z_{1-\alpha}] = 1 - \alpha \tag{195}$$

We claim that this further implies that:

$$\lim_{n \to \infty} P[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}] = 1 - \alpha \tag{196}$$

Suppose, by contradiction, $\lim_{n \to \infty} P[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}] = 1 - \alpha$. It means that $\varepsilon > 0$ such that $N \quad N \quad n \geq N$ such that:

$$|P[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}] - (1 - \alpha)| > \varepsilon \tag{197}$$

Thus, we can construct a subsequence n_k such that:

$$|P[H_{n_k}(A_{n_k}, v_{n_k}, x_{n_k}) \leq z_{1-\alpha}] - (1 - \alpha)| > \varepsilon \tag{198}$$

for all $k \quad N$. Noting that $A_{n_k}, v_{n_k}, x_{n_k}$ still belongs to a compact metric space, we can find a further subsequence n_{k_j} such that $A_{n_{k_j}}, v_{n_{k_j}}, x_{n_{k_j}}$ is convergent. But for this subsequence our previous result, (195), yields that:

$$P[H_{n_{k_j}}(A_{n_{k_j}}, v_{n_{k_j}}, x_{n_{k_j}}) \to (1 - \alpha), \tag{199}$$

which yields a contradiction. Thus, for any (A_n, v_n, x_n) satisfying $x_n \quad X$, $\underline{v} < ||v_n|| \leq \overline{v}$ and $\sigma(A_n, v_n, x_n, \Sigma) \geq \underline{\sigma}$ for all $n \quad N$:

$$\lim_{n \to \infty} P[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}] = 1 - \alpha \tag{200}$$

We can therefore pick an arbitrarily small $\underline{\sigma}$ and $\underline{v}$ and consider event:

$$E_n \equiv \{\sigma(\hat{A}, \hat{v}, \hat{x}, \Sigma) < \overline{\sigma}\} \quad \{||\hat{v}|| < \underline{v}\} \tag{201}$$

It is straightforward to see that continuity of $\sigma(\cdot)$ with respect to the first three arguments (considered in the product metric), Assumption B4 and the fact that $\hat{A}, \hat{x}, \hat{v}$ converge in probability to a random sequence in the set of optimal triplets by the previous results imply that $P[E_n] \to 0$. Observe that:

$$\mathbb{1}_{E_n} \inf_{A,v,x \quad (\underline{v},\underline{\sigma})} P[H_n(A, v, x) \leq z_{1-\alpha}] \leq P[H_n \leq z_{1-\alpha} | \hat{A}, \hat{v}, \hat{x}] \leq \tag{202}$$

$$\leq \sup_{A,v,x \quad (\underline{v},\underline{\sigma})} P[H_n(A, v, x) \leq z_{1-\alpha}] + \mathbb{1}_{E_n} \tag{203}$$

It follows that:

$$\lim_{n \to \infty} \mathsf{P}[H_n \leq z_{1-\alpha} | \hat{A}, \hat{v}, \hat{x}] = 1 - \alpha + o_p(1) \tag{204}$$

Therefore:

$$\mathsf{P}[H_n \leq z_{1-\alpha} | \hat{A}, \hat{v}, \hat{x}] = 1 - \alpha + o_p(1) \tag{205}$$

From where, integrating over $\boldsymbol{D}_n^{(1)}$, one concludes that:

$$\mathsf{P}[H_n \leq z_{1-\alpha}] = 1 - \alpha + o(1) \tag{206}$$

Now, note that:

$$G_n = O_p\left(\frac{1}{\sqrt{n}}\right) \tag{207}$$

From where it follows that, for any $\varepsilon > 0$:

$$o(1) + \mathsf{P}[H_n \leq z_{1-\alpha} - \varepsilon] \leq \mathsf{P}[H_n - G_n \leq z_{1-\alpha}] \leq \mathsf{P}[H_n \leq z_{1-\alpha} + \varepsilon] + o(1) \tag{208}$$

Letting $\alpha^+(\varepsilon) \equiv 1 - \Phi(z_{1-\alpha} - \varepsilon)$ and $\alpha^-(\varepsilon) \equiv 1 - \Phi(z_{1-\alpha} + \varepsilon)$, applying (206), one obtains:

$$o(1) + 1 - \alpha^+(\varepsilon) \leq \mathsf{P}[H_n - G_n \leq z_{1-\alpha}] \leq o(1) + 1 - \alpha^-(\varepsilon) \tag{209}$$

Taking $\varepsilon \to 0$ and using continuity of the normal's cdf, we obtain:

$$\mathsf{P}[H_n - G_n \leq z_{1-\alpha}] = 1 - \alpha + o(1) \tag{210}$$

6.9. Proof of Lemma 7

Proof. Let $\delta > 0$ be a jump at P_0. Construct a sequence $\{\mathsf{P}_n\} \quad P$ such that for some $0 < \vartheta < 1$:

$$||\mathsf{P}_0 - \mathsf{P}_n||_{TV} < \vartheta n^{-1} \tag{211}$$

While $||V(\mathsf{P}_0) - V(\mathsf{P}_n)|| > \delta$. Recall that:

$$||\mathsf{P}_0^n - \mathsf{P}_n^n||_{TV} \leq n||\mathsf{P}_0 - \mathsf{P}_n||_{TV} \tag{212}$$

It follows that:

$$||\mathsf{P}_0^n - \mathsf{P}_n^n||_{TV} \leq \vartheta \tag{213}$$

Using the binary Le Cam's method[39], one obtains n:

$$\inf_{\hat{V}_n} \sup_{\mathsf{P} \in P} \mathsf{E}_{\mathsf{P}}[||V(\mathsf{P}) - \hat{V}_n(X(\mathsf{P}^n))||] \geq \frac{\delta(1-\vartheta)}{2} \tag{214}$$

Recalling that $0 < \vartheta < 1$ and δ were chosen arbitrarily and taking supremum over δ as well as sending $\vartheta \to 0$ yields the result.

6.10. Proof of Theorem 6

Proof. We proceed in three steps.
i) Notice that any $w_1 > ||p||\delta^{-1}$ satisfies assumption A1 for a given $\mathsf{P} \in P$, if the δ-condition holds for P. Therefore, $B(\theta(\mathsf{P})) = \tilde{B}(\theta(\mathsf{P}), w_1)$ for this P.
ii) Using the same arguments as in the proof of Theorem 1:

$$|\tilde{B}(\hat{\theta}_n; w_1) - \tilde{B}(\theta(\mathsf{P}); w_1)| \leq ||x||_\infty (||\hat{p}_n - p|| + w_1||\hat{M}_n - M||) + w_1||\hat{c}_n - c|| \tag{215}$$

iii) Using i and ii, we have $\mathsf{P} \in P$:

$$|\tilde{B}(\hat{\theta}_n; w_n) - \tilde{B}(\theta(\mathsf{P}); w_n)| \leq \tag{216}$$

$$\underbrace{\mathbb{I}\{||p|| - ||p_n|| < \delta\zeta\} \left[||x||_\infty \left(||\hat{p}_n - p|| + w_n||\hat{M}_n - M||\right) + w_n||\hat{c}_n - c||\right)\right]}_{\eta_n} + \gamma_n, \tag{217}$$

where $\gamma_n = 0$ if $||p|| - ||p_n|| < \delta\zeta$. Using triangle inequality and the properties of supremum, we get that $\mathsf{P} \in P$:

$$\sup_{m \geq n} |\tilde{B}(\hat{\theta}_m, w_m) - B(\theta(P))| \leq \sup_{m \geq n} |\gamma_m| + \sup_{m \geq n} |\eta_m| \tag{218}$$

Therefore, using the union bound and the properties of supremum:

$$\sup_{\mathsf{P} \in P} \mathsf{P}[\sup_{m \geq n} ||\hat{\theta}_m - \theta(\mathsf{P})|| \geq \varepsilon] \leq \sup_{\mathsf{P} \in P} \mathsf{P}[\sup_{m \geq n} |\gamma_m| > \frac{\varepsilon}{2}] + \sup_{\mathsf{P} \in P} \mathsf{P}[\sup_{m \geq n} |\eta_m| > \frac{\varepsilon}{2}] \tag{219}$$

We now note that:

$$\mathsf{P}[\sup_{m \geq n} |\gamma_m| > \frac{\varepsilon}{2}] \leq \mathsf{P}[\sup_{m \geq n} ||p|| - ||p_m|| > \delta\zeta] \leq \mathsf{P}[\sup_{m \geq n} ||p - p_m|| > \delta\zeta], \tag{220}$$

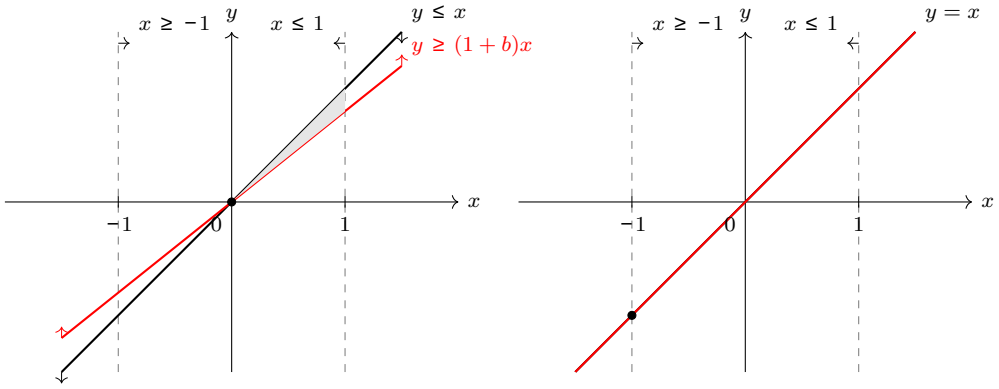
where the latter probability goes to 0 uniformly over P by a.s. uniform consistency of $\hat{\theta}_n$. Using the same arguments and employing the boundedness of $||\hat{p}_n||$ over P, one shows that the first term in (220) also goes to 0 uniformly over P.

6.11. Penalty parameter selection

To develop an intuition for the tradeoff involved in selecting $\delta > 0$ and therefore the w_n penalizing sequence in Theorem 6, let us return to the example in Proposition 5:

In this case, the smallest singular value at the binding constraint for $b < 0$ is simply $|b|$. Therefore, as $b \to 0^-$, the underlying measure belongs to a progressively smaller-δ

[39] See Chapter 15 of Wainwright (2019)



(a) $b < 0,\ B(b) = 0$ (b) $b = 0,\ B(b) = -1$

Figure 6: $B(b) = \min_{x,y} x \quad \text{s.t.} : y \geq (1+b)x,\ y \leq x,\ x \quad [-1; 1]$

set. For a given sample size, a higher w_n is then required to appropriately penalize the deviations, because the population Lagrange multiplier that needs to be dominated by it is equal to $-1/b$ (see A1 and Lemma 2). On the other hand, if the true measure is the one on the right, i.e. $b \geq 0$, the Lagrange multiplier that needs to be dominated by w_n is fixed at $\frac{-1+\sqrt{5}}{2}$.

An arbitrarily large w_n will perform well in case the identified set has a 'sharp angle' ($b \approx 0^-$). However, if $b = 0$, for example, in 50% of the cases the sample identified set will look like Figure 2.a), delivering the exploding sample Lagrange multiplier $\frac{-1}{b_n}$. If it happens to be dominated by w_n in the sense of A1, the incorrect minimum at 0, which is selected by the plug-in, is also picked by the penalty function estimator.

The aim of this section is to develop a prescription for the selection of a reasonable δ parameter that balances finite sample performance of the estimator with sufficient robustness. As a starting point, let us note that the scale of δ clearly depends on the scale of the singular values of M_J matrices. Any reasonable prescription for δ parameter selection should then first normalize the constraint matrix M. More precisely, we suggest that the constraint matrix first be normalized row-wise, setting the norm of each row to 1. We further suggest rescaling it by s, where:

$$s^2 \equiv \frac{1}{Qd-1}\sum_{i,j}(\hat{M}_{ij} - \overline{\hat{M}_{..}})^2, \tag{221}$$

Once the singular values of our matrix are thus normalized, δ may be interpreted as the degree of irregularity of the sharp identified set that one is willing to allow at the optimal solution. While uniformly and consistently estimating the sufficient δ is infeasible (because otherwise the uniformly consistent estimator would exist), we attempt to formulate a notion of what values of δ are *regular*. One possibility is to imagine that the population matrix of binding constraints M_J , in turn, is generated by some prior over the space of all measures. In particular, we can think of a prior such that each entry of M_J is a normalized mean zero variable, independent from other entries (but not necessarily identically distributed). In terms of the lower bound on the singular value, this prior turns

out to be rather *conservative*, because it can be shown that $\sigma_d(M_J)$ goes to 0 at the rate $\sqrt{d}$. We therefore view this as a prudent way to characterize the irregularity of a given matrix. The random matrix theory provides the following version of the 'Central Limit Theorem' for this general prior:

Theorem 11 (Tao and Vu (2010))**.** Let Ξ_n be a sequence of $n \times n$ matrices with $[\Xi_n]_{ij} \quad \xi_{ij}$, independently across i, j where ξ_{ij} are such that $\mathsf{E}[\xi] = 0$, $Var(\xi) = 1$ and $\mathsf{E}[|\xi|^{C_0}] < \infty$ for some sufficiently large C_0, then:

$$\sqrt{n}\sigma_n(\Xi_n) \xrightarrow{d} \Sigma \tag{222}$$

with the cdf of Σ given by:

$$\mathsf{P}[\Sigma \leq t] = 1 - e^{-t/2 - \sqrt{t}} \tag{223}$$

Remark. The distribution of mean-zero normalized ξ_{ij} in Theorem 6 is arbitrary, possibly discrete, and not necessarily identical.

This gives us the benchmark of what is 'reasonable' for a singular value of a $d \times d$ matrix. We suggest selecting the $0 < \alpha < 1$ quantile of this distribution, so that:

$$w_n = ||\hat{p}_n|| \delta^{-1} d_n \tag{224}$$

$$\delta = \frac{\left(\sqrt{1 - 2\ln(1-\alpha)} - 1\right)^2}{\sqrt{d}} \tag{225}$$

Where $d_n \to \infty$ is some sequence that diverges slowly enough, as in Theorem 5. For example, one could set $d_n = \ln \ln n / \ln \ln 100$ and $\alpha = 0.15$, seeing as the prior we selected appears rather 'conservative'. In our simulations of the example in Proposition 5, this choice of parameters delivers good uniform performance of the penalty function estimator, see Figure 3.

6.12. Proof of Theorem 9

Proof. We first show that:

$$\Theta \quad = \{\beta \quad \mathsf{R}| \quad x \quad \Theta_I : \beta = p\, x + \overline{p}\, \overline{x}\} \tag{226}$$

Fix $x \quad \Theta_I$. It follows that the quantity $m = P_m x + \overline{P}_m \overline{x}$ satisfies (55) by construction. To see that there exists at least one $P \quad \mathsf{P}$ that supports this m by generating $m(P) = m$, consider P under which the marginal distribution $F_{T,Z}(\cdot)$ is as observed, and the potential outcomes have the form:

$$Y(t) = \mathsf{I}\{t \quad O\}(\mathsf{I}\{T = t\} f(t, T, Z) + \mathsf{I}\{T = t\}\eta(t)) + \mathsf{I}\{t \quad U\} f(t, T, Z), \tag{227}$$

where $f : \mathsf{T}^2 \times \mathsf{Z} \to \mathsf{R}$ is a deterministic function with $f(t, d, z)$ that maps to the component of x corresponding to the conditional moment indexed by t, d, z: $\mathsf{E}[Y(t)|T = d, Z = z]$ if this moment is counterfactual and to 0 otherwise. $\eta(t)$ is some variable such that it aligns

with $Y(t)$ across the observed dimension: $F_{\eta(t)|T=t,Z=z}(y) = F_{Y(t)|T=t,Z=z}(y)$, y R and t O, z Z. By construction, this DGP generates $m(P) = m$ and delivers the required identified distribution across observed dimensions, $F_{Y|T=t,Z}(\cdot)$ for t O. Therefore:

$$x \quad \Theta_I = \quad \mu \quad (P_m x + \overline{P}_m \overline{x}) = p\, x + \overline{p}\, \overline{x} \quad \Theta \tag{228}$$

The other direction holds by construction: β Θ x $\Theta_I : p\, x + \overline{p}\, \overline{x} = \beta$.
The claim of the theorem is then established by showing that the identified set is indeed an interval, a ray, or the whole line. This follows, since if β_0, β_1 Θ with $\beta_0 < \beta_1$, then x_0, x_1 Θ_I such that $\beta_i = p\, x_i + \overline{p}\, \overline{x}$ for $i = 0, 1$. Because Θ_I is convex, for arbitrary β $[\beta_0, \beta_1]$ setting $\alpha = \frac{\beta_1 - \beta}{\beta_1 - \beta_0}$, one obtains $\alpha x_0 + (1 - \alpha) x_1$ Θ_I = β Θ .

6.13. Proof of Theorem 10

Lemma 10. Fix K_0, μ_v, μ_w, K_1 R: $K_0 \leq \mu_v \leq \mu_w \leq K_1$ and $F_w(\cdot)$ that is a valid c.d.f. with expectation μ_w. Suppose the probability space (P, Ω, S) can support a $U[0;1]$ random variable, and $P[W \leq w] = F_w(w)$. Then, there exists a random variable V s.t. $K_0 \leq V \leq W \leq K_1$ a.s. and $E[V] = \mu_v$.

Proof. Proof. Suppose $\mu_w > K_0$ as otherwise the statement is trivial. W can be represented as:

$$W = F_w^{-1}(U) \tag{229}$$

Where $F_w^{-1}(t) \equiv \inf\{w : F_w(w) \geq t\}$ is a generalized inverse. Consider a CDF $G(x) \equiv \mathbb{I}\{x \geq K_0\}$ on $[K_0; K_1]$. Notice that by definition:

$$\int x dG(x) = K_0 \tag{230}$$

Moreover, by linearity of the Lebesgue integral α $[0;1]$ we have:

$$\int x d(\alpha G(x) + (1 - \alpha) F_w(x)) = \alpha K_0 + (1 - \alpha)\mu_w \tag{231}$$

Let $F_v(x) \equiv \alpha\ G(x) + (1 - \alpha\ ) F_w(x)$ where α $\equiv \frac{\mu_w - \mu_v}{\mu_w - K_0}$. Then, notice that:

$$V = F_v^{-1}(U) \tag{232}$$

Yields the required random variable.

To prove the inverse inclusion in (62) for some $\tilde{M}, \tilde{b}$, note that from Theorem 1:

$$\{\beta \quad \mathbb{R} |\ P \quad P : \beta = \mu \quad m(P) \quad b \quad + M \quad m(P) \geq 0\} = \{\beta \quad \mathbb{R} | x : Mx \geq b : \beta = p\, x + \overline{p}\, \overline{x}\} \tag{233}$$

Where:

$$\overline{p} \equiv \overline{P}_m \mu \ , \quad p \equiv P_m \mu \tag{234}$$

$$M \equiv M \quad P_m \quad b \equiv -b \quad - M \quad \overline{P}_m \overline{x} \tag{235}$$

Therefore proving the inclusion consists in finding such data-consistent Y (or, equivalently, the measure P P) for any given $x : Mx \geq b$ that it generates $m(P) = p\, x + \overline{p}\, \overline{x}$ with $M \quad m(P) + b \quad \geq 0$ and $\tilde{M}\mathsf{Y} \geq \tilde{b}$ P - a.s.

1) Bounds For any $x : Mx \geq b$ we can once again construct the d.g.p. P from the Proof of Theorem 1:

$$Y(t) = \mathsf{I}\{t \quad O\}(\mathsf{I}\{T = t\} f(t, T, Z) + \mathsf{I}\{T = t\}\eta(t)) + \mathsf{I}\{t \quad U\} f(t, T, Z), \tag{236}$$

Where $f(t, d, z), \eta(t)$ are defined as in the proof of Theorem 1 and the distribution of T, Z is as observed. Clearly, $b \quad + M \quad m(P) \geq 0$ and $P \quad \mathsf{P}$ for this P holds by construction, and: $Y(t) \quad [K_0; K_1] \quad t \quad T$ a.s., therefore $\tilde{M}\mathsf{Y} \geq \tilde{b}$ a.s. by construction.

2) MTR In this case it is clear that (236) fails, because it does not necessarily satisfy monotonicity almost surely. Consider:

$$\begin{aligned} \mathsf{Y} = (\mathsf{I}\{t \quad O\}(\mathsf{I}\{T = t\} f(t, T, Z) + \mathsf{I}\{T = t\}\eta(t)) + \mathsf{I}\{t \quad U\} f(t, T, Z))_{t \ T} + \\ + \sum_{t \ O} (\iota_{N_T} - e_t)\mathsf{I}\{T = t\}(\eta(t) - \mathsf{E}[Y(t)|T = t, Z]) \end{aligned} \tag{237}$$

Where e_t is the standard basis vector with 1 in the position of the potential outcome corresponding to t in Y. Notice that the process in (237) has the same conditional means as the deterministic process of form (236), and therefore the corresponding $m(P)$ satisfies $M \quad m(P) + b \quad \geq 0$. Furthermore, by construction of $M \quad$ it must be that $\quad t \quad O$ and $\quad d \quad T : d = t$, we have:

$$\mathsf{E}[Y(d)|T = t, Z] = f(d, t, Z) \leq \mathsf{E}[Y(t)|T = t, Z] \text{ iff } d < t \tag{238}$$

and for $d_0, d_1 \quad T \setminus \{t\} : d_0 < d_1$:

$$\mathsf{E}[Y(d_0)|T = t, Z] = f(d_0, t, Z) \leq f(d_1, t, Z) = \mathsf{E}[Y(d_1)|T = t, Z] \tag{239}$$

Consider Y constructed in (237) over some element of the partition of Ω induced by T, where $T = t$.

i) If $t \quad U$, it is simply:

$$\mathsf{Y} = \begin{pmatrix} f(1, t, Z) \\ f(2, t, Z) \\ \dots \\ f(N_T, t, Z) \end{pmatrix} \tag{240}$$

Which satisfies $\tilde{M}\mathsf{Y} + \tilde{b} \geq 0$ over this element of the partition a.s., by construction of f.

ii) If $t \quad O$:

$$\mathsf{Y} = \begin{pmatrix} f(1,t,z) + \eta(t) - \mathsf{E}[Y(t)|T=t,Z] \\ \cdots \\ f(t-1,t,z) + \eta(t) - \mathsf{E}[Y(t)|T=t,Z] \\ \mathsf{E}[Y(t)|T=t,Z] + \eta(t) - \mathsf{E}[Y(t)|T=t,Z] \\ f(t+1,t,z) + \eta(t) - \mathsf{E}[Y(t)|T=t,Z] \\ \cdots \\ f(N_T,t,z) + \eta(t) - \mathsf{E}[Y(t)|T=t,Z] \end{pmatrix} \tag{241}$$

Notice that by (238) and (239) the MTR is then satisfied, i.e. $\tilde{M}\mathsf{Y} + \tilde{b} \geq 0$.

3) MTR + Bounds It is clear that the process given in (237) does not necessarily satisfy boundedness. We therefore resort to a different constructive argument. Consider the element of the partition wrt to T corresponding to $T = t$. For $t \quad U$ we can again set Y as in (240). Because each $f(d,t,Z)$ satisfies MTR and boundedness by construction, we have $\tilde{M}\mathsf{Y} + \tilde{b} \geq 0$ over this element of the T-partition.
Suppose $t \quad O$. The solution of the linear programming results in some moments that are given by our map $f(d,t,Z)$ that satisfies (238) and (239). Observe that constructing Y over the considered element of partition consists in constructing the counterfactual $Y(d)$ s.t. $d \quad T : d = t$ such that:

$$\mathsf{E}[Y(d)|T=t,Z] = f(d,t,Z) \quad d \quad T \setminus \{t\} \tag{242}$$

$$Y(1) \leq Y(2) \leq \cdots \leq Y(t) \leq \cdots \leq Y(N_T) \ a.s. \tag{243}$$

Where the distribution of $Y(t)$ over this element of the partition is identified. Repeated application of Lemma 7 yields this result. To construct the variables on the left, one starts from $Y(t-1)$, invokes Lemma 7 to construct it given the cdf of $Y(t)$ (which is identified over this element of the partition), and proceeds to use the obtained cdf to construct $Y(t-2)$, etc., descending to $Y(1)$. For the variables 'above' $Y(t)$, the Lemma is simply applied with the negative sign. All of the variables can be constructed using the same U random variable in the proof of Lemma 7, which yields that there exists a probability space such that (242)-(243) hold jointly a.s. This concludes the proof of the Theorem.

6.14. Failure of the converse inclusion for almost sure inequalities

Consider a binary treatment $T \quad \{0,1\}$ and suppose we estimate the sharp lower bound for $\mathsf{E}[Y(1)|T=0]$. Suppose that conditional on $T=0$, $Y(0)$ is 1 and -1 with equal probability. Assume that there is the only conditional restriction that $\mathsf{E}[Y(1)|T=0] \geq 0$. Further suppose that there is an almost sure restriction:

$$\begin{pmatrix} 1 & 1 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} Y(0) \\ Y(1) \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{244}$$

Note that this restriction defines the lower bound on $Y(1)$ of 2 if $Y(0) = 1$ and 1 if $Y(0) = -1$, and thus $\mathsf{E}[Y(1)|T=0] \geq 1.5$. Taking the expectation of this system

conditional on $T = 0$, however, yields that $\mathsf{E}[Y(1)|T = 0] = 0$ is a solution. Therefore, although 0 is a lower bound, it is not sharp.

6.15. Identification under cMIV

Sharp identification results for cMIV conditions follow from Theorem 2. cMIV-w, however, allows for a more explicit characterization of the bounds, which may better illustrate the source of the identifying power of cMIV-w relative to MIV. This characterization is also useful in binary settings, when cMIV assumptions coincide. For didactic purposes, in this section we also show how to construct the restriction matrix M and vector b under cMIV-s and cMIV-p. While we focus on bounding potential outcomes or ATEs, other choices of β can be accommodated by applying Theorem 2.

In what follows, I_k stands for the identity matrix of dimension k, and ι_k is the vector of ones of size k. These subscripts may be dropped in what follows without further notice. All vectors are column vectors, and $\mathsf{R}^{n \times m}$ refers to the space of real-valued $n \times m$ matrices. Notice that we can consider each $t \in \mathcal{T}$ separately, because cMIV conditions do not impose any restrictions across potential outcomes.

6.15.a. Recursive bounds under cMIV-w. Construct the ordering on the support of Z: $\mathcal{Z} = \{z_1, z_2, \ldots, z_{N_Z}\}$, s.t. $z_i < z_j$ for $i < j$. Denote by $l_i(t)$, $u_i(t)$ the sharp lower and upper bounds for the conditional moment over the whole treatment support, $\mathsf{E}[Y(t)|Z = z_i]$. Similarly, let $l_i^{-t}(t)$, $u_i^{-t}(t)$ be the sharp upper and lower bounds for the counterfactual subset, $\mathsf{E}[Y(t)|T = t, Z = z_i]$. We shall suppress the dependence on t whenever it does not cause confusion.

The only bound of interest is the bound on unconditional expectation, l_i. However, it turns out to be instructive to also consider the bound for the counterfactual subset, l_i^{-t}.

Proposition 12. *If i) cMIV-w holds or ii) treatment is binary and cMIV-s or cMIV-p hold, the sharp bounds for* $\mathsf{E}[Y(t)|Z=z_j]$ *are obtained through the following recursion for* $j \geq 2$*:*

$$l_j = l_{j-1} + \Delta_j \tag{245}$$

$$l_j^{-t} = l_{j-1}^{-t} + \Delta_j^{-t} \tag{246}$$

Where $\Delta_j, \Delta_j^{-t} \geq 0$ *are defined as follows:*

$$\Delta_j \equiv \left(\underbrace{\frac{\Delta P[T=t|Z=z_j]}{P[T=t|Z=z_{j-1}]} \left(l_{j-1} - P[T=t|Z=z_{j-1}]\mathsf{E}[Y(t)|T=t,Z=z_{j-1}]\right)}_{\Delta P[T=t|Z=z_j] l_{j-1}^{-t}} + \delta_j \right)^+ \tag{247}$$

$$\Delta_j^{-t} \equiv \frac{1}{P[T=t|Z=z_j]} \left(-\Delta P[T=t|Z=z_j] l_{j-1}^{-t} - \delta_j\right)^+ \tag{248}$$

$$\delta_j \equiv \Delta(P[T=t|Z=z_j]\mathsf{E}[Y(t)|T=t,Z=z_j]) \tag{249}$$

Sharp upper bounds u_i, u_i^{-t} *are obtained analogously. Moreover,*

$$\sum_{i=1}^{N} P[Z=z_i] l_i(t) \leq \mathsf{E}[Y(t)] \leq \sum_{i=1}^{N} P[Z=z_i] u_i(t) \tag{250}$$

In the absence of additional information, these bounds are sharp.

Proof. Note that $l_1^{-t} = K_0$ and $u_N^{-t} = K_1$. Moreover, $l_1 = \mathsf{P}[T=t|Z=z_1]\mathsf{E}[Y(t)|T=t,Z=z_1] + \mathsf{P}[T=t|Z=z_1]K_0$, $u_N = \mathsf{P}[T=t|Z=z_N]\mathsf{E}[Y(t)|T=t,Z=z_N] + \mathsf{P}[T=t|Z=z_N]K_1$. First, we note that the equations above may be rearranged to yield:

$$l_j^{-t} = max\left\{ \frac{1}{P[T=t|Z=z_j]} \left(l_{j-1} - \mathsf{E}[Y(t)|T=t,Z=z_j]P[T=t|Z=z_j]\right),\ l_{j-1}^{-t} \right\} \tag{251}$$

$$l_j = \mathsf{E}[Y(t)|T=t,Z=z_j]P[T=t|Z=z_j] + l_j^{-t}P[T=t|Z=z_j] \tag{252}$$

We consider the sharp lower bounds and proceed by induction on j. The proof for the sharp upper bounds is identical.
Consider $j=2$. The only information about lower bounds provided by assumption cMIV-w at $j=2$ is[40]:

$$\begin{cases} \mathsf{E}[Y(t)|Z=z_2] \geq \mathsf{E}[Y(t)|Z=z_1] \\ \mathsf{E}[Y(t)|T=t,Z=z_2] \geq \mathsf{E}[Y(t)|T=t,Z=z_1] \end{cases}$$

[40] Note that we can ignore the information that $Y(t) \geq K_0$, as it will be implied by the bound l_1^{-t} and l_1

Which can be rewritten as a single condition on $\mathsf{E}[Y(t)|T=t,Z=z_2]$:

$$\mathsf{E}[Y(t)|T=t,Z=z_2] \geq max\Big\{\mathsf{E}[Y(t)|T=t,Z=z_1],$$
$$P[T=t|Z=z_2]^{-1}\big(\mathsf{E}[Y(t)|Z=z_1] - P[T=t|Z=z_2]\mathsf{E}[Y(t)|T=t,Z=z_2]\big)\Big\}$$

Because l_1^{-t} is a sharp lower bound on $\mathsf{E}[Y(t)|T=t,Z=z_1]$, we get:

$$l_2^{-t} = max\Big\{l_1^{-t},\ P[T=t|Z=z_2]^{-1}\big(l_1 - P[T=t|Z=z_2]\mathsf{E}[Y(t)|T=t,Z=z_2]\big)\Big\}$$
$$l_2 = P[T=t|Z=z_2]\mathsf{E}[Y(t)|T=t,Z=z_2] + P[T=t|Z=z_2]l_2^{-t}$$

The base is thus proven. Now suppose that for some $j \geq 2$, and sharp lower bounds for $i<j$ are defined. The information we have at j is:

$$\begin{cases}\mathsf{E}[Y(t)|Z=z_j] \geq \mathsf{E}[Y(t)|Z=z],\ z<z_j\\ \mathsf{E}[Y(t)|T=t,Z=z_j] \geq \mathsf{E}[Y(t)|T=t,Z=z],\ z<z_j\end{cases}$$

Or, equivalently,

$$\mathsf{E}[Y(t)|T=t,Z=z_j] \geq max\Big\{\max_{i<j}\{\mathsf{E}[Y(t)|T=t,Z=z_i]\},$$
$$P[T=t|Z=z_j]^{-1}\big(\max_{i<j}\{\mathsf{E}[Y(t)|Z=z_i]\} - P[T=t|Z=z_j]\mathsf{E}[Y(t)|T=t,Z=z_j]\big)\Big\}$$

Because l_i, l_i^{-t} are sharp and non-decreasing in i by inductive hypothesis, it follows that sharp lower bounds at j are given by:

$$l_j^{-t} = max\Big\{l_{j-1}^{-t}, P[T=t|Z=z_j]^{-1}\big(l_{j-1} - P[T=t|Z=z_j]\mathsf{E}[Y(t)|T=t,Z=z_j]\big)\Big\}$$
$$l_j = \mathsf{E}[Y(t)|T=t,Z=z_j]P[T=t|Z=z_j] + l_j^{-t}P[T=t|Z=z_j]$$

The characterization in the proposition is obtained by rearranging these two equations.

To see that these bounds are indeed sharp, consider a process, for which $\mathsf{E}[Y(t)|T=d,Z=z_j] = l_j^{-t}, d=t, j \quad \overline{1,N}$. For such process cMIV-w will hold by construction and l_j and l_j^{-t} are both attained for all j. An example of such process is given by:

$$Y(w) = \sum_t \mathsf{I}\{t=w\}\left(\sum_j\left\{\mathsf{I}\{Z=z_j,T=t\}\,\eta(t) + \sum_{d=t}\mathsf{I}\{Z=z_j,T=d\}\,l_j^{-t}\right\}\right) \tag{253}$$

Where $\eta(t)$ is as defined in the proof of Theorem 1.

The intuition for Proposition 2 is that MIV bounds are obtained by 'ironing' the bounds on the population moment $\mathsf{E}[Y(t)|Z=z]$, which can be seen in equation (247). cMIV-w additionally 'irons' the counterfactual moments $\mathsf{E}[Y(t)|T=t,Z=z]$, as evident from (248). Figure 1 plots the derived sharp bounds as well as the benchmark MIV sharp bounds for a simulation exercise.

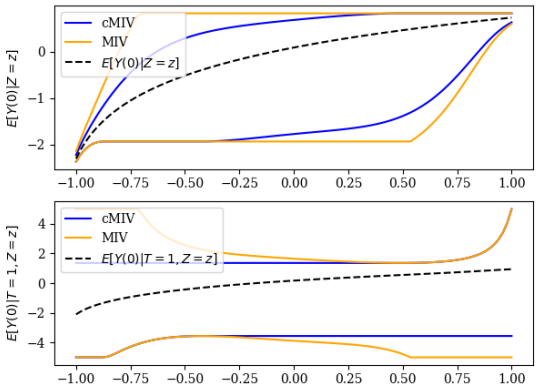


Figure 7: Bounds for the d.g.p. in Appendix 6.17.

6.15.b. Constructing M and b for cMIV-s and cMIV-p. Bounds given in Proposition 2 are not necessarily sharp under cMIV-s. Intuitively, cMIV-s allows us to 'iron' more moments than cMIV-w. cMIV-p, however, does not imply nor is implied by cMIV-w, so the bounds under the two conditions can compare arbitrarily. To characterize the sharp bounds under cMIV-s and cMIV-p, it is useful to introduce some notation first.

Let $\mathcal{F} \equiv 2^{\mathcal{T}} \setminus \{\{t\}, \emptyset\}$ and its cardinality, $Q \equiv |\mathcal{F}| = 2^{N_{\mathcal{T}}} - 2$. Fix an ordering on $\mathcal{F}$, so that $\mathcal{F} = \{A^1, A^2, \dots A^Q\}$.

Then all information under cMIV-s can be written as:

$$\mathsf{E}[Y(t)|T \quad A^k, Z = z_j] \geq \mathsf{E}[Y(t)|T \quad A^k, Z = z_{j-1}], \ k = 1, \dots, Q, \ j = 2, \dots N_Z \tag{254}$$

$$\mathsf{E}[Y(t)|T = d, Z = z_N] \leq K_1, d \quad \mathcal{T} \setminus \{t\} \tag{255}$$

$$\mathsf{E}[Y(t)|T = d, Z = z_1] \geq K_0, d \quad \mathcal{T} \setminus \{t\} \tag{256}$$

Where notice that the LHS of (255) is the largest marginal moment due to monotonicity in Z, while the LHS of (256) is the smallest marginal moment. Therefore, once almost sure bounds for these two moments are imposed $d \quad \mathcal{T} \setminus \{t\}$, these are also implied for all other moments through equation (254) and the law of total probability.

We now rewrite the expectations in (254) in terms of pointwise conditional moments. Let the vector of unobserved treatment responses be $x^j \equiv (\mathsf{E}[Y(t)|T = d, Z = z_j])_{d=t}$ and $p^j \equiv (P[T = d|Z = z_j])_{d=t}$ be the vector of respective probabilities at $Z = z_j$. Denote the element of x^j corresponding to $T = d$ as $x^j_d = \mathsf{E}[Y(t)|T = d, Z = z_j]$.

For $k = 1, \ldots, Q$ and $j = 2, \ldots, N_Z$, we can rewrite inequality (254) as follows:

$$\sum_{d=t} \mathsf{I}\{d \quad A^k\} \frac{P[T = d|Z = z_j]}{P[T \quad A^k|Z = z_j]} x_d^j +$$
$$+\mathsf{I}\{t \quad A^k\} \frac{P[T = t|Z = z_j]}{P[T \quad A^k|Z = z_j]} \mathsf{E}[Y(t)|T = t, Z = z_j] \geq$$
$$\geq \sum_{d=t} \mathsf{I}\{d \quad A^k\} \frac{P[T = d|Z = z_{j-1}]}{P[T \quad A^k|Z = z_{j-1}]} x_d^{j-1} +$$
$$+\mathsf{I}\{t \quad A^k\} \frac{P[T = t|Z = z_{j-1}]}{P[T \quad A^k|Z = z_{j-1}]} \mathsf{E}[Y(t)|T = t, Z = z_{j-1}]$$

Inequalities (255)-(256) are just $x_d^N \leq K_1, d = t$ and $x_d^1 \geq K_0, d = t$. This can be written succinctly in matrix notation. Introdude the following:

$$G_j \equiv \left(\mathsf{I}\{d \quad A^k\} \frac{P[T = d|Z = z_j]}{P[T \quad A^k|Z = z_j]} \right)_{k \ \overline{1,Q}, d=t} \quad \mathsf{R}^{Q \times N_T - 1} \tag{257}$$

$$c_j \equiv \left(\mathsf{I}\{t \quad A^k\} \frac{P[T = t|Z = z_j]}{P[T \quad A^k|Z = z_j]} \mathsf{E}[Y(t)|T = t, Z = z_j] \right)_{k \ \overline{1,Q}} \quad \mathsf{R}^Q \tag{258}$$

The whole set of information given by cMIV-s can be represented as follows:

$$G_j x^j - G_{j-1} x^{j-1} \geq -\Delta c_j, j = 2, \ldots, N_Z \tag{259}$$
$$x^N \leq K_1 \iota \tag{260}$$
$$x^1 \geq K_0 \iota \tag{261}$$

The procedure for cMIV-p is similar. First, we note that all the information under it is given by:

$$\mathsf{E}[Y(t)|Z = z_j] \geq \mathsf{E}[Y(t)|Z = z_{j-1}], \ j = 2, \ldots N_Z \tag{262}$$
$$\mathsf{E}[Y(t)|T = d, Z = z_j] \geq \mathsf{E}[Y(t)|T = d, Z = z_{j-1}], \ d \quad T \setminus \{t\}, \ j = 2, \ldots N_Z \tag{263}$$
$$\mathsf{E}[Y(t)|T = d, Z = z_N] \leq K_1, d \quad T \setminus \{t\} \tag{264}$$
$$\mathsf{E}[Y(t)|T = d, Z = z_1] \geq K_0, d \quad T \setminus \{t\} \tag{265}$$

Where (262) is just MIV and (263) is the monotonicity of the pointwise conditional moments. In this case, we can once again represent all information in the matrix form (259)-(261) with the following matrices:

$$G_j \equiv \begin{pmatrix} p^j \\ I_{N_T - 1} \end{pmatrix} \quad \mathsf{R}^{N_T \times N_T - 1} \tag{266}$$

$$c_j \equiv \begin{pmatrix} P[T = t|Z = z_j] \mathsf{E}[Y(t)|T = t, Z = z_j] \\ 0_{N_T - 1} \end{pmatrix} \quad \mathsf{R}^{N_T - 1} \tag{267}$$

Corollary 2. Under cMIV-s and cMIV-p, sharp bounds on $\mathsf{E}[Y(t)]$ take the form:

$$\min_{Mx \geq c} \left\{ \sum_{j=1}^{N} P[Z = z_j] \cdot p^j \; x^j \right\} + \sum_{j=1}^{N} P[T = t, Z = z_j] \mathsf{E}[Y(t) | T = t, Z = z_j] \leq \mathsf{E}[Y(t)] \leq$$
$$\leq \max_{Mx \geq c} \left\{ \sum_{j=1}^{N} P[Z = z_j] \cdot p^j \; x^j \right\} + \sum_{j=1}^{N} P[T = t, Z = z_j] \mathsf{E}[Y(t) | T = t, Z = z_j],$$

where:

$$M \equiv \begin{bmatrix} -I_{N_T - 1} & \dots & 0 & 0 \\ G_N & -G_{N-1} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & G_2 & -G_1 \\ 0 & \dots & 0 & I_{N_T - 1} \end{bmatrix}, \quad c \equiv \begin{pmatrix} -K_1 \cdot \iota_{N_T - 1} \\ -\Delta c_N \\ \vdots \\ -\Delta c_2 \\ K_0 \cdot \iota_{N_T - 1} \end{pmatrix}, \quad x = \begin{pmatrix} x^N \\ \vdots \\ x^1 \end{pmatrix}, \tag{268}$$

and G_j and c_j are given by (257) and (258) for cMIV-s and by (266) and (267) for cMIV-p respectively.

6.16. Proof of Proposition 8

Let $\Gamma(z) \equiv \sum_{d \; T} P[T = d | Z = z] \mathsf{E}[\psi(z, \eta) | Z = z]$.

a) Let $\tilde{g}(t) \equiv \mathsf{E}[g(t, \xi) | T = d, Z = z] = \mathsf{E}[g(t, \xi) | Z = z]$, where we use independence of ξ and T, Z.
MIV implies:

$$\mathsf{E}[Y(t) | Z = z] = \tilde{g}(t) + h(t)\Gamma(z) \;\; - \;\; increasing \tag{269}$$

Since inequality is strict for some z, z , it follows that $h(t) = 0$ and $h(t)/h(d) > 0$. Note that:

$$\mathsf{E}[Y(t) | T = d, Z = z] - \tilde{g}(t) = \frac{h(t)}{h(d)} \left(\mathsf{E}[Y(d) | T = d, Z = z] - \tilde{g}(d) \right) \tag{270}$$

Therefore, cMIV-p holds iff all observed moments are monotone.

b) Let $\tilde{g}(t, d) \equiv \mathsf{E}[g(t, \xi) | T = d, Z = z]$, where we use independence of ξ and T, Z. We can write:

$$\mathsf{E}[Y(t) | T = d, Z = z] - \tilde{g}(t, d) = \frac{h(t)}{h(d)} \left(\mathsf{E}[Y(d) | T = d, Z = z] - \tilde{g}(d, d) \right) \tag{271}$$

Using b): ii) yields the result.

6.17. Simulation exercise

We now consider the following parametric example:

$$Y(t) = c + \alpha t + \beta\eta + Z \tag{272}$$

$$T = \mathsf{I}\{\varepsilon + f(Z) \geq 0\} \tag{273}$$

$$\eta = min\{u, max\{\varepsilon, l\}\} \tag{274}$$

$$\varepsilon \quad \mathsf{N}(0,1) \tag{275}$$

Where $u, l, \alpha, \beta, c \quad \mathsf{R}$ and $u > l$. Moreover, ε is independent of all other variables. Also suppose for simplicity that $Z \quad [l;u]$ a.s. Consider:

$$\mathsf{E}[Y(t)|T=1, Z=z] = c + \alpha t + z + \beta\mathsf{E}[min\{u,\varepsilon\}|\varepsilon > -f(z)] = \tag{276}$$

$$= c + \alpha t + z + \beta\left(\frac{1-\Phi(u)}{\Phi(f(z))}u + \frac{\phi(f(z)) - \phi(u)}{\Phi(f(z))}\right) \tag{277}$$

$$\mathsf{E}[Y(t)|T=0, Z=z] = c + \alpha t + z + \beta\mathsf{E}[max\{u,\varepsilon\}|\varepsilon \leq -f(z)] = \tag{278}$$

$$= c + \alpha t + z + \beta\left(\frac{\Phi(l)}{\Phi(-f(z))}l + \frac{\phi(l) - \phi(f(z))}{\Phi(-f(z))}\right) \tag{279}$$

For the Figure, suppose:

$$t = 0$$

$$[l,u] = [-4,2]$$

$$Z \quad U[-1,1]$$

$$f(z) = -2z^4$$

$$g(z) = ln(z+1.1)$$

$$\beta = 0.1$$

6.18. Empirical analysis

	$ATE(3,2)$	$ATE(2,1)$	$ATE(1,0)$
cMIV-s	(0.059, 3.768)	(0.09, 3.761)	(0.103, 3.742)
	{0.053, 3.801}	{0.082, 3.81}	{0.094, 3.791}
cMIV-p	(0.036, 4.163)	(0.042, 4.185)	(0.053, 4.058)
	{0.033, 4.176}	{0.039, 4.225}	{0.049, 4.099}
cMIV-w	(0, 4.162)	(0, 4.072)	(0, 4.087)
	{0, 4.176}	{0, 4.102}	{0, 4.118}
MIV	(0, 4.163)	(0, 4.227)	(0, 4.108)
	{0, 4.175}	{0, 4.25}	{0, 4.134}
ETS	0.092	0.012	0.017

Table 2: Estimation results under various assumptions. CI in curly brackets are two-sided 95%, see Proposition 11.

6.19. Uniform rate of the debiased penalty function estimator

Our theoretical results show that under a polytope δ-condition the debiased penalty function estimator is at least $\sqrt{n}/w_n$ uniformly consistent. We now attempt to see if that rate is sharp uniformly, or whether the pointwise rate of $\sqrt{n}$ is achievable. This subsection describes the design of simulations that allow us to study the uniform rate of convergence of the debiased penalty function estimator.

The proof of pointwise $\sqrt{n}-$consistency of the debiased penalty function estimator relies on the fact that the value $L(x;\theta,w)$ at x outside the argmin set $\tilde{A}(\theta;w)$ is sufficiently well-separated from the optimal value $B(\theta)$. While at any fixed measure, including those that result in 'flat faces', there exists some 'separation constant' for a given distance from the argmin, this statement becomes problematic uniformly. In particular, around some $\overline{\theta}$ at which there occurs a flat face, there exist sequences θ_n, along which for any given distance of x from the argmin the difference between objective functions grows arbitrarily small.

It is worth emphasizing that the situation of an exact flat face is not problematic by itself, which is easy to see by drawing the picture of the example below at $a = 0$. Instead, the issue seems to occur when the measure grows arbitrarily close to a flat face. However, it seems that this is also not enough to undermine uniform $\sqrt{n}-$consistency: Slater's condition must also fail. Intuitively, if Slater's condition holds in the vicinity of $\overline{\theta}$, the estimator eventually becomes insensitive to w_n and delivers $\sqrt{n}-$consistency.

We consider the following linear program:

$$B(a,b,c,d) \equiv \min_{x,y} y - (1+a)x, \quad \text{s.t.:} \begin{cases} y \leq (1+b)x + d \\ y \geq (1+c)x \\ x \quad [-1;1] \end{cases}, \tag{280}$$

Where we take a to be fixed and indexing a probability measure. $b = 0, c = 0, d = 0$ are estimated via b_n, c_n, d_n as sample averages of independent $U[-0.5, 0.5]$ random variables. We now describe the design of our simulations:

1. We set $w_n = \frac{\ln n}{\ln 100}(\delta/1.5)^{-1}$, where δ is the biggest value for which the delta condition is satisfied over $a \quad [-0.1, 0.1]$.

2. For any fixed n, we take the grid of 9 points:

$$\mathsf{G}_n \equiv \{-0.1, 0, 0.1\} \quad \{-0.1C_1 n^{-1/2}, 0.1C_1 n^{-1/2}\}$$
$$\{-0.1C_2 w_n n^{-1/2}, 0.1C_2 w_n n^{-1/2}\} \quad \{-0.1C_3 w_n^{-1}, 0.1C_3 w_n^{-1}\},$$

   where C_i are chosen so that each point is equal to -0.1 at $n = 100$.

3. At each n, we run $N_{sim} = 10000$ simulations, each time computing b_n, c_n, d_n and plugging in to obtain:

$$\sup_{a \quad \mathsf{G}_n} |\tilde{B}(a, b_n, c_n, d_n; w_n) - B(a,0,0,0)| \tag{281}$$

4. We then compute the standard deviation of (281) across simulations at each n

5. We consider multiplying the resulting standard deviations by two rates: $\sqrt{n}$ and $\sqrt{n}/w_n$.

In all figures below the level of the red curve is equated to the level of the blue one at the smallest n to illustrate the growth rate.

From Figure 8, it appears that standard deviations multiplied by $\sqrt{n}$ are indeed exploding,

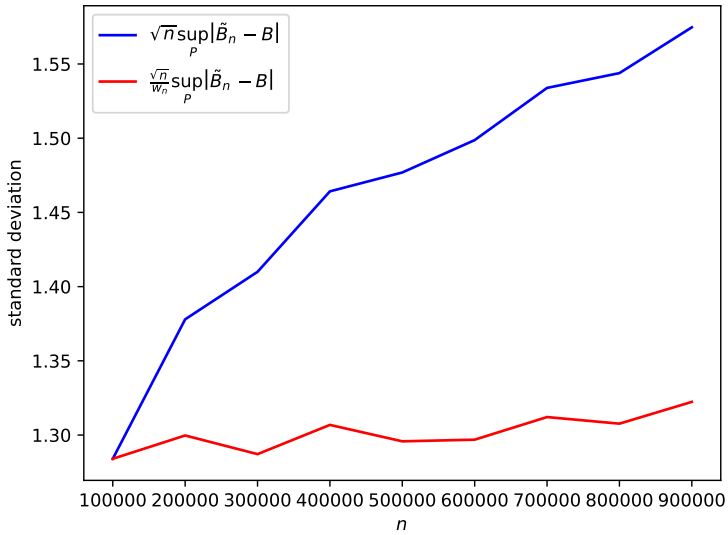


Figure 8: Uniformity of the penalized estimator: continuous vicinitiy of a flat face

although very slowly, while those multiplied by $\sqrt{n}/w_n$ are stable. It may be the case that the rate of $\sqrt{n}/w_n$ is sharp uniformly.

We next consider the grid that includes the flat face itself, but restricts the measures from approaching it from the left and right. In other words, we conduct the same simulation exercise with:

$$\mathsf{G}_n \equiv \{-0.1, 0, 0.1\} \quad \{-0.05(1 + C_1 n^{-1/2}), 0.05(1 + C_1 n^{-1/2})\}$$
$$\{-0.05(1 + C_2 w_n n^{-1/2}), 0.05(1 + C_2 w_n n^{-1/2})\} \quad \{-0.05(1 + C_3 w_n^{-1}), 0.05(1 + C_3 w_n^{-1})\}$$

In this case, Figure 9 suggests that uniform $\sqrt{n}$-consistency is achieved.

Finally, we return to the original grid G_n, but consider the case in which Slater's condition holds. For that reason, we take the true value of $d = 0.5$ by sampling d_n from $U[0, 1]$ instead. Once again, it appears that we obtain uniform $\sqrt{n}-$consistency.

Our simulation evidence thus suggests that while our estimator is only $\sqrt{n}/w_n$ uniformly consistent in general, it is $\sqrt{n}$-consistent uniformly apart from the sequences of probability measures, along which both Slater's condition fails and where a flat-face is 'approached' monotonically. It appears possible to rule out the latter scenario by considering a uniform condition similar to the $\delta-$condition we imposed before. This condition would restrict the set of measures under consideration to those at which the 'distance' from a flat face is either 0 or bounded away from 0 in some metric. Accordingly, it would likely cover the

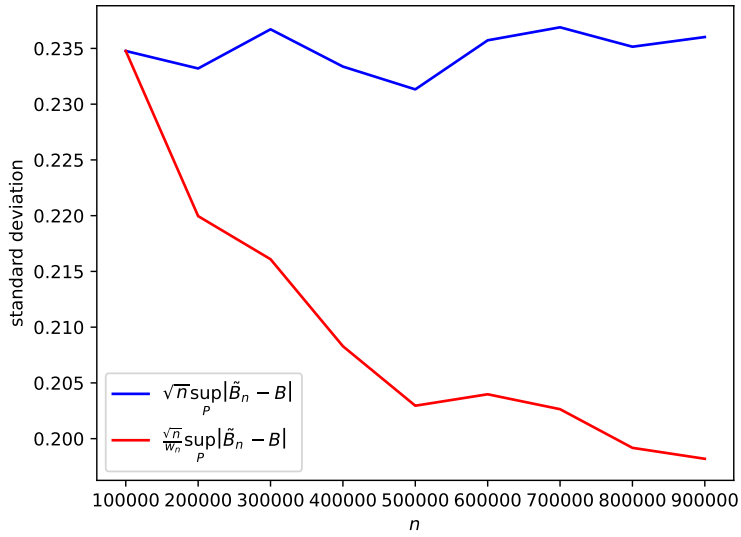


Figure 9: Uniformity of the penalized estimator: restricted vicinitiy of a flat face, flat face included.

unrestricted set of measures in the limit. These considerations, however, are the topic of a separate exploration, and space does not permit us to include them in this paper.

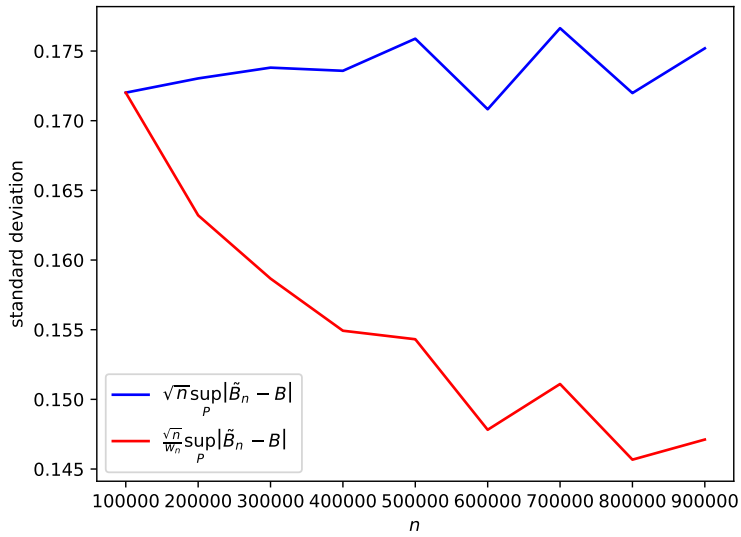


Figure 10: Uniformity of the penalized estimator: continuous vicinitiy of a flat face, Slater's condition holds.