Linear programming approach to partially identified econometric models

Andrei Voronin*

February 22, 2024

PRELIMINARY AND INCOMPLETE

Link to latest version

Abstract

Sharp bounds on partially identified parameters are often given by the values of linear programs (LPs). This paper introduces a novel estimator of the LP value $B(\theta) = \min_{Mx \geq c} p'x$, where $\theta = (p, M, c)$ is estimated. Unlike existing procedures, our estimator is $\sqrt{n}$-consistent whenever the true LP is feasible and finite, and remains valid under point-identification, over-identifying constraints, and solution multiplicity. Turning to uniform properties, we prove that the LP value cannot be uniformly consistently estimated over the unrestricted set of measures. We then show that our estimator achieves uniform consistency under a condition that appears minimal for the existence of any such estimator. We also obtain computationally efficient, asymptotically normal inference with exact asymptotic coverage. To complement our estimation results, we derive LP sharp bounds in a general identification setting. Our approach allows applied work to employ previously intractable conditions. We apply our findings to estimating returns to education. To that end, we propose the conditionally monotone IV assumption (cMIV) that tightens the classical monotone IV (MIV) bounds. We argue that cMIV remains unrestrictive relative to MIV 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%, whereas classical conditions fail to produce an informative bound.

*Department of Economics, UCLA. Email: avoronin@g.ucla.edu. I am grateful to Denis Chetverikov, Andres Santos, 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.

1. Introduction

In many partial identification frameworks[1], the sharp bounds on parameters correspond to the values of linear programs (LPs) that depend on identified functionals of the underlying probability measure. Specifically, the bounds take the form $B(\mathbb{P}) = B(\theta_0(\mathbb{P}))$, where

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

and $\theta_0(\mathbb{P})$ is the true value of parameter $\theta = (p', \text{vec}(M)', c')'$, estimated via a $\sqrt{n}$-consistent estimator $\hat{\theta}_n$. However, optimization problem (1) exhibits non-regular behavior, particularly when the underlying model is rich enough that some linear functionals of x are nearly or exactly point-identified over $\Theta_I = \{x \in \mathbb{R}^d : Mx \geq c\}$. In such cases, existing estimators of $B(\mathbb{P})$ are either inconsistent or rate-conservative, creating an undesirable tradeoff in empirical work: richer models provide tighter bounds but complicate their estimation.

To address this issue, we develop a novel estimator of $B(\mathbb{P})$. Our estimator takes the form $\hat{B}(\hat{\theta}_n; w_n)$, where

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

and $w_n \to \infty$ is the penalty parameter, with $\frac{w_n}{\sqrt{n}} \to 0$. We refer to $\hat{B}(\hat{\theta}_n; w_n)$ as *the debiased penalty function estimator*. Only assuming that Θ_I is non-empty and contained in a known compact $\mathcal{X}$, we show that $\hat{B}(\hat{\theta}_n; w_n)$ is $\sqrt{n}-$consistent for any w_n satisfying the above conditions[2]. In contrast, the plug-in estimator is not generally consistent and may fail to exist with non-vanishing probability, while the alternative set-expansion estimator based on Chernozhukov et al. (2007) is rate-conservative and may fail to exist in finite samples. Figure 1 gives a preview of the comparative performance of these estimators.

We obtain an asymptotically normal version of our estimator via sample-splitting and construct confidence regions with exact asymptotic coverage. By comparison, existing procedures either rely on further conditions (Gafarov, 2024), or lead to asymptotically conservative inference (Cho and Russell, 2023). Notably, the approach most commonly used in applied research—combining plug-in estimation with bootstrap (De Haan (2017), Cygan-Rehm et al. (2017), Siddique (2013), Kreider et al. (2012), Gundersen et al. (2012), Blundell et al. (2007))—may not provide valid confidence intervals even when the underlying model is far from point-identification.

[1]Including conditional moment inequalities (Andrews et al., 2023), generalized IV models (Mogstad et al., 2018), revealed preference restrictions (Kline and Tartari, 2016), intersection bounds (Honoré and Lleras-Muney, 2006), dynamic discrete choice panels (Honoré and Tamer, 2006) and shape restrictions Manski and Pepper (2000).

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

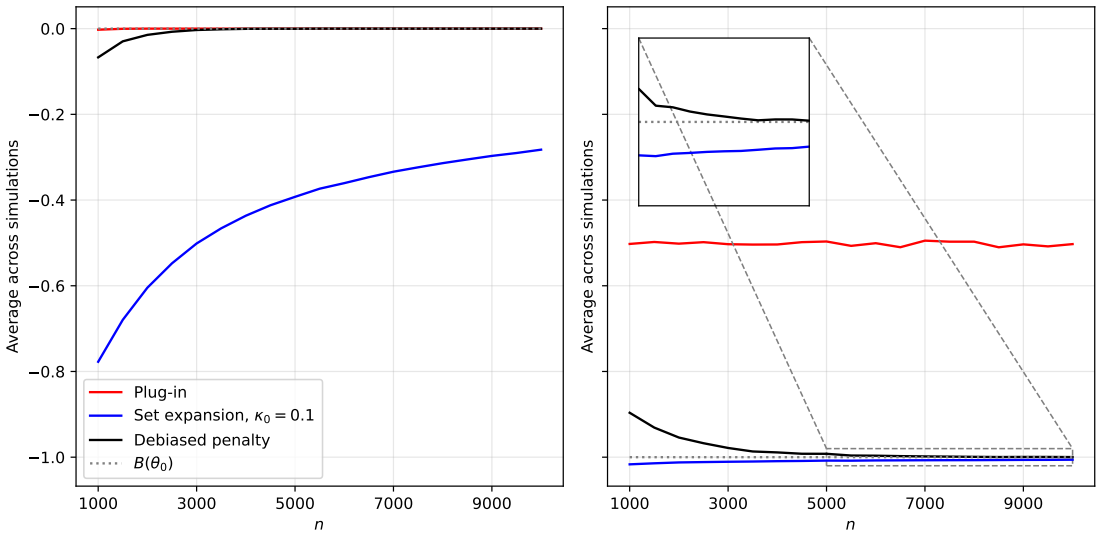


Figure 1: Comparison of estimators for two measures with the true values of 0 and −1, left to right. Average estimate across 10000 simulations. See Section 2.6 for details.

Using an equivalence between unconstrained piecewise-linear problems and auxiliary LPs, one can compute our estimator in polynomial time with 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[3].

Turning to uniform asymptotic theory, we first establish a general impossibility result: using Le Cam's binary testing method, we show that no uniformly consistent estimator exists when the estimated functional is discontinuous in the total variation norm. Applied to LP, this implies that $B(\mathbb{P})$ cannot be uniformly consistently estimated over the unrestricted set of measures $\mathcal{P}$. To make progress, we introduce the '$\delta-$condition' that parametrizes $\mathcal{P}$ by restricting it to the measures at which the smallest singular value of some full-rank submatrix of constraints binding at an optimal vertex is lower-bounded by a $\delta > 0$. This condition is minimal in the sense that any measure from $\mathcal{P}$ satisfies it for some δ, ensuring the family of restricted measures' sets covers $\mathcal{P}$ as δ grows small. Unlike the conditions in Gafarov (2024), it does not exclude economically relevant *problematic* cases, such as point-identification and over-identification, nor does it preclude solution multiplicity. Under the $\delta-$condition, our estimator is shown to be uniformly consistent.

To complement our estimation procedure, we develop a general identification framework that results in LP bounds. In particular, we derive novel sharp bounds for a broad class of

[3]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 the latter work. Our approach requires computing a LP once, which takes around 0.0022 seconds on average.

treatment parameters[4] 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

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

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. In AICM models, θ from (1) is usually a function of identified conditional moments $(\mathbb{E}[Y|T=t,Z=z])_{t,z}$ and the identified joint distribution of T, Z, while x collects relevant unobserved conditional moments. 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[5].

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

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

where $(Y(t))_{t\in\mathcal{T}}$ are potential log-wage schedules corresponding to education levels T, and Z is a proxy for ability based on Saber test scores. Sets A parametrize different versions of cMIV and are chosen by the researcher[6]. While an explicit form for the sharp bounds under some versions of cMIV may not be feasible, their LP representation follows from our general identification result for (3). The cMIV conditions we consider yield tighter bounds than the classical monotone instrumental variables (MIV) assumption of Manski and Pepper (2000). We argue, however, that they remain unrestrictive in many applications, including ours. While empirical literature (e.g., De Haan (2017)) has visually examined the monotonicity of observed conditional moments, i.e. $A = \{t\}$, to justify applying MIV, 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). Using Saber test scores as a cMIV, we find that obtaining a university degree increases average wages by at least 5.91% in Colombia. In contrast, the classical conditions fail to produce an informative bound.

This paper also contributes two 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

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

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

[6]We explore various choices, such as letting A run through all singletons $\{d\}$ and the full support $\mathcal{T}$. See Section 4 for details.

restrictions, empirical work has used it even in the absence of a theoretical justification, obtaining bounds that were either not sharp, or invalid (see Lafférs (2013)). Our method yields sharp bounds and a valid estimation procedure for that setting even under continuous outcomes. Another auxiliary contribution consists in a novel lower bound on the ℓ_1-deviation from a non-empty bounded polytope in terms of Euclidean distance from the polytope. It may provide insights into the behavior of ℓ_1-penalized solutions of systems of linear inequalities studied in the control theory literature (e.g. Pinar and Chen (1999)).

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 the Appendix. A theoretically $\sqrt{n}-$uniformly consistent estimator follows from our analysis, but it depends on an unobserved parameter δ that is difficult to estimate, so we 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 $\hat{c}_n$, but a known matrix M and a coefficient vector p. Bhattacharya (2009) considers a LP with estimated $\hat{p}_n$ and known M, c. Methods developed under a known M assumption cannot be easily extended to the setting when M is estimated, as will become evident in Section 2. Mogstad et al. (2018) construct a set-expansion estimator and prove its consistency. Syrgkanis et al. (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 (SC) hold. Gafarov's conditions may be restrictive in some applications - for example, under AICM, see Section 2. Andrews et al. (2023) develop inference in a special case of LP estimation framework, which arises from their model. In their problem, SC holds and θ has a particular structure, making their findings hard to generalize. Cho and Russell (2023) add random distortions to p and introduce random non-vanishing expansions to Θ_I to enforce uniform Hadamard differentiability of the perturbed LP. Their approach yields uniformly valid, yet conservative confidence regions for $B(\mathbb{P})$ when θ is affine

[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.

in unconditional population moments, and their practical procedure implicitly assumes that the SC holds[9]. We conduct Monte Carlo simulations to compare our approach with existing methods in Section 2.6.

A number of frameworks that result in bounds of form (1) are described in the review article by Kline and Tamer (2023). Other examples include conditional moment inequalities (Andrews et al., 2023), generalized IV models (Mogstad et al., 2018) and a strand of models nested in (3), pioneered by Manski (1997) and Manski and Pepper (2000, 2009) (MP). The novel LP sharp bounds in Theorem 3.1 coincide with or tighten the bounds in Blundell et al. (2007), Boes (2009), Siddique (2013), Kreider et al. (2012), De Haan (2017) and Cygan-Rehm et al. (2017). To compute the bounds, the above papers use the plug-in estimator $B(\hat{\theta}_n)$ combined with bootstrap for inference. Our results show that this procedure relies on rather strong assumptions. The exact inference procedure in Theorem 2.3 could be used instead.

AICM approach complements the method of Mogstad et al. (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 (3) 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 et al. (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 $M \in \mathbb{R}^{n\times m}$. 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. For $M \in \mathbb{R}^{m\times n}$ and $A \subseteq \overline{1,m}$, $M_A \in \mathbb{R}^{|A|\times n}$ is the submatrix of the rows

[9]We discuss this in more detail in Section 2.6.

[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 et al. (2018) may be used. If any of these conditions fail, our approach is an alternative.

of M 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. In a normed space S, the distance between $x \in S$ and $A \subseteq S$ is written as $d(x, A) \equiv \inf_{a\in A} ||x - a||$, 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 $A, B \subseteq S$. For $A \subseteq S$ the open expansion is $A^\varepsilon \equiv \{s \in S : d(s, A) < \varepsilon\}$. $Int(A)$ and $Cl(A)$ are the interior and closure of $A \subseteq \mathbb{R}^d$, while $Cone(A)$ is its conical hull. If A is a matrix, $Cone(A)$ is the conical hull of its columns. $s(x, A) \equiv \max_{a\in A} x'a$ for a compact $A \subseteq \mathbb{R}^d$ and $x \in \mathbb{R}^d$ 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 $v, u \in \mathbb{R}^d$ 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. $\iota_d \in \mathbb{R}^d$ is a vector of ones, $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix, and the subscript is dropped occasionally. Operator $\mathbb{E}_\mathbb{P}$ is the expectation under a measure $\mathbb{P}$, and the subscript is dropped whenever it does not cause confusion. The statement $w_n \to \infty$ w.p.a.1 means that $\forall M > 0$, $\lim_{n\to\infty} \mathbb{P}[w_n > M] = 1$. We adopt the convention $\inf \emptyset = +\infty$, and $\sup \emptyset = -\infty$.

2. LP estimation framework

In many partial identification settings, bounds on the parameters of interest can be expressed as LP values (see Kline and Tamer (2023) for a review). Readers who prefer to first see an identification framework resulting in such bounds may refer to Section 3, where LP sharp bounds are derived for a general class of AICM models. This section focuses on the estimation theory for such problems. The LP value function is given by

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

where $M \in \mathbb{R}^{q\times d}$, $c \in \mathbb{R}^q$ and $p \in \mathbb{R}^d$. The vector $\theta \equiv (p', c', \text{vec}(M)')'$ collects parameters of the LP. The estimable value of these parameters at a fixed true measure is denoted by $\theta_0 \in \mathbb{R}^S$, with $S = qd + q + d$. The value of interest is therefore $B(\theta_0)$.

Remark 2.1. (4) does not rule out equality constraints, as $Ax = b \iff Ax \geq b \wedge -Ax \geq -b$.

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

Assumption A0 (Pointwise setup). *Suppose that at the fixed true parameter θ_0: i) The identified set is non-empty, $\Theta_I(\theta_0) \neq \emptyset$; ii) $\Theta_I(\theta_0) \subseteq \mathcal{X}$ for a known compact $\mathcal{X} \subseteq \mathbb{R}^d$ and iii) There is an estimator $\hat{\theta}_n \equiv (\hat{p}'_n, \hat{c}'_n, vec(\hat{M}_n)')'$: $||\hat{\theta}_n - \theta_0|| = O_p(1/\sqrt{n})$*

Assumption A0 is maintained throughout this section, while the rest of the conditions are stated explicitly. A0.i means that the underlying model *cannot be rejected*. It does not imply that the identifying restrictions are correctly specified. A0.ii is a mild restriction that usually

holds in applications, for example under bounded outcomes in AICM models (see Section 3). The $\sqrt{n}-$consistent estimator[11] in A0.iii typically follows from CLT and the Delta-Method.

The following primal and dual solution sets will prove useful in our discussion:

$$\mathcal{A}(\theta) \equiv \arg\min_{Mx \geq c} p'x, \quad \Lambda(\theta) = \arg\max_{M'\lambda = p} c'\lambda.$$

Assumption A0 implies that a finite $B(\theta_0)$ is attained as a minimum in (4), $\Theta_I(\theta_0)$ and $\mathcal{A}(\theta_0)$ are non-empty compacts, and $\Lambda(\theta_0)$ is non-empty. We now briefly discuss the typically imposed regularity conditions.

Definition. Slater's condition (SC) is the assertion that $\text{Int}(\Theta_I) \neq \emptyset$.[12]

SC rules out point-identification of any linear functional of x. In particular, it precludes exact point-identification of the target $B(\theta_0)$ and point-identification of x, i.e. the case when $|\Theta_I| = 1$. Most existing methods rely on SC explicitly (Gafarov (2024)) or implicitly (Cho and Russell (2023), Andrews et al. (2023)). Even an 'approximate' failure of SC, when the true identified set Θ_I becomes 'thin', may be problematic for the existing methods in finite samples. Our simulation evidence illustrates this, see Section 2.6. This creates an undesirable tradeoff: high identification power implies poor estimation quality.

Definition. Linear independence constraint qualification (LICQ) is the assertion that the submatrix of binding inequality constraints at any $x \in \mathcal{A}(\theta_0)$ is full-rank.

LICQ precludes the existence of overidentifying constraints at the optimum. It may be hard to justify in 'bigger' models, like the one we develop and apply in Sections 4 and 5. These feature a larger number of inequality constraints that may have similar identifying power, so it is not ex-ante clear why there must not be overidentification at the optimum. One may also interpret LICQ as ruling out parameters-on-the-boundary, as the following example clarifies.

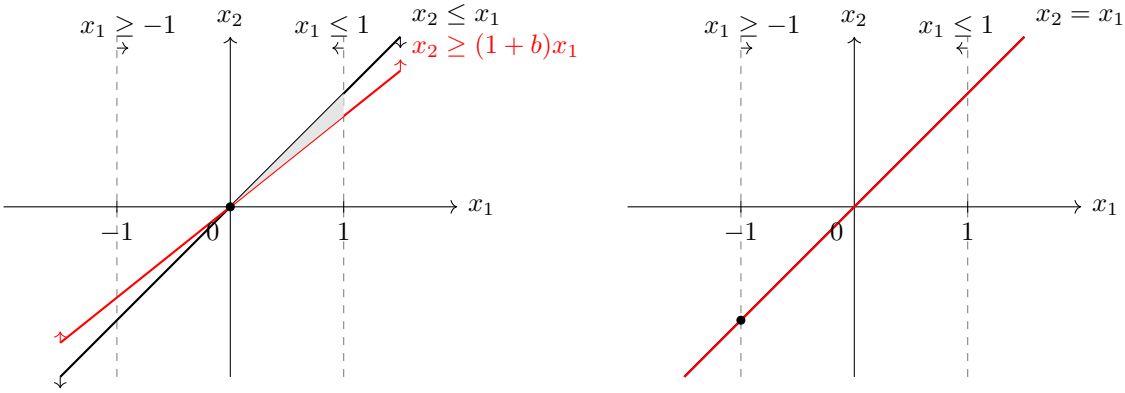
Example 2.1. Example 2.1 in Fang and Santos (2018) can be restated as a LP: $B(\theta_0) = \max\{0, \mathbb{E}[X]\} = \min_{t \in \mathbb{R}} t$ s.t. $t \geq 0, t \geq \mathbb{E}[X]$. LICQ fails in that program if $\mathbb{E}[X] = 0$, which corresponds to the parameter-on-the-boundary case from Andrews (1999, 2000).

Definition. No flat faces condition (NFF) is the assertion that $|\mathcal{A}(\theta_0)| = 1$.

The notion of flat faces thus corresponds to $|\mathcal{A}(\theta_0)| \neq 1$, i.e. the situation in which the bound on the target parameter $p'x$ is achieved at multiple partially identified features x.

[11] $\sqrt{n}$ rate straightforwardly generalizes to any $r_n \to \infty$. We focus on $\sqrt{n}$ throughout for expositional simplicity.

[12] We give simplified versions of assumptions here for simplicity of exposition. In the presence of 'true' equalities $Ax = b$, SC should be stated in terms of *Relint*, allowing for point-identification along 'true' equalities. LICQ should similarly be restated to account for equalities, as in Gafarov (2024).



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

Figure 2: The feasible region in (5) for two values of b.

Assumption A0 does not impose LICQ or SC, nor does it rule out flat faces. Estimating $B(\theta_0)$ without these conditions is challenging due to the irregular behavior of $B(\cdot)$. If SC fails, $B(\cdot)$ may be discontinuous at θ_0, and the plug-in estimator $B(\hat{\theta}_n)$ may not be pointwise consistent.

Proposition 2.1. *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.*

Proof. We provide a simple example for the first part. Consider

$$B(b) = \min_x x_1 \quad \text{s.t.} : x_2 \geq (1+b)x_1,\ x_2 \leq x_1,\ x_1 \in [-1;1], \tag{5}$$

where b is estimated via $\hat{b}_n = b + \frac{1}{n}\sum_{i=1}^n U_i$ with $U_i \sim U[-1;1]$ i.i.d. Suppose in population $b = 0$, as in Figure 2b. The true value is then $B(0) = -1$, attained at $x^* = -\iota$. The plug-in estimator collapses to

$$B(\hat{b}_n) = -\mathbb{1}\{\hat{b}_n \geq 0\} \nrightarrow -1 \text{ in probability.}$$

For the second part, consider the family

$$B(a) = \min_x x_1 \quad \text{s.t.} : x_2 \geq x_1 + a,\ x_2 \leq x_1,\ x_1 \in [-1;1],$$

where a is estimated via $\hat{a}_n = a + \frac{1}{n}\sum_{i=1}^n U_i$ with U_i as before. Suppose $a = 0$. If $\hat{a}_n > 0$, the plug-in estimator $B(\hat{a}_n)$ does not exist. This occurs with probability $1/2$ for any $n \in \mathbb{N}$. ■

In some special cases (e.g., Honoré and Tamer (2006)), SC may be argued to hold, ensuring the continuity of $B(\cdot)$. However, an additional challenge arises: $B(\cdot)$ is not necessarily

Hadamard differentiable unless LICQ and NFF also hold. This complicates inference, as Proposition 2.4 in Section 2.2 illustrates.

This section addresses Propositions 2.1 and 2.4. Section 2.1 introduces the penalty function estimator and its debiased version, which we show to be $\sqrt{n}$-pointwise consistent under A0. Section 2.2 develops a computationally efficient inference procedure with exact coverage under A0. Turning to uniform properties, Section 2.3 presents a general impossibility result for discontinuous functionals. It implies that the LP value cannot be uniformly consistently estimated under a uniform version of A0 alone. We then characterize a broad class of measures over which a uniformly consistent estimator exists. Sections 2.4 and 2.5 establish the uniform rates of the penalty function estimators over this class. Section 2.6 provides simulation evidence.

2.1. Consistency

2.1.a. *Penalty function estimator.* We now develop a consistent estimator that is inspired by the theory of exact penalty functions. The idea is to restate (4) as an unconstrained penalized problem. Define the L_1-penalized version of the LP objective as

$$L(x;\theta,w) \equiv p'x + w'(c - Mx)^+,$$

and consider the unconstrained problem

$$\tilde{B}(\theta;w) \equiv \min_{x\in\mathcal{X}} L(x;\theta,w), \quad \tilde{\mathcal{A}}(\theta;w) \equiv \arg\min_{x\in\mathcal{X}} L(x;\theta,w). \tag{6}$$

We use $\tilde{B}(\cdot)$ to obtain a preliminary estimator, which we term *the penalty function estimator*. Note that $L(x;\theta,w) = p'x$ at any $x \in \Theta_I$, i.e. the penalized function is equal to the objective function whenever the constraint in (4) holds.

Assumption A1 (Penalty parameter). *The penalty vector $w \in \mathbb{R}^q$ is such that in the initial LP there exists a KKT vector $\lambda^* \in \Lambda(\theta_0)$ such that $w > \lambda^*$.*

Remark 2.2. Note that Assumption A1 does not require w to be component-wise larger than all KKT vectors. In any solvable LP there exists at least one $\lambda^* < \infty$, so at any fixed θ_0 there always exists a large enough w that satisfies A1.

Remark 2.3. If it is known that i) $B(\theta_0) < K$ for some $K > 0$ and ii) $c > \underline{c} > 0$ for some known $\underline{c} > 0$, Assumption A1 is satisfied for $w = \iota K/\underline{c}$, which is known.

The following Lemma is key to understanding the penalty function approach. It asserts that under Assumption A1 the L_1-penalty function is *exact* for the LP in (4).

Lemma 2.1. For any $(\theta, w) \in \mathbb{R}^S \times \mathbb{R}^q_+$, if $\Theta_I(\theta) \subseteq \mathcal{X}$, then:

$$\tilde{B}(\theta; w) \leq B(\theta) \tag{7}$$

If the pair (θ_0, w) satisfies Assumption A1, then: i) (7) holds with an equality, and ii) optimal solutions coincide: $\tilde{\mathcal{A}}(\theta_0; w) = \mathcal{A}(\theta_0)$.

The deterministic result in Lemma 2.1, combined with the observation that the objective function in probabiloty converges uniformly in x under A0, establish that the penalty function estimator with a fixed w is consistent under A1:

Proposition 2.2. *Under Assumption A1,*

$$\tilde{B}(\hat{\theta}_n; w) \xrightarrow{p} B(\theta_0).$$

In what follows, the variation across coordinates of w will not be of interest. We shall thus treat $w \in \mathbb{R}_+$ as a scalar that induces the penalty vector $w\iota$. Based on Lemma 2.1 and Proposition 2.2, it might seem that w should be selected to be as large as possible. This, however, yields a generally inconsistent estimator if SC fails.

Remark 2.4. Consider (5) with $b = 0$ and suppose $w > 2$. If $\hat{b}_n < 0$, there exists a sample KKT vector $\hat{\lambda}_n \in \Lambda(\hat{b}_n)$, whose largest coordinate is $||\hat{\lambda}_n||_\infty = |\hat{b}_n^{-1}|$. So, if also $w > |\hat{b}_n^{-1}|$, the penalty estimator selects an incorrect optimum $(0, 0)$ in light of Lemma 2.1. Since $\hat{b}_n \xrightarrow{p} 0$, at a large enough sample size $|\hat{b}_n^{-1}|$ will exceed any fixed w with high probability and the correct minimum of -1 will be estimated. However, that logic fails in finite samples if w is 'large'.

This observation justifies the need to study $w \to \infty$ asymptotic theory. We show that the penalty parameter can be allowed to diverge at the rate dominated by $\sqrt{n}$.

Theorem 2.1. For any $w_n \to \infty$ w.p.a.1 with $\frac{w_n}{\sqrt{n}} \xrightarrow{p} 0$, we have

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

The estimator in Theorem 2.1 does not rely on Assumption A1, as the latter is always satisfied at a fixed measure for a large enough n when $w_n \to \infty$.

2.1.b. *Debiased penalty estimator.* The $\frac{w_n}{\sqrt{n}}$ rate of convergence in Theorem 2.1 is determined by the slowly vanishing penalty term. This term is a product of the deviation from the true polytope, that vanishes at $\frac{1}{\sqrt{n}}$, and an exploding sequence w_n. It is thus reasonable to

ask whether the $\sqrt{n}-$rate could be restored by dropping the penalty term, i.e. *debiasing* the penalty function. We show that this can be done. Before we proceed, let us make the following simplification. Without loss of generality, suppose that

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

To see why this is w.l.g., note that one can set $p = e_1 = (1\ 0 \dots 0)'$ and add an auxiliary variable for the value of the problem in the first position of x (see Gafarov (2024)).

The following theorem is one of the main contributions of this paper.

Theorem 2.2. Suppose $\mathcal{A}(\theta_0) \subseteq Int(\mathcal{X})$. For any $w_n \to \infty$ w.p.a.1 with $\frac{w_n}{\sqrt{n}} \xrightarrow{p} 0$,

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

Before we discuss the Theorem, let us introduce some notation and jargon. For $x \in \mathbb{R}^d$, define the set of constraints that bind at it, when evaluated at $\tilde{\theta}$, by $J(x; \tilde{\theta}) \equiv \{j \in [q] : \tilde{M}_j x = \tilde{c}_j\}$.

Definition (Vertex)**.** We call $x \in \tilde{\mathcal{A}}(\hat{\theta}_n; w_n)$ a *vertex-solution* if the corresponding matrix of binding constraints, $\hat{M}_{nJ(x;\hat{\theta}_n)}$, has full column rank.

Definition (Nice face)**.** We say that a set $A \subseteq [q]$ corresponds to *a nice face* $F \equiv \{x \in \mathbb{R}^d : M_A x = c_A\}$ if $p'x = B(\theta_0)$ for any $x \in F$.

It should be noted that a nice face F is not necessarily a valid $k-$face of the polytope Θ_I.

Intuitively, the proof of Theorem 2.2 proceeds in two steps. First, by anti-concentration arguments we establish that with high probability asymptotically the penalty function estimator manages to select a vertex-solution $\hat{x}_n \in \tilde{\mathcal{A}}(\hat{\theta}_n; w_n)$ with $\hat{A}_n = J(\hat{x}_n; \theta_n)$, such that $\hat{A}_n$ corresponds to a nice face $F = \{x \in \mathbb{R}^d : M_{\hat{A}_n} x = c_{\hat{A}_n}\}$.

Once a nice face has been selected, the $\sqrt{n}-$convergence of $p'\hat{x}_n$ to $B(\theta_0)$ obtains as a consequence of $(\hat{M}_{nA}, \hat{c}_{nA})$ converging to (M_A, c_A) at this rate for a fixed $A \subseteq [q]$. To illustrate this idea, suppose for a moment that with high probability asymptotically the penalty function estimator only selects nice faces corresponding to $\hat{A}_n$, such that $M_{\hat{A}_n}$ is also an invertible square matrix. Then, $\hat{x}_n = \hat{M}^{-1}_{\hat{A}_n} \hat{c}_{\hat{A}_n} + o_p(n^{-0.5})$, which converges to $x = M^{-1}_{\hat{A}_n} c_{\hat{A}_n} + o_p(n^{-0.5})$ at rate $\sqrt{n}$. Because $x \in F$, with F being a nice face w.p.a.1, one has $p'x = B(\theta_0) + o_p(n^{-0.5})$, so that also $p'\hat{x}_n \to B(\theta_0)$ at $\sqrt{n}$ by CMT.

Remark 2.5. The result in Theorem 2.2 is uniform over the argmin set, so in the context of lower/upper bound estimation one may use $\max / \min_{\tilde{\mathcal{A}}(\hat{\theta}_n; w_n)} p'x$ to obtain the tightest bound.

We define *the debiased penalty function estimator* as

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

The discussion of uniform asymptotic theory in Section 2.3 sheds light on the role of w_n and the trade-off involved in its selection. The practical guidance on selecting w_n is then developed on the basis of our results and random matrix theory, see the Appendix.

Remark 2.6. An alternative estimator can be constructed using a set-expansion argument: $\check{B}_n = \min \hat{p}_n' x$ s.t. $\hat{M}_n x \geq \hat{c}_n - \sqrt{\frac{\kappa_n}{n}}\iota$. In the Appendix, we show that the results from Chernozhukov et al. (2007) and the geometry of polytopes imply that $\check{B}_n$ with an appropriately chosen, diverging κ_n, is consistent for $B(\theta_0)$. However, it can be rate-conservative, converging at $\sqrt{n}\kappa_n^{-1/2}$. It appears to perform worse than the debiased penalty function estimator $\hat{B}_n$ in our simulations, see Section 2.6.

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) or Bhattacharya (2009).

Assumption B0 (Random sample). *Suppose $\hat{\theta}_n = \hat{\theta}_n(\mathcal{D}_n)$ is a measurable function of the sample $\mathcal{D}_n \equiv \{W_1, W_2, \ldots, W_n\}$, where $W_i \in \mathbb{R}^{d_W}$, $i \in [n]$ are i.i.d. random vectors.*

We suppose that Assumption B0 holds throughout Section 2.2, whereas the rest of the conditions are imposed explicitly.

Assumption B1 (Asymptotic normality). *The estimator $\hat{\theta}_n$ is such that, for $\Sigma < \infty$,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} \mathbb{G}_0 \sim \mathcal{N}(0, \Sigma).$$

Assumption B1 is typically warranted by reference to CLT and the Delta Method when $\hat{\theta}_n = g(n^{-1}\sum_{i=1}^n W_i)$ for some smooth $g(\cdot)$, as in the AICM models (3), see Section 3.

2.2.a. *Bootstrapping the plug-in fails even under SC.* To further justify the need for our inferential procedure, we first examine the properties of the approach that combines bootstrap on $\hat{\theta}_n$ with the plug-in estimator $B(\hat{\theta}_n)$. This method is widely used in empirical literature applying AICM conditions (Blundell et al. (2007), Kreider et al. (2012), Gundersen et al. (2012), Siddique (2013), De Haan (2017), and Cygan-Rehm et al. (2017)).

In light of Proposition 2.1, this approach is inapplicable when SC fails. In practice, researchers attempting to apply it to a LP with a small or empty interior of Θ_I will likely encounter frequent bootstrap failures.

In some cases, SC may be established. This is true, for example, if the bound of interest can be expessed as an intersection bound $B(\theta_0) = \max\{c_1, c_2, \ldots, c_q\} = \min_{t\in\mathbb{R}} t \;\; \text{s.t. } t \geq c_i, \; i \in [q]$, where $\theta_0 = (1\ 1\ c')'$ (as in Chernozhukov et al. (2013)).

Yet, even if SC holds, we demonstrate that bootstrap inference based on the plug-in estimator is not valid, unless NFF and LICQ also hold. This observation can be viewed as a generalization of the parameter-on-the-boundary problem of Andrews (1999, 2000).

Definition. Let $\mathbb{D}$ and $\mathbb{E}$ be Banach spaces, and $f : \mathbb{D}_f \subseteq \mathbb{D} \to \mathbb{E}$. The map f is said to be Hadamard directionally differentiable at $\upsilon \in \mathbb{D}_f$ tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$, if there is a continuous map $f'_\upsilon : \mathbb{D}_0 \to \mathbb{E}$, such that

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

for all sequences $\{h_n \subset \mathbb{D}\}$ and $\{t_n\} \subset \mathbb{R}_+$ such that $t_n \to 0^+$, $h_n \to h \in \mathbb{D}_0$ as $n \to \infty$ and $\upsilon + t_n h_n \in \mathbb{D}_f$ for all n. If, moreover, $f'_\upsilon(h)$ is linear in h, the map f is said to be fully Hadamard differentiable.

For simplicity of exposition, we abstract from the case of 'true equalities' in Θ_I. The results extend trivially to this case if SC is defined in terms of relative interior.

Lemma 2.2. Under SC, $B(\cdot)$ is Hadamard directionally differentiable at θ_0. The directional derivative is given by

$$B'_{\theta_0}(h) = \inf_{x\in\mathcal{A}(\theta_0)} \sup_{\lambda\in\Lambda(\theta_0)} h'_p x + \sum_{i=1}^{Q} \lambda_i (h_{c_i} - h'_{M_i} x), \tag{9}$$

where $h = (h'_p, h'_{M_1}, \ldots, h'_{M_q}, h_{c_1}, \ldots, h_{c_q})'$ is the direction of the increment in θ.

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

Hadamard directional differentiability of $B(\theta)$ is sufficient for convergence in law.

Proposition 2.3. *Under SC and Assumption B1, it follows that*

$$\sqrt{n}(B(\hat{\theta}_n) - B(\theta_0)) \xrightarrow{L} B'_{\theta_0}(\mathbb{G}_0)$$

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

Gaussianity of $B'_{\theta_0}(\mathbb{G}_0)$ is a necessary condition for bootstrap consistency (Fang and Santos, 2018). Consequently, the empirical literature using bootstrap with the plug-in estimator has implicitly relied on this assumption. However, $B'_{\theta_0}(\mathbb{G}_0)$ is not normal unless full Hadamard

differentiability holds, i.e. $B'_{\theta_0}(h)$ is linear in h. As (9) suggests, this is not generally the case. Theorem 3.1 in Fang and Santos (2018) establishes that bootstrap is inconsistent for the distribution when $B'_{\theta_0}(h)$ fails to be linear. The typically applied plug-in and bootstrap combination is then only valid under further restrictive assumptions[13]:

Proposition 2.4. *If Assumption B1 holds, Supp*$(\mathbb{G}_0) = \mathbb{R}^S$ *and* θ_n^* *satisfies Assumption 3 in Fang and Santos (2018), bootstrap is consistent for the distribution of* $B(\hat{\theta}_n)$ *in the sense that*

$$\sup_{f \in BL_1(\mathbb{R})} \left| \mathbb{E}[f(\sqrt{n}(B(\theta_n^*) - B(\hat{\theta}_n)))|\mathcal{D}_n] - \mathbb{E}[f(B'_{\theta_0}(\mathbb{G}_0))] \right| = o_p(1),$$

if and only if i) SC, ii) NFF, iii) LICQ all hold at θ_0.

Remark 2.7. A consistent estimator for the distribution of the plug-in under SC can be obtained by combining the Functional Delta Method (FDM) of Fang and Santos (2018) with the Numerical Delta Method (NDM) given in Hong and Li (2015), see the Appendix.

Remark 2.8. The penalty function estimator is H.d.d. in θ (see the Appendix), so FDM + NDM combination yields exact inference for it. This approach still relies on an arbitrarily selected sequence ϵ_n and features a fixed w, and so does not appear satisfactory.

Remark 2.9. In the Appendix, we show that the set-expansion estimator has a Lipschitz-bounded bias. A conservative inference procedure based on it can then be obtained using FDM + NDM.

2.2.b. *Exact inference on a debiased estimator.* In this section, we develop our approach to statistical inference on $B(\theta_0)$, which achieves exact asymptotic coverage. Our method relies on an asymptotically normal version of the debiased penalty estimator with $w_n \to \infty$ and is outlined in Algorithm 1. Before presenting the main result, we introduce auxiliary constructions and discuss our assumptions.

For the true $\theta_0 = (c', \text{vec}(M)')'$ and some subset of indices $A \subseteq [q]$, consider

$$\exists x \in \mathcal{A}(\theta_0) : M_A x = c_A, \tag{10}$$

$$p \in \mathcal{R}(M'_A). \tag{11}$$

Equation (10) is satisfied if constraints A may bind simultaneously at some solutions of the original LP, while (11) holds if the objective function's gradient p is a linear combination of the gradients of inequalities from A. For example, if A is a set of *all binding constraints* at some $x \in \mathcal{A}(\theta_0)$, equation (11) follows from KKT conditions.

[13]The full-support condition in Proposition 2.4 is imposed for expositional purposes. Sufficiency of conditions i, ii, iii holds generally, whereas necessity obtains whenever the derivative $B'_{\theta_0}(h)$ is not linear over Supp$(\mathbb{G}_0)$ when $\mathcal{A}(\theta_0), \Lambda(\theta_0)$ are not singletons.

Algorithm 1 (Inference procedure)

Given data $\mathcal{D}_n$, estimators $\hat{\theta}(\mathcal{D}_n)$ and $\hat{\Sigma}_n = \hat{\Sigma}(\mathcal{D}_n)$, penalty vector $w(n, \mathcal{D}_n) \in \mathbb{R}^q$ and constants $\gamma \in (0; 1)$, $\overline{v} > 0$, follow the steps below to obtain confidence intervals for $B(\theta_0)$.

Step 1 (Split the sample):

1: Randomly split $\mathcal{D}_n$ into two folds $\{\mathcal{D}^{(f)}\}_{f=1,2}$ of sizes $n_1 = \lfloor \gamma n \rfloor$, $n_2 = n - n_1$

2: Compute $\hat{\theta}^{(f)} \equiv \hat{\theta}(\mathcal{D}^{(f)}) = (\hat{c}^{(f)\prime}, \text{vec}(\hat{M}^{(f)})')'$ for $f = 1, 2$

Step 2 (Find the vertex):

1: On the first fold, compute the penalty estimator's arg min as

$$\hat{\mathcal{A}} \equiv \arg \min_{x \in \mathbb{R}^d, a \in \mathbb{R}^q} p'x + w(n_1, \mathcal{D}^{(1)})'a, \quad \text{s.t.: } a \geq \hat{c}^{(1)} - \hat{M}^{(1)}x, \ a \geq 0.$$

2: Find the (finite) set of vertex-solutions $\hat{\mathcal{V}}_x \equiv \{x \in \mathbb{R}^d : (x, a) \in \hat{\mathcal{A}}, \ \text{rk}\left(\hat{M}^{(1)}_{J(x;\hat{\theta}^{(1)})}\right) = d\}$

3: Find the optimal vertex-solution $\hat{x} \in \arg\max_{x \in \hat{\mathcal{V}}_x} p'x$

4: Compute the set of binding inequalities $\hat{A} \equiv J(\hat{x}; \hat{\theta}^{(1)})$

5: Compute $\check{v} = \arg\min_{v \in \mathbb{R}^{|\hat{A}|}} ||p - \hat{M}^{(1)\prime}_{\hat{A}} v||^2, \ \text{s.t. } ||v|| \leq \overline{v}$

Step 3 (Construct the C.I.)

1: Compute $\hat{\sigma}^2_n = \sigma^2(\hat{A}, \hat{x}, \hat{v}, \hat{\Sigma}_n)$ using the formula in Lemma 6.6.

2: Compute an updated estimate $\check{B} \equiv \check{v}'(\hat{c}^{(2)}_{\hat{A}} - \hat{M}^{(2)}_{\hat{A}}\hat{x}) + p'\hat{x}$

3: The right, two-side and left $\alpha-$confidence intervals for $B(\theta_0)$ are given by

$$(\check{B} - \frac{\hat{\sigma}_n}{\sqrt{n_2}} z_{1-\alpha}; +\infty), \quad (\check{B} - \frac{\hat{\sigma}_n}{\sqrt{n_2}} z_{1-\alpha/2}; \check{B} + \frac{\hat{\sigma}_n}{\sqrt{n_2}} z_{1-\alpha/2}), \quad (-\infty; \check{B} + \frac{\hat{\sigma}_n}{\sqrt{n_2}} z_{1-\alpha/2})$$

Continuing the discussion in Section 2.1.b, we note that subsets A that satisfy (10) and (11) correspond to the nice faces (see p.12).

Lemma 2.3. If $A \subseteq [q]$ satisfies (10) and (11), then $F = \{x \in \mathbb{R}^d : M_A x = c_A\}$ is a nice face,

$$B(\theta_0) = p'x, \quad \forall x \in F.$$

Define the set $\mathbb{A} \equiv \{A \in 2^{\overline{1,q}} : |A| \geq d, A \text{ satisfies (10) and (11)}\}$. With probability approaching 1, the penalty function estimator manages to select a vertex-solution $\hat{x}_n \in \tilde{\mathcal{A}}(\hat{\theta}_n; w_n)$, determined by the binding constraints $\hat{A}_n = J(\hat{x}_n; \hat{\theta}_n) \in \mathbb{A}$ that satisfy (10) and (11) and thus correspond to a nice face by Lemma 2.3.

The debiased estimator may hence be understood as a two-stage procedure: one first finds the set of binding inequalities $\hat{A}_n \in \mathbb{A}$, and then estimates $B(\theta_0)$ as $p'(\hat{M}_{n\hat{A}_n})^\dagger \hat{c}_{n\hat{A}_n}$. Performing inference on that object directly requires working with the joint distribution of $\hat{A}_n$, and $\hat{\theta}_n$, leading to a complex and likely non-normal asymptotic distribution.

We address this by 'disentangling' the variation in $\hat{A}_n$ and $\hat{\theta}_n$ via sample splitting. Intuitively, a vertex is estimated on one part of the sample, while the noise in the parameter estimation comes from the other. We now state our assumptions and present the main result.

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

Assumption B2 requires the researcher to possess a consistent estimator of the asymptotic variance of $\hat{\theta}_n$. If $\hat{\theta}_n = g\left(n^{-1}\sum_{i=1}^n W_i\right)$ for some smooth and known $g(\cdot)$, such estimator can typically be obtained from the estimated covariance matrix of W_i using via Delta-method. In more complicated scenarios, bootstrap on $\hat{\theta}_n$ may be employed.

Define the set $\mathcal{S}_A \equiv \{v \in \mathbb{R}^{|A|} : p = M_A' v\}$ and note that (11) is equivalent to $\mathcal{S}_A \neq \emptyset$.

Assumption B3. *For a constant* $\overline{v} > 0$, $\max_{A \in \mathbb{A}} \min_{v \in \mathcal{S}_A} ||v|| \leq \overline{v}$.

Assumption B3 is a technical condition ensuring that we can find a sequence approaching $\mathcal{S}_A$ asymptotically, i.e. $d(\check{v}, \mathcal{S}_A) = o_p(1)$ for $\check{v}$ defined in (12). Practical guidance on choosing $\overline{v}$ is provided in the Appendix. Unlike the penalty parameter w_n, our simulations suggest that the specific choice of $\overline{v}$ has little impact on inference, as long as it is sufficiently large.

Definition (Optimal triplet)**.** We call $(A, x, v) \in 2^{[q]} \times \mathbb{R}^d \times \mathbb{R}^q$ an optimal triplet if i) $|A| \geq d$, ii) $x \in \mathcal{A}(\theta_0)$, iii) $M_A x = c_A$, iv) $p = M_A' v_A$, and v) $A = \text{Supp}(v)$.

We randomly split $\mathcal{D}_n$ into two disjoint, collectively exhaustive folds $\mathcal{D}_n^{(f)}$ of size n_f for $f = 1, 2$, with $n_1 = \lfloor \gamma n \rfloor$ and $n_2 = n - \lfloor \gamma n \rfloor$ for some fixed $\gamma \in (0; 1)$. Our inference procedure uses the data from $\mathcal{D}^{(1)}$ to estimate an optimal triplet $(\hat{A}, \hat{x}, \hat{v})$. The vertex[14] is estimated as

$$\hat{x} \in \arg \max_{x \in \tilde{\mathcal{A}}(\hat{\theta}^{(1)}; w_{n_1})} p'x \quad \text{s.t.:} \quad \text{rk}(\hat{M}_{J(x;\hat{\theta}^{(1)})}) = d,$$

the set of binding constraints that define it is denoted by $\hat{A} \equiv J(\hat{x}; \hat{\theta}^{(1)})$, and

$$\check{v} \in \arg \min_{||v|| \leq \overline{v}} ||\hat{M}_{\hat{A}}^{(1)\prime} v - p||^2. \tag{12}$$

Our procedure is then based on showing that, for large n,

$$\sqrt{n_2}\left(\check{v}'(\hat{c}_{\hat{A}}^{(2)} - \hat{M}_{\hat{A}}^{(2)}\hat{x}) + p'\hat{x} - B(\theta_0)\right) \approx \mathcal{N}(0, \sigma^2(\hat{A}, \hat{x}, \hat{v}, \Sigma)),$$

where $\sigma^2(\cdot)$ is derived in the Appendix, and $\hat{v} \in \mathbb{R}^q$ satisfies $\hat{v}_{\hat{A}} = \check{v}$ and $\hat{v}_j = 0$ for $j \notin \hat{A}$.

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

[14] While we assume that $\tilde{\mathcal{A}}(\hat{\theta}^{(1)}; w_{n_1})$ is estimated precisely, the results do not change if one is only able to estimate a single optimum. This may occur if numerical errors do not allow the LP-solver to find all of the LP solutions. Such optimum will satisfy $\text{rk}(\hat{M}_{J(x;\hat{\theta}^{(1)})}) = d$ by definition, and so will be a valid vertex-solution.

An inspection of the proof of Theorem 2.3 below reveals that Assumption B4 rules out the scenarios when finding a $A \in \mathbb{A}$ exactly determines the value $B(\theta_0)$. This may occur, for example, if the corresponding $\hat{c}_A^{(2)}, \hat{M}_A^{(2)}$ are deterministic. In this case, if A is also unique, meaning $|\mathbb{A}| = 1$, the debiased estimator has 0 asymptotic variance, because A and therefore $B(\theta_0)$ are correctly estimated with probability approaching approaching 1.

Theorem 2.3. Suppose $\mathcal{A}(\theta_0) \subseteq \text{Int}(\mathcal{X})$ and Assumptions B1, B3, B4 hold. Moreover,

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

for any optimal triplet (A, x, v) with $||v|| \leq \overline{v}$, which holds for $\hat{\sigma}_n(A, x, v) = \sigma(A, x, v, \hat{\Sigma}_n)$ under Assumption B2. Then, for any $\alpha \in (0; 1)$, and any $w_n \to \infty$ w.p.a.1 such that $w_n = o_p(\sqrt{n})$,

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

Remark 2.10. Following Gafarov (2024), one can drop Assumption B4 by using $\max\{\hat{\sigma}_n(\cdot), \underline{\sigma}\}$ for some small $\underline{\sigma} > 0$ instead of $\hat{\sigma}_n(\cdot)$ in Theorem 2.3. In that case, the test would have a correct level, but potentially conservative size.

Remark 2.11. If an estimator $\hat{\Sigma}_n$ is not available, one may alternatively compute the quantiles by performing bootstrap on $\sqrt{n}_2\hat{v}_{\hat{A}}(\hat{c}_{\hat{A}}^{(2)} - c_{\hat{A}} - (\hat{M}_{\hat{A}}^{(2)} - M_{\hat{A}})\hat{x})$, where $\hat{A}, \hat{x}, \hat{v}$ are fixed, while $\hat{c}^{(2)}$ and $\hat{M}^{(2)}$ are bootstrapped by resampling from the second fold $\mathcal{D}^{(2)}$.

2.3. Uniform asymptotic theory

The optimization problem (1) is challenging to study under no further assumptions, since it may feature instability with respect to arbitrary parameters' perturbations. We now show that this not only leads the plug-in $B(\hat{\theta}_n)$ to fail, but also precludes the existence of uniformly consistent estimators in general. The following auxiliary result establishes that there exists no uniformly consistent estimator for any real-valued functional from a space of probability measures if it is discontinuous in the total variation norm.

Lemma 2.4. Suppose a functional $V : (\mathcal{P}, ||\cdot||_{TV}) \to (\mathbb{R}, |\cdot|)$ is discontinuous at $\mathbb{P}_0 \in \mathcal{P}$. Then, there exists no uniformly consistent estimator $\hat{V}_n = \hat{V}_n(X)$, which is a sequence of measurable functions of the data $X \sim \mathbb{P}^n$. Moreover, if $\delta > 0$ is the jump at $\mathbb{P}_0$, then

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

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

In this section, we treat the parameter θ_0 as a functional of the underlying probability measure $\mathbb{P} \in \mathcal{P}$. We then make the following assumption on the pair $\theta_0(\cdot), \mathcal{P}$:

Assumption U0 (Uniform setup). *The functional $\theta_0(\cdot)$ and the set of probability measures $\mathcal{P}$ are such that: i) $\theta_0 : (\mathcal{P}, ||\cdot||_{TV}) \to (\mathbb{R}^S, ||\cdot||_2)$ is continuous; ii) $\theta_0(\mathcal{P}) = \{y \in \mathbb{R}^S$ s.t. $\Theta_I(y) \neq \emptyset, \Theta_I(y) \subseteq \mathcal{X}\}$ for a known and fixed compact $\mathcal{X}$*

Assumption U0 defines what is meant by 'the unrestricted set of measures'. U0.i demands that the true parameter be continuous in P, which holds, for example, in AICM models (see Section 3, Assumption I0). U0.ii assumes that θ_0 has full support over $\mathcal{P}$, meaning that any θ generating a non-rejectable model with $\Theta_I(\theta) \subseteq \mathcal{X}$ is attained at some $P \in \mathcal{P}$.

Theorem 2.4. Under U0 there exists no uniformly consistent estimator $\hat{B}_n$ of $B(\theta_0)$.

Proof. Combining U0, Lemma 2.4 and trivially extending the example in (5) to $\mathbb{R}^d$. ■

Given this negative result, it seems natural to seek a minimal restriction on $\mathcal{P}$ for which a uniformly consistent estimator may exist. We now show that the condition ensuring uniform consistency of the penalty function approach can be considered minimal in the sense to be made precise in Proposition 2.6.

Remark 2.4 illustrates that w_n cannot be allowed to diverge faster than $\sqrt{n}$, as otherwise the penalty approach may fail at measures where SC fails. At the same time, if all KKT vectors $\lambda^* \in \Lambda(\theta_0)$ grow large, an arbitrarily large w is needed for Assumption A1 to hold. This occurs when optimal vertices become 'sharp', i.e. all relevant full-rank submatrices of binding inequality constraints grow closer to being degenerate. The condition that ensures uniform consistency of the penalty function approach should therefore bound such 'sharpness'.

We begin with an existence result based on the Caratheodory's Conical Hull Theorem.

Proposition 2.5. *The problem (4) admits a solution x^* and the associated KKT vector λ^* such that for some index subset $J^* \subseteq \{1, \ldots, q\}$ with $|J^*| = d$, M_{J^*} is invertible and:*

$$
\begin{aligned}
x^* &= M_{J^*}^{-1} c_{J^*}, \\
\lambda^*_{J^*} &= M_{J^*}^{-1\prime} p, \\
\lambda^*_i &= 0 \text{ if } i \notin J^*.
\end{aligned}
$$

Proposition 2.5 asserts that any finite and feasible LP has an optimal vertex x^* at which there is a subset J^* of binding constraints, such that i) the corresponding gradients form a full-rank square matrix, and ii) the objective function gradient belongs to the conical hull formed by the gradients of the constraints from J^*.

Assumption U1 (δ-condition). *The class of measures $\overline{\mathcal{P}}$ satisfies the $\delta-$condition for a given $\delta > 0$, if*

$$\inf_{\mathbb{P}\in\overline{\mathcal{P}}} \max_{J^*\in\mathcal{J}^*(\theta(\mathbb{P}))} \sigma_d(M_{J^*}(\mathbb{P})) > \delta, \tag{13}$$

where $\mathcal{J}^*(\theta(\mathbb{P}))$ collects all J^* defined in Proposition 2.5 at a given $\theta(\mathbb{P})$.

The δ-condition does not rule out the failure of LICQ, SC or NFF, and is weaker than the conditions usually imposed to establish uniform consistency of LP estimators. To formalize this, let us introduce three families of measures. Firstly, denote the family of measures satisfying U1 for a given $\delta > 0$ by $\mathcal{P}^\delta$. A measure satisfies the Slater's condition if $\mathbb{P} \in \mathcal{P}^{SC} \equiv \{\mathbb{P} \in \mathcal{P} | \text{Int}(\Theta_I(\theta(\mathbb{P}))) \neq \emptyset\}$. Similarly, a measure satisfies a uniform ε-LICQ condition (as in Gafarov (2024)) if $\mathbb{P} \in \mathcal{P}^{LICQ;\varepsilon}$, where

$$\mathcal{P}^{LICQ;\varepsilon} \equiv \{\mathbb{P} \in \mathcal{P} | M(\mathbb{P})_A \in \mathbb{R}^{d\times d}, \sigma_d(M(\mathbb{P})_A) > \varepsilon \ \forall A \in \mathcal{V}(\mathbb{P})\},$$

where the set $\mathcal{V}(\mathbb{P}) \subseteq 2^{[q]}$ consists of sets indices of binding inequalities that define vertices of the polytope $\Theta_I(\theta(\mathbb{P}))$.

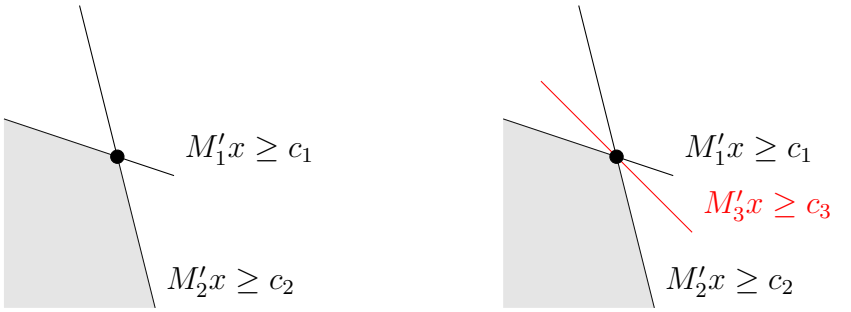
Proposition 2.6. *The following hold:*

1. *$\mathcal{P}^{Slater} \cup \mathcal{P}^{LICQ;0} \subset \mathcal{P} = \bigcup_{\delta>0} \mathcal{P}^\delta$, where the inclusion is strict*
2. *$\mathcal{P}^{LICQ;\varepsilon} \subset \mathcal{P}^\delta$ for any $\delta \leq \varepsilon$, where the inclusion is strict*

Proof. Part 1 follows from a trivial extension of the example in (5) to $\mathbb{R}^d$ and Proposition 2.5. Part 2 is true by definition of $\mathcal{P}^\delta, \mathcal{P}^{LICQ;\varepsilon}$. ■

Intuitively, the $\delta > 0$ in Assumption U1 merely parametrizes the degree of irregularity that the researcher is willing to allow for the identified polytope over the considered set of measures. The resulting family of sets 'covers' the unconstrained set of measures asymptotically as δ decreases to 0. Uniform LICQ and SC, on the contrary, both restrict the set of measures. That is because measures like the one in Figure 2b do not belong to either $\mathcal{P}^{LICQ;\varepsilon}$ for any $\varepsilon \geq 0$, or $\mathcal{P}^{SC}$.

Example 2.1. Figure 4 plots the range of $b-$values that correspond to the measures satisfying the $\delta-$condition for a given $\delta > 0$ in the example (5). The $\delta-$condition cuts off an interval of b along which the optimal vertex becomes 'too sharp'. Recall that the case $b = 0$ leads to the failure of SC. However, it satisfies the $\delta-$condition for a relatively large δ, because at the optimum $x^* = (-1 \ -1)'$ there is a set $J^* = \{1, 3\}$ from Proposition 2.5 such that the relevant matrix of binding constraints M_{J^*} has the smallest singular value $\sigma_2(M_{J^*}) \approx 0.62 \gg 0$.



(a) $J = \{1, 2\}$, LICQ holds (b) $J = \{1, 2, 3\}$, LICQ fails

Figure 3: Illustrating the difference between LICQ and $\mathcal{P}^\delta$. In (a) and (b) the $\delta-$condition holds with the same $\delta = \sigma_2(M_{\{1,2\}}) \gg 0$.

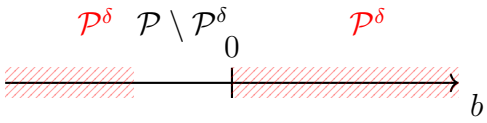


Figure 4: The set of b in problem (5) satisfying a $\delta-$condition.

Example 2.2. There are LPs in which SC, LICQ and NFF all fail, but the $\delta-$condition is satisfied for a relatively large δ. One example is the problem:

$$\min -x_1 + x_2, \quad \text{s.t.} \quad x_2 \leq x_1, x_2 \geq x_1, x_1 \in [-1; 1].$$

The $\delta-$condition is satisfied for $\delta > \sigma_2(M_{\{1,3\}}) \approx 0.62$, which is large relative to the penalty we suggest (see Section 2.6). There are flat faces, as any pair with $x_2 = x_1$ and $x_1 \in [-1; 1]$ is a solution, SC fails, as $\text{Int}(\Theta_I) = \emptyset$, and LICQ fails, as at an optimal $x_2 = x_1 = -1$ there are three binding constraints.

2.4. Uniform consistency of penalty function estimator

Theorem 2.5. Suppose i) $\hat{\theta}_n = \hat{\theta}_n(\mathbb{P})$ converges to $\theta_0(\mathbb{P})$ a.s. uniformly over $\mathcal{P}$ at rate $\overline{r}_n \uparrow \infty$, i.e. for all $r_n \uparrow \infty$ with $r_n = o(\overline{r}_n)$ and any $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}[\sup_{m\geq n} r_m ||\hat{\theta}_m - \theta(\mathbb{P})|| \geq \varepsilon] = 0, \tag{14}$$

and ii) $w_n(\mathbb{P}) = w_n \to \infty$ w.p.a.1 a.s. uniformly over $\mathcal{P}$, i.e. for any $M > 0$,

$$\lim_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}[\inf_{m\geq n} w_m > M] = 1.$$

Then, for all $r_n \uparrow \infty$ with $r_n = o(\frac{\overline{r}_n}{w_n})$ and any $\varepsilon > 0$,

$$\sup_{\delta>0} \lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}^\delta} \mathbb{P}[\sup_{m\geq n} r_m ||\tilde{B}(\hat{\theta}_m; w_m) - B(\theta_0(\mathbb{P}))|| \geq \varepsilon] = 0. \tag{15}$$

Example 2.3. In the AICM models (3), θ_0 is linear in moments of interactions of $Y(t)$ with treatment indicators and linear or hyperbolic in probabilities of $\{T = t, Z = z\}$ (see Section 3). Thus, condition ii) in Theorem 2.5 is established by, firstly, imposing that:

$$\lim_{C\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} ||\mathbb{Y}|| \mathbb{I}\{||\mathbb{Y}|| > C\} = 0, \quad \mathbb{Y} \equiv (Y(t))_{t\in\mathcal{T}},$$

which holds whenever bounded outcomes are assumed. If also a full-support condition holds:

$$\inf_{\mathbb{P}\in\mathcal{P},\ (t,z)\in\mathcal{T}\times\mathcal{Z}} \mathbb{P}[T = t, Z = z] > C$$

for some $C > 0$, then $\theta(\cdot)$ is uniformly continuous in population moments. Combining this with the LLN uniform in probability measure (see Proposition A.5.1 on p. 456 of Van Der Vaart et al. (1996)) yields condition (ii). If one additionally assumes that:

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

the rate in (14) is $r_n = \sqrt{n}$ (see Proposition A.5.2 on p. 457 of Van Der Vaart et al. (1996)).

Remark 2.12. The existence of a uniformly consistent estimator over $\mathcal{P}$ implies that $B(\cdot)$ is continuous over $\theta_0(\mathcal{P})$. However, as Proposition (2.1) illustrates, it is not necessarily continuous over the support of $\hat{\theta}_n$. It is straightforward to see that the case $b = 0$ in (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 estimator

We need to introduce the following two objects, which we term the face and polytope condition numbers.

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

$$\tilde{\kappa}(f) \equiv \min_{B \subseteq A:\ \text{rk}(M_B)=d-k} \sigma_{d-k}(M_B)$$

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

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

To see why the second equality above holds, note that 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) $f^* \subseteq f$, so the condition number of f is greater or equal than that of f^* by the interlacing inequality for singular values.

Assumption U2 (Polytope δ-condition). *The class of measures $\overline{\mathcal{P}}$ satisfies the polytope $\delta-$condition for a given $\delta > 0$, if:*

$$\inf_{\mathbb{P} \in \overline{\mathcal{P}}} \kappa(\Theta_I(\mathbb{P})) \geq \delta$$

The polytope $\delta-$condition thus lower bounds the smallest singular values of all full-rank matrices that can be constructed from the vertices of the polytope. Similarly to Assumption U1, it parametrizes the unconstrained set of probability measures, since at any fixed $\mathbb{P} \in \mathcal{P}$ we have $\kappa(\Theta_I(\mathbb{P})) > 0$ by definition. Proposition 2.6 continues to hold for U2, if one substitutes $\mathcal{P}^\delta$ with $\mathcal{P}_p^\delta$ - the family of measures satisfying U2 for a given $\delta > 0$.

The following 'anticoncentration' Lemma appears to be mathematically novel.

Lemma 2.5. For any non-empty and bounded polytope $\Theta = \{x \in \mathbb{R}^d | Mx \geq c\}$:

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

By Theorem 2.5, the biased penalty estimator is uniformly consistent at rate $\frac{\sqrt{n}}{w_n}$ under the δ-condition. The following Theorem asserts that the debiased estimator converges at least at the same rate.

Theorem 2.6. Suppose $\hat{\theta}_n$ converges to $\theta(\mathbb{P})$ a.s. uniformly over $\mathcal{P}$, i.e. for any $\varepsilon > 0$:

$$\lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}[\sup_{m\geq n} r_m ||\hat{\theta}_m - \theta(\mathbb{P})|| \geq \varepsilon] = 0$$

Then:

$$\sup_{\delta>0} \lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_p^\delta} \mathbb{P}[\sup_{m\geq n} \frac{r_m}{w_m} ||\hat{B}(\hat{\theta}_m; w_m) - B(\theta(\mathbb{P}))|| \geq \varepsilon] = 0$$

It is unclear if the $\frac{\sqrt{n}}{w_n}$ rate is uniformly sharp. Our simulation evidence (see the Appendix) suggests that this may be the case, but only along the sequences of measures for which SC, LICQ and NFF all fail in the limit. Once such sequences are ruled out, the rate of $\sqrt{n}$ appears to be restored uniformly.

2.6. Monte Carlo

Consistency and inference simulations use 10^4 and 10^3 repetitions respectively.

2.6.a. *Consistency.* We first describe the example from Proposition 2.1 in more detail. Consider a setup with $d = 2$ variables and $q = 4$ constraints. Recall that $\theta = (p', \text{vec}(M)', c')'$. In this case, the parameters are defined as

$$p = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad M = \begin{pmatrix} -(1+b) & 1 \\ 1 & -1 \\ 1 & 0 \\ -1 & 0 \end{pmatrix}, \quad c = \begin{pmatrix} 0 \\ 0 \\ -1 \\ -1 \end{pmatrix}. \tag{16}$$

The sample analogues $\hat{M}_n, \hat{c}_n$ of parameters in (16) are obtained by substituting b with its estimator $\hat{b}_n$. In our simulation study, that estimator is given by $\hat{b}_n = b + n^{-1} \sum_{i=1}^n U_i^b$, where $U_i^b \sim U[-1; 1]$ for $i \in [n]$ are i.i.d.

The true parameter b is in one-to-one correspondence with the underlying measure, and θ is completely described by it: $\theta(\mathbb{P}) = \theta(b(\mathbb{P}))$. Consequently, it is sufficient to index the parameter θ, the identified polytope Θ_I and the program's value B by b only.

Observe that the example (16) is so engineered that $\Theta_I(\hat{b}_n)$ would never be empty or unbounded, ensuring the existence of the plug-in estimator $B(\hat{b}_n)$. A slight modification to that setup, that we consider in the next subsection, would lead the identified polytope to be empty with a potentially non-vanishing probability.

Measures We consider two values of b: $b = 0$ and $b = -0.05$. At $b = 0$, SC fails, because $\Theta_I(0)$ has an empty interior. LICQ also fails at $b = 0$, as at the optimum $x = (-1, -1)$,

inequalities $1, 2, 4$ are binding. There are no flat faces at $b = 0$. At $b = -0.05$, SC, LICQ and NFF all hold. The values $b = 0$ and $b = -0.05$ result in the smallest singular values of $M_{J^*(b)}(b)$ matrices that correspond to the 75−th and the 19−th percentiles of the Tao-Vu distribution given in Theorem 6.1.

If $b < 0$, the norm of the 'smallest' KKT vector λ in the true LP corresponding to (16) is proportional to $|b^{-1}|$. So, for a small negative b, the δ-condition is only satisfied for a small value of $\delta > 0$. In this case the 'optimal' w is large. By that logic, the plug-in estimator, which in this case obtains as the limit $w \to \infty$, should perform well at such b, and potentially outperform the debiased penalty estimator. This is partly an artifact of the setup in (16): additional noise in one of the slated lines' intercepts would render the problem infeasible with positive probability, which would worsen the performance of the plug-in estimator in finite samples (see Figure 8).

At $b \geq 0$, the δ-condition is satisfied with a relatively large $\delta = \frac{\sqrt{5}-1}{2}$, and so the 'optimal' w is relatively small. If $b = 0$, in 50% of the cases, $\hat{b}_n < 0$ is estimated. If, moreover, $w_n > \text{const} \times |\hat{b}_n^{-1}|$, the debiased penalty estimator would select the incorrect maximum of 0 in such cases. A larger w at $b = 0$ thus hampers the performance of the debiased penalty estimator.

Parameters We set $w_n = \delta_{0.2}^{-1}(d)||p|| \frac{\ln \ln n}{\ln \ln 100}$, where $\delta_\alpha(d)$ is the $\alpha-$quantile of the $d-$dimensional Tao-Vu distribution given in Theorem 6.1 in the Appendix. To ensure the same expansion rate for the set-expansion estimator, we set $\sqrt{\kappa_n} = \kappa_0 \times \ln \ln n$. There is no guidance as to the selection of κ_0. Our baseline is $\kappa_0 = 0.1$, and we explore other values in Figure 7.

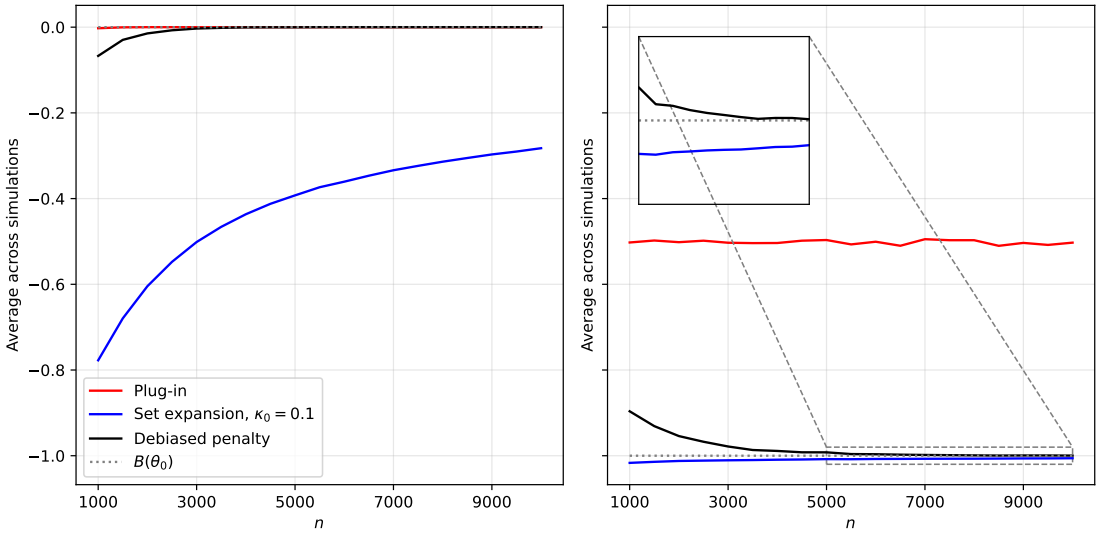


Figure 5: Simulation of example in (16) for $b = -0.05$ and $b = 0$ respectively.

Discussion Consider the right panel of Figure 5, corresponding to $b = 0$. The plug-in estimator is inconsistent, while the set-expansion estimator performs well, approaching -1 from below for larger n. The debiased penalty-function estimator is slightly upward-biased for smaller n, but yields the value of almost exactly -1 in larger samples. In contrast, the set-expansion estimator has a conservative rate, and remains slightly downward-biased even in larger samples.

The case of $b = -0.05$ is depicted in the left panel of Figure 5. The plug-in estimator is consistent and appears to be the best estimator out of the three. The set-expansion estimator is severely downward-biased. This is because when the optimal vertex has a 'sharp' angle, a small expansion of the inequalities' RHS may lead to a large shift of the vertex. To see that, consider shifting both inequalities outwards in Figure 12 for a small and a large absolute value of b, when b is negative. Once the expansion grows smaller, the set-expansion estimator slowly converges to the true value of 0 from below. While selecting a smaller κ_0 parameter would improve the performance of the set-expansion estimator at $b = -0.05$, in the next part of our analysis we demonstrate that in this example $\kappa_0 = 0.1$ is close to being optimal in the uniform sense, because smaller κ_0 worsens the estimator's performance at $b = 0$. The debiased penalty estimator, in contrast, converges rather quickly. It is slightly conservative at smaller n, as it selects the incorrect vertex of -1 whenever $\hat{b}_n < 0$ and $|\hat{b}_n^{-1}| > \text{const} \times w_n$, i.e. when the penalty parameter is not large enough.

Robustness of the debiased penalty estimator For $b < 0$, the debiased estimator may be expected to perform better at measures with a larger $|b|$. Figure 7 illustrates that point:

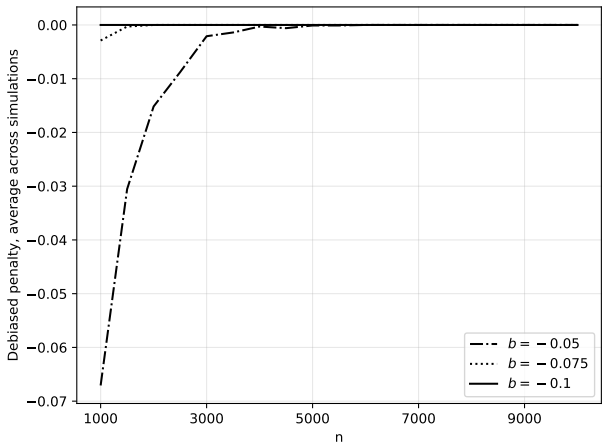


Figure 6: Performance of the debiased penalty estimator for different b in example (16).

Researchers applying any LP estimator should exercise caution when operating in smaller samples due to irregularity inherent in 4. In example (16), the debiased penalty estimator exhibits desirable behavior even along highly irregular measures for sample sizes of order

$n = 5000$. Such sample sizes are not uncommon in partially identified settings. In our application, estimation is performed on 664633 observations.

Alternative κ_0 We now investigate to which extent the behavior of the set-expansion estimator can be improved by selecting an alternative κ_0 parameter. Clearly, a larger κ_n makes the set-expansion estimator more conservative. However, κ_n cannot be selected as small is possible. If the expansion is too small, it may be insufficient to counteract the noise involved in estimating Θ_I, leading to poor performance of the estimator at measures where SC fails to hold.

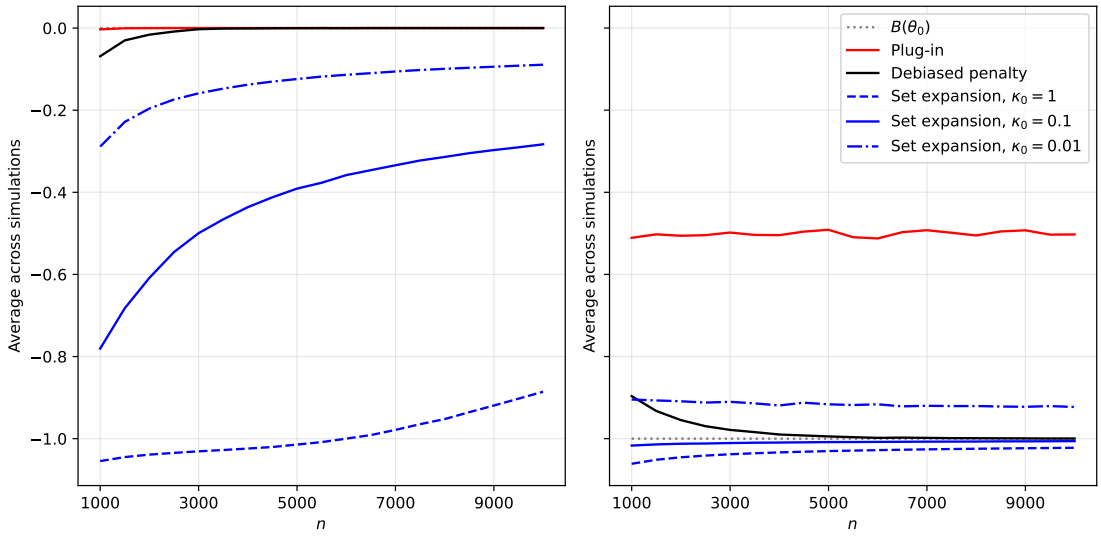


Figure 7: Simulation of (16) for $b = -0.05$ and $b = 0$ resp., different κ_0.

That tradeoff is very clear in Figure 7. Compare the performance of the estimator with $\kappa_0 = 0.01$ to the baseline of $\kappa_0 = 0.1$. Decreasing κ_0 down to 0.01 makes the estimator less conservative at $b = -0.05$, but results in an upward bias at $b = 0$. This occurs, because at $\kappa_0 = 0.01$ the resulting set-expansion sequence κ_n is too small to counteract the estimation noise for the considered sample sizes. This logic is 'monotone', meaning that selecting an even smaller κ_0 would worsen the performance at $b = 0$ further. Even at $\kappa_0 = 0.01$, the set-expansion estimator is quite conservative for the measure $b = -0.05$, and the parameter can clearly not be reduced any further without affecting the validity of the estimator at $b = 0$. It appears that the baseline choice of $\kappa_0 = 0.1$ is close to being optimal in our example. Therefore, the conservative behavior of the set-expansion estimator at $b = -0.05$ is not explained by a poor choice of the tuning parameter, but rather is a feature of the estimator itself.

Noise in $\hat{c}_n$ We also report the result of consistency simulations for the DGP described in (17). This DGP features noise on the right-hand-side of the inequalities that describe the polytope, which allows the estimated polytope to be empty.

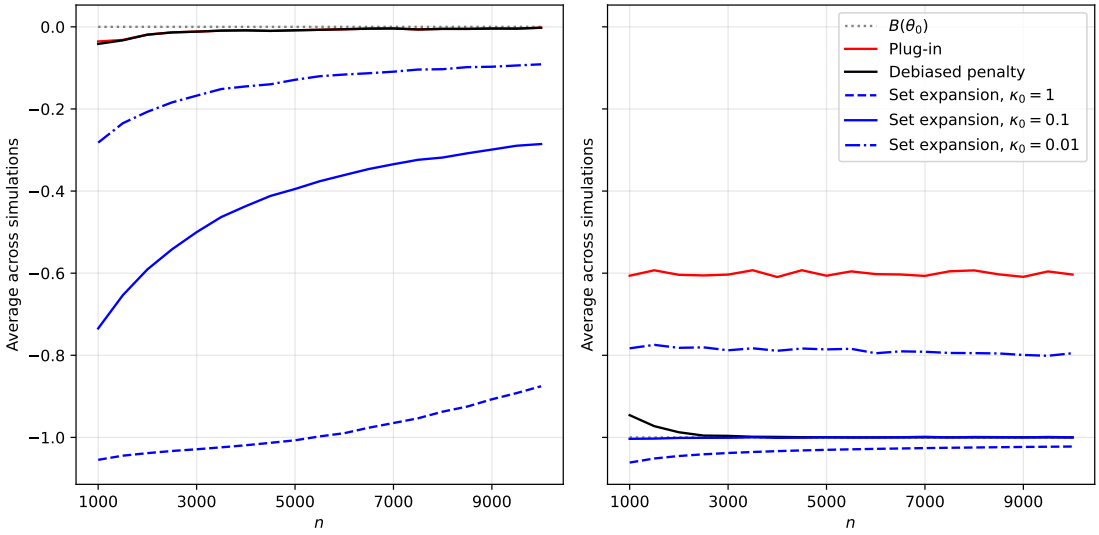


Figure 8: Simulation of (17) for $b = -0.05$ and $b = 0$ resp., different κ_0. Averages for the plug-in and set-expansion estimators ignore failed iterations.

In Figure 8, the plug-in and the debiased penalty estimators perform equally well at $b = -0.05$. The rest of the conclusions are qualitatively unchanged.

2.6.b. *Inference.* In this section, we assess the performance of our inferential procedure. We compare it to the performance of two recently developed methods, Cho and Russell (2023) (CR) and Gafarov (2024) (BG). We consider a slightly modified version of example (16):

$$M = \begin{pmatrix} -(1+b) & 1 \\ 1+\zeta & -1 \\ 1 & 0 \\ -1 & 0 \end{pmatrix}, \quad c = \begin{pmatrix} \nu \\ \zeta \\ -1-\nu \\ -1 \end{pmatrix} \tag{17}$$

The sample analogues $\hat{M}_n, \hat{c}_n$ of parameters in (17) are obtained by substituting b, ζ, ν with their estimators:

$$\hat{b}_n = b + n^{-1}\sum_{i=1}^{n} U_i^b, \quad \hat{\zeta}_n = n^{-1}\sum_{i=1}^{n} U_i^\zeta, \quad \hat{\nu}_n = n^{-1}\sum_{i=1}^{n} U_i^\nu,$$

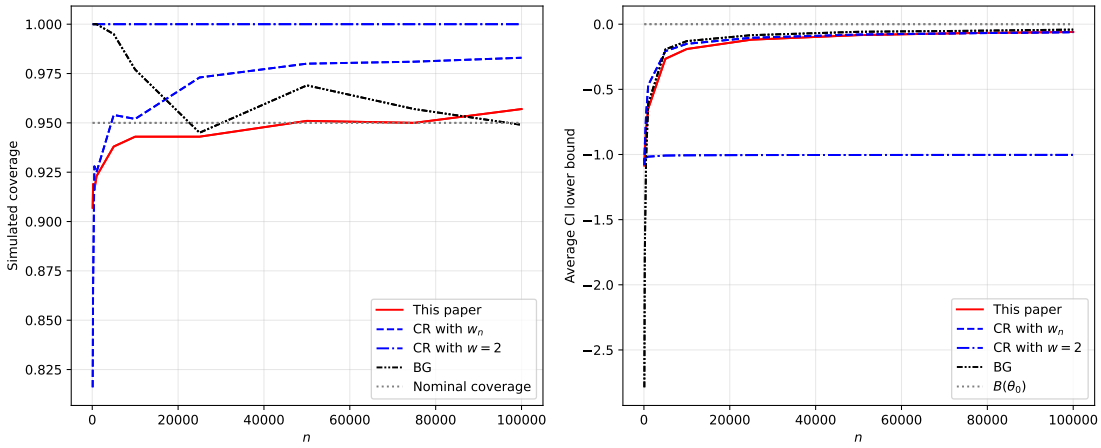
where $U_i^k \sim U[-0.5; 0.5]$, i.i.d. across $i \in [n]$ and $k \in \{b, \zeta, \nu\}$. We consider measures, for which the true values of $\zeta = \nu = 0$, whereas we still vary b as in the example before. As before, we mainly consider $b = -0.05$ and $b = 0$.

Note that in (17) the plug-in estimator of the polytope $\hat{\Theta}_I$ may be empty for some realizations of $\{U_i^k\}_{i,k}$. That is the case, for example, if $\hat{b}_n = \hat{\zeta}_n = 0$ and $\hat{\nu}_n > 0$.

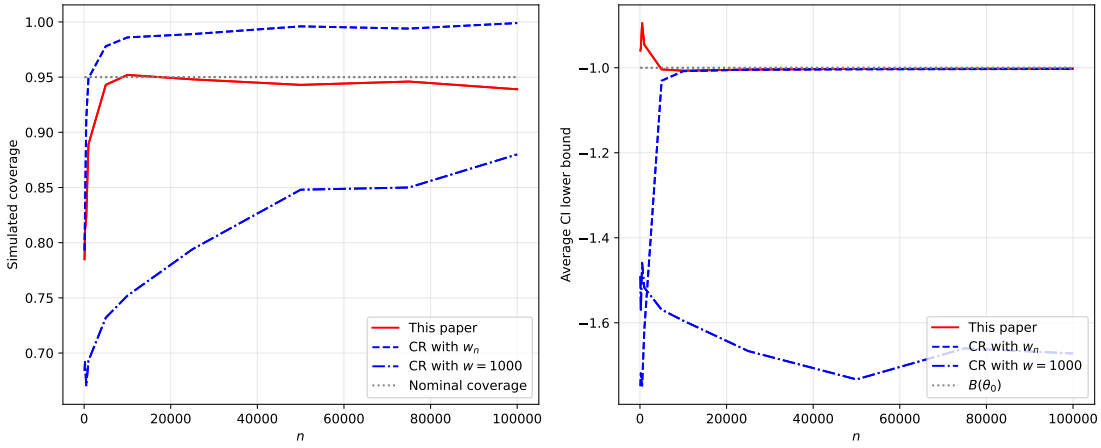
Other methods In case SC fails, the BG estimator is not applicable. It fails to exist in around 25% of cases at $b = 0$. The CR estimator in the main text of the paper also relies on $B(\hat{\theta}_n)$ and is not applicable if SC fails. The procedure described on p.47 of the Supplementary Appendix in Cho and Russell (2023) would not be practically applicable without the results contained in the present paper. It can be shown that the CR augmented procedure combines a random, non-vanishing set expansion and objective function perturbation with the penalty function approach[15]. The authors, however, treat the analogue of the penalty parameter as 'some [fixed] large value'. They proceed to argue that the inequalities, which establish the size of their inference procedure, hold for any value of w, which appears to suggest that its selection is unimportant. There is no practical guidance on w selection, and CR do not implement the augmented estimator in their simulations. In this paper, we studied penalized estimation in great detail, and our results suggest that the appropriate choice of w is critical. Both our previous findings and simulations in Figure 9 demonstrate that for different values of w the performance of the CR procedure ranges from highly conservative to invalid. Selecting 'a large value' of w does not yield a valid procedure in finite samples. Our simulations also suggest that CR augmented approach can perform relatively well if combined with our results on the rate and level of w_n. Unlike our approach, however, it remains asymptotically conservative due to the use of non-vanishing random expansions.

Implementation details As mentioned in Section 2.2, one can estimate the as. covariance matrix $\hat{\Sigma}_n$ of $\hat{\theta}_n$ with resampling. One then plugs it into the expression for $\sigma(\cdot)$ to obtain the required s.e. An even simpler approach is to compute $\sigma(\cdot)$ directly by bootstrapping the quantities estimated on the second fold, while keeping the first fold quantities fixed. We have verified that the performance of the two approaches is similar to using the closed-form estimator for Σ, namely $\hat{\Sigma}_n = G\hat{\Omega}_n G'$, where $G \equiv \frac{\partial \theta}{\partial (b \ \zeta \ \nu)'}$ and $\hat{\Omega}_n$ is the sample covariance matrix of $(b_i \ \zeta_i \ \nu_i)'$. We employ the latter estimator in our simulations. When implementing the procedure in Cho and Russell (2023) (CR), we use uniform noise with the support size of 0.001, as recommended in the paper. Note that we refer to their parameter M as w. The estimator in Gafarov (2024) (BG) was implemented using the code kindly provided by the author.

[15] Because the penalty function estimator can be rewritten in the form of an equivalent problem w.p.a.1.



(a) $b = -0.05$. For BG, the number of failed simulations was $206, 158, 46, 6$ at $n = 100, 200, 500, 1000$ respectively, none at larger n.



(b) $b = 0$. BG omitted, as around 250 simulations fail.

Figure 9: Performance of different inferential procedures over 1000 simulations of (17). Left panel - estimated coverage of a 95% one-sided C.I.; right panel - average lower confidence bound.

Discussion Results of our simulations are given in Figure 9. Overall, it appears that our estimator has the correct nominal level even at smaller sample sizes, whereas the CR with penalty w_n overcovers asymptotically. BG estimator achieves nominal coverage asymptotically, although it over-covers in smaller samples and may yield a very conservative left confidence bound, as is evident from the right panel. It fails at $b = 0$, because the SC fails[16]. For different fixed values of w the Cho-Russell procedure's performance may range from highly conservative to invalid. We illustrate this by adding lines with $w = 2$ and $w = 1000$ to the figures corresponding to $b = -0.05$ and $b = 0$ respectively.

[16]We have still run the corresponding simulations, but are not displaying them. BG fails to exist in around 25% of cases, while the remaining simulations result in highly conservative bounds with incorrect coverage

3. Special case of LP bounds

Let $Y \in \mathbb{R}$ denote the outcome of interest[17], $T \in \mathbb{R}$ stand for the treatment, and $Z \in \mathbb{R}^{d_Z}$ be the candidate instrument. Here *treatment* is any variable which effect on Y we attempt to infer, whereas the term *instrument* refers to an auxiliary variable that allows us to partially identify the treatment effect of interest. Our approach nests the case when Z is the usual IV. The reader seeking an economic intuition may refer to the classical example from Manski and Pepper (2000), where Y is the wage, T is an indicator of educational degree and Z is the level of ability.

Denote the supports Y, T, Z as $\mathcal{Y}$, $\mathcal{T}$ and $\mathcal{Z}$ respectively. Throughout this section we consider the case of continuous outcomes and discrete treatment and instrument, namely $\mathcal{Y}$ is uncountable, while $N_T \equiv |\mathcal{T}| < \infty$ and $N_Z \equiv |\mathcal{Z}| < \infty$. In non-parametric bounds literature it is rather conventional to employ a discrete instrument at the estimation stage (see Manski and Pepper (2009)). While the 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. Namely, we split the set of treatments into two disjoint subsets $\mathcal{T} = \mathcal{O} \sqcup \mathcal{U}$. Whenever $T \in \mathcal{O}$, the researcher observes Y, T, Z, whereas if $T \in \mathcal{U}$, 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 $Y(t),\ t \in \mathcal{T}$:

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

Continuing the wages and education example, the value of $Y(t)$ for a fixed $t \in \mathcal{T}$ 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 $\mathbb{Y} \equiv (Y(t))_{t \in \mathcal{T}} \in \mathbb{R}^{N_T}$. Variables $(\mathbb{Y}, T, Z, Y)$ 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 set of considered measures throughout this section:

Assumption I0 (Conditions on $\mathcal{P}$). *$\mathcal{P}$ is such that $P \in \mathcal{P}$ if: i) P generates $F_{T,Z}(\cdot)$ and $\{F_{Y|T=t,Z}(\cdot)\}_{t \in \mathcal{O}}$; ii) $P[T = d, Z = z] > 0\ \forall d, z \in \mathcal{T} \times \mathcal{Z}$ and iii) $|\mathbb{E}_P[Y(t)|T = d, Z = z]| < \infty$ for all $z \in \mathcal{Z}$ and $t, d \in \mathcal{T}$*

Part i) of the Assumption I0 formalizes the assumed identification pattern. It says that the joint distribution of T, Z is always identified and the researcher also observes the joint

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

distribution of Y, T, Z whenever $T \in \mathcal{O}$. Parts ii) and iii) of I0 ensure that all conditional expectations and probabilities are well-defined and finite[18].

Remark 3.1. Under no missing data, i.e. $\mathcal{T} = \mathcal{O}$, condition i) is equivalent to P generating the identified joint distribution $F_{Y,T,Z}(\cdot)$.

We define the vector m collecting all elementary conditional moments as

$$m(P) \equiv (\mathbb{E}_P[\mathbb{Y}|T = d, Z = z])_{d\in\mathcal{T}, z\in\mathcal{Z}} \in \mathbb{R}^{N_T^2 N_Z},$$

and suppose that the researcher is interested in the target parameter β^*, given by

$$\beta^* = \mu^*(\mathbb{P})' m(\mathbb{P}), \tag{18}$$

where $\mu^* : \mathcal{P} \to \mathbb{R}^{N_T^2 N_Z}$ is *identified* and *chosen* by the researcher. It parametrizes the choice of the outcome of interest, as the following remark clarifies.

Remark 3.2. The form (18) nests i) $\mathbb{E}[Y(t)]$, ii) $ATE_{td} = \mathbb{E}[Y(t) - Y(d)]$ and iii) $CATE_{td,A,B} = \mathbb{E}[Y(t) - Y(d)|T \in A, Z \in B]$.

3.1. Affine inequalities over conditional moments

We now introduce the general class of identifying conditions described by affine inequalities over conditional moments, potentially augmented with affine a.s. restrictions. These restrict the set of admissible measures to $\mathcal{P}^*$:

$$\mathcal{P}^* \equiv \{P \in \mathcal{P} | (M^* m + b^*)(P) \geq 0 \wedge (\tilde{M}\mathbb{Y} + \tilde{b})(P) \geq 0 \quad P\text{-a.s.}\}, \tag{19}$$

where $b^* : \mathcal{P} \to \mathbb{R}^R$, $M^* : \mathcal{P} \to \mathbb{R}^{R \times N_T^2 N_Z}$, $\tilde{b} : \mathcal{P} \to \mathbb{R}^{\tilde{R}}$ and $\tilde{M} : \mathcal{P} \to \mathbb{R}^{\tilde{R} \times N_T}$ are *identified* parameters, *chosen* by the researcher. These parametrize the choice of $R \in \mathbb{N}$ identifying inequalities on conditional moments of potential outcomes as well as $\tilde{R} \in \mathbb{N}$ almost sure inequalities on the potential outcomes. In general, $\psi(P) \equiv (\mu^*(P), b^*(P), M^*(P), \tilde{b}(P), \tilde{M}(P))$ and $m(P)$ are functionals of P. We omit this dependence whenever it does not cause confusion.

The family of models that can be written in the form (19) is very rich, as illustrated by the following examples.

Example 3.1. MIV with $Z \in \mathbb{R}$ (Manski and Pepper, 2000) imposes that for each $t \in \mathcal{T}$ and $z, z' \in \mathcal{Z}$, $z' \geq z \implies \mathbb{E}[Y(t)|Z = z'] \geq \mathbb{E}[Y(t)|Z = z]$. It is nested for an appropriate choice of matrix $M^* = M_{MIV}$ and $b^* = 0$. MTS from (Manski and Pepper, 2000) obtains when $Z = T$.

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

Example 3.2. IV with $Z \in \mathbb{R}$ imposes that for each $t \in \mathcal{T}$ and $z, z' \in \mathcal{Z}$, $\mathbb{E}[Y(t)|Z = z'] = \mathbb{E}[Y(t)|Z = z]$. It is nested for an appropriate choice of matrix M_{IV} that can, for example, be constructed as $M^* = M_{IV} = \begin{pmatrix} M_{MIV} \\ M_{MIV} \end{pmatrix}$ and $b^* = 0$.

Example 3.3. MTR (Manski and Pepper, 2000) imposes that for each $t, t' \in \mathcal{T}$: $t' > t$, $Y(t') \geq Y(t)$ a.s. It is nested for an appropriate choice of matrix $\tilde{M} = \tilde{M}_{MTR}$ with $\tilde{b} = 0$.

Example 3.4. Roy model (Lafférs, 2019) imposes that for each $t \in \mathcal{T}$, the individual's choice is, on average, optimal $\mathbb{E}[Y(t)|T = t, Z = z] = \max_{d \in \mathcal{T}} \mathbb{E}[Y(d)|T = t, Z = z]$. 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 $Y(t) \equiv \mathbb{I}\{W \leq w\}$, so that $\mathbb{E}[Y(t)|X = x] = F(w|x)$. Our approach allows to accommodate all identifying conditions in the original paper by appropriately choosing M^*, b^* and $\tilde{M}, \tilde{b}$.

Remark 3.3. Combinations of assumptions are obtained by stacking the respective matrices, as in Example 3.2. Sensitivity analysis can be performed via relaxations $b^*_\ell = b^* + \ell$, or $\tilde{b}_\ell = \tilde{b} + \ell$ for some $\ell \geq 0$. For example, given some $\ell = \{\ell(t, z, z')\}_{t,z,z'} \geq 0$, inequalities $\mathbb{E}[Y(t)|Z = z'] - \mathbb{E}[Y(t)|Z = z] \geq -\ell(t, z, z')$ for $z, z' \in \mathcal{Z}$ with $z' > z$ and $t \in \mathcal{T}$ yield a relaxation of MIV. In De Haan (2017) the shape of observed moments may suggest a failure of monotonicity near the boundaries of Supp(Z). Selecting positive $\ell(z, z')$ for values of z, z' close to the boundaries could constitute a meaningful robustness check.

3.2. Linear programming bounds

We now provide the general identification result for the models described by $\mathcal{P}^*$. Let us construct x that collects unobserved pointwise-conditional moments and the vector $\overline{x}$ of those pointwise-conditional moments that are identified:

$$x \equiv (\mathbb{E}[Y(t)|T = d, Z = z])_{z \in \mathcal{Z} \wedge (t,d \in \mathcal{T}: t \neq d \vee t,d \in \mathcal{U}: t = d)},$$
$$\overline{x} \equiv (\mathbb{E}[Y(t)|T = t, Z = z])_{z \in \mathcal{Z}, t \in \mathcal{O}}.$$

For *known* selector matrices $P_m, \overline{P}_m$, one can then decompose m as

$$m = \underbrace{P_m x}_{\text{partially identified}} + \underbrace{\overline{P}_m \overline{x}}_{\text{identified}}.$$

It is also straightforward to observe that $\tilde{M}\mathbb{Y} + \tilde{b} \geq 0$ a.s. implies

$$(I_{N_T N_Z} \otimes \tilde{M})m + \iota_{N_T N_Z} \otimes \tilde{b} \geq 0. \tag{20}$$

Before we state the main identification result, let us construct the matrix M^{**} and the vector b^{**} that combine the conditional restrictions with the implications of the almost sure restrictions in (20), as

$$M^{**} \equiv \begin{pmatrix} I_{N_T N_Z} \otimes \tilde{M} \\ M^* \end{pmatrix}, \quad b^{**} \equiv \begin{pmatrix} \iota_{N_T N_Z} \otimes \tilde{b} \\ b^* \end{pmatrix}. \tag{21}$$

Theorem 3.1. Suppose Assumption I0 holds. For any ψ, the sharp identified set $\mathcal{B}^*$ for β^* satisfies

$$\mathcal{B}^* \equiv (\mu^{*\prime} m)(\mathcal{P}^*) \subseteq \{\beta \in \mathbb{R} | \inf_{x: Mx \geq b} p'x \leq \beta - \overline{p}'\overline{x} \leq \sup_{x: Mx \geq b} p'x\}, \tag{22}$$

where

$$\overline{p} \equiv \overline{P}'_m \mu^*, \quad p \equiv P'_m \mu^*, \quad M \equiv M^{**} P_m, \quad b \equiv -b^{**} - M^{**} \overline{P}_m \overline{x}.$$

The converse inclusion in (22) holds if $\tilde{M} = 0'_{N_T}$, $\tilde{b} = 0$, or:

1. MTR holds, i.e. $Y(t_1) \geq Y(t_0)$ a.s. $\forall t_1, t_0 \in \mathcal{T}$ s.t. $t_1 > t_0$:

$$\tilde{M} = \tilde{M}_{MTR}, \quad \tilde{b} = 0_{N_T - 1}$$

2. Outcomes are bounded, $Y(t) \in [K_0; K_1]$, $\forall t \in \mathcal{T}$ a.s. for known $K_1 > K_0$:

$$\tilde{M} = \tilde{M}_b \equiv \begin{pmatrix} I_{N_T} \\ -I_{N_T} \end{pmatrix}, \quad \tilde{b} = \tilde{b}_b \equiv \begin{pmatrix} -K_0 \cdot \iota_{N_T} \\ K_1 \cdot \iota_{N_T} \end{pmatrix}. \tag{23}$$

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

$$\tilde{M} = \begin{pmatrix} \tilde{M}_{MTR} \\ \tilde{M}_b \end{pmatrix}, \quad \tilde{b} = \begin{pmatrix} \tilde{b}_{MTR} \\ 0_{N_T - 1} \end{pmatrix}. \tag{24}$$

The matrix $\tilde{M}_{MTR}$ is defined in the Appendix.

Theorem 3.1 postulates that bounds on the target parameter β^* under $\mathcal{P}^*$ can be obtained by solving two linear programs. The LP bounds are sharp if there are no a.s. inequalities in the model, or if the a.s. inequalities parametrize three special cases that are typically used in the literature. Otherwise, the LP bounds may not be sharp, as we show in the Appendix. This is because, in general, the entire distribution of $\mathbb{Y}, T, Z$ is relevant for β^* under a.s. restrictions. The naive approach of searching over such joint distributions, however, would

involve infinite-dimensional optimization, because $|\mathcal{Y}| = \infty$.

Remark 3.4. We are not aware of affine a.s. restrictions $\tilde{M}, \tilde{b}$ used in applied work that are not special cases 1-3 in Theorem 3.1. The special cases 1-3 appear in Blundell et al. (2007), Kreider et al. (2012), Gundersen et al. (2012), Siddique (2013).

Remark 3.5. The empirical literature has extensively relied on the MIV + MTR + MTS combination of Manski and Pepper (2000) assumptions, as it yields the tightest bounds out of all classical conditions. In the absence of a theoretical justification, this has led to errors (Lafférs, 2013). Theorem 3.1 provides the first available sharp bounds under this combination when $|\mathcal{Y}| = \infty$.

4. Conditional monotonicity assumptions

A particular family of identifying conditions that can be written in the form (19) 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 $Y(t) \in [K_0; K_1]$ *a.s.* for known $K_0, K_1 \in \mathbb{R}$, $K_0 < K_1$. We also suppose that there are no missing data[19], i.e. $\mathcal{T} = \mathcal{O}$.

We argue that cMIV assumptions are reasonable in classical applications, discuss 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 imply 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 yields informative bounds even when the classical conditions fail to do so.

While we only discuss three variations of cMIV, the class of such assumptions is potentially richer[20], and Theorem 3.1 applies in any such framework.

Assumption cMIV-s. *Suppose that for any* $t \in \mathcal{T}$, $A \subseteq \mathcal{T} : A \neq \{t\}$ *and* $z, z' \in \mathcal{Z}$ *s.t.* $z' > z$ *we have:*

$$\mathbb{E}[Y(t)|T \in A, Z = z'] \geq \mathbb{E}[Y(t)|T \in A, Z = z],$$

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

[20] One can consider the class of conditional restrictions $\mathbb{E}[Y(t)|T \in A, Z = z'] \geq \mathbb{E}[Y(t)|T \in A, Z = z]$, $\forall A \in \mathcal{F}_t$ for all $t \in T$ where subcollections $\mathcal{F}_t \subseteq \mathcal{T}$ are chosen by the researcher.

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 considered cMIV conditions. To see that cMIV-s implies MIV, set $A = \mathcal{T}$ in the definition above.

Assumption cMIV-w. *Suppose MIV holds and for any $t \in \mathcal{T}$ and $z, z' \in \mathcal{Z}$ s.t. $z' > z$ we have:*

$$\mathbb{E}[Y(t)|T \neq t, Z = z'] \geq \mathbb{E}[Y(t)|T \neq t, Z = z],$$

i.e. the potential outcomes are, on average, non-decreasing in Z for the non-treated and 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, see the Appendix.

Assumption cMIV-p. *Suppose MIV holds and for any $t \in \mathcal{T}, d \in \mathcal{T} \setminus \{t\}$ and $z, z' \in \mathcal{Z}$ s.t. $z' > z$ we have:*

$$\mathbb{E}[Y(t)|T = d, Z = z'] \geq \mathbb{E}[Y(t)|T = d, Z = z],$$

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

The pointwise conditional monotonicity assumption is directly testable under a mild homogeneity condition, see Section 4.2.

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 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 4.1. 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 it is possible for the general apporach of form (19), 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. On the other hand, large violations of the observed outcomes' monotonicity will lead the test developed in Section 4.2 to reject cMIV-p and cMIV-s.

The following observation motivates the use of cMIV assumptions.

Proposition 4.1. *Manski and Pepper (2000) MIV bounds are not sharp under either cMIV-s, cMIV-w or cMIV-p.*

Proof. Consider a binary treatment T, three levels of the instrument $Z \in \{z_0, z_1, z_2\}$ with $z_0 < z_1 < z_2$ and $-K_0 = K_1 = 1$. Suppose for a fixed $t \in \{0, 1\}$, we have $\mathbb{E}[Y(t)|T = t, Z = z_i] = 0$, with $P[T \neq t|Z = z_0] = 0.125$, $P[T \neq t|Z = z_1] = 0.5$, $P[T \neq t|Z = z_2] = 0.25$. The no-assumptions lower bounds on $\mathbb{E}[Y(t)|Z = z_i]$ are $(-0.125,\ -0.5,\ -0.25)$. MIV 'irons' the no-assumptions bounds to $(-0.125,\ -0.125,\ -0.125)$, which also implies the lower bounds on $\mathbb{E}[Y(t)|T \neq t, Z = z_i]$: $(-1,\ -0.25,\ -0.5)$. Under cMIV, one can further 'iron' these to improve the lower bound for z_2 up to -0.25, so that the lower bound on $\mathbb{E}[Y(t)|Z = z_2]$ becomes $-1/16 > -1/8$. ■

Sharp bounds for all versions for cMIV follow from Theorem 3.1. 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 provide the detailed construction of the triplet M, c, p from Theorem 3.1 under cMIV-s and cMIV-p. All details on the identification under cMIV are provided in the Appendix.

4.1. Discussion of cMIV

This section illustrates the difference between MIV and cMIV by considering two parametric examples with classical applications.

4.1.a. *Education selection.* 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.

MIV assumption on Z implies that more able individuals can do better both with and without a college degree on average: $\mathbb{E}[Y(t)|Z = z]$ - 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: $\mathbb{E}[Y(0)|Z = z, T = 1]$ - 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: $\mathbb{E}[Y(1)|Z = z, T = 0]$ - 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. an IQ score), the employer additionally observes the employee effort level $\eta + \varepsilon$ with $\varepsilon \perp\!\!\!\perp (Z, T, \eta)$. Suppose $Var(Z) = Var(\eta) = 1$ and $\mathbb{E}[Z] = \mathbb{E}[\eta] = \mathbb{E}[\varepsilon] = 0$. Suppose that, on average, employees choose T to maximize their expected earnings. This

motivates a stylized Roy selection model with:

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

where $\nu \perp\!\!\!\perp (Z, \eta, \varepsilon)$ is remaining heterogeneity, and $\varepsilon(t) \equiv \beta_2(t)\varepsilon$. MIV demands that:

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

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 effort 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 differentials in the effects of ability and effort respectively, the additional requirement of cMIV is that:

$$\underbrace{\beta_1(0)z}_{\text{direct effect}} + \underbrace{\beta_2(0)\mathbb{E}[\eta|\delta_z z + \delta_\eta \eta + \tilde{\nu} \geq 0]}_{\text{selection given } T=1} - \text{increasing} \tag{25}$$

$$\underbrace{\beta_1(1)z}_{\text{direct effect}} + \underbrace{\beta_2(1)\mathbb{E}[\eta|\delta_z z + \delta_\eta \eta + \tilde{\nu} \leq 0]}_{\text{selection given } T=0} - \text{increasing}, \tag{26}$$

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 (25)-(26) 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[21]:

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

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

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

necessitates that:

$$\beta_1(0)z + \beta_2(0)\mathbb{E}[\eta|\eta \geq -z] - \text{increasing} \tag{27}$$
$$\beta_1(1)z + \beta_2(1)\mathbb{E}[\eta|\eta \leq -z] - \text{increasing} \tag{28}$$

In this case, conditional on each level of education, effort level η and ability Z are negatively associated, so the conditional selection terms in (27)-(28) 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 as her counterpart with a lower ability. Similarly, if an individual is capable, but does not have a degree, she has to be of lower effort as otherwise she would have selected into education.

Even if MIV holds, cMIV can thus fail if employer prefers effort over ability to the extent that the conditional negative association between the two 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 (25) and (26) 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*.

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:

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

The observed log-price P clears the market:

$$P \in \{p \in \mathbb{R}|\mathbb{E}[q^s(p)|Z, \eta] = \mathbb{E}[q^d(p)|Z, \eta]\}, \tag{29}$$

where η, Z are continuous unobserved and observed random variables respectively, with $\mathbb{E}[\eta|Z = z] = 0$[22] and $\mathbb{E}[\varepsilon^k] = 0$ with $\varepsilon^k \perp\!\!\!\perp (\eta, Z, \varepsilon^{-k})$ for $k \in \{s, d\}$. Further assume that all functions of p are continuous.

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

Potential price p indexes the potential outcomes, giving rise to the demand and supply *schedules*. Suppose we aim to identify the elasticity of supply, $\mathbb{E}[\frac{q^s(p_1)-q^s(p_0)}{p_1-p_0}]$ for some $p_1 > p_0$, and Z is a monotone instrument for $q^s(p)$, 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 $p \neq P$, 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 (29) 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:

$$P = \delta_p^{-1}(-\delta_z Z - \delta_\eta \eta) \tag{30}$$

Equation (30) 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:

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

In this setup, MIV requires:

$$(\text{MIV}): \ \beta^s(p) \geq 0, \ \forall p \in \mathbb{R}$$

Whereas cMIV-p additionally imposes that:

$$\beta^s(p)z + \gamma^s(p)\mathbb{E}[\eta|\delta_z z + \delta_\eta \eta = -\delta_p(d)] \geq 0 - \text{increasing in } z, \ \forall p, d \in \mathbb{R} : d \neq p \tag{31}$$

Suppose that $\delta_z, \delta_\eta \neq 0$ to rule out uninteresting cases. (31) rewrites as:

$$\beta^s(p) \geq \gamma^s(p)\frac{\beta^s(p) - \beta^d(p)}{\gamma^s(p) - \gamma^d(p)} \tag{32}$$

For concreteness, consider two positive supply shocks, i.e. $\beta^s(p), \gamma^s(p) > 0$. Equation (32)

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:

$$sgn(\delta_\eta) \neq sgn(\delta_z) \quad \text{or} \quad \left|\frac{\beta^s(p) - \beta^d(p)}{\beta^s(p)}\right| \leq \left|\frac{\gamma^s(p) - \gamma^d(p)}{\gamma^s(p)}\right| \tag{33}$$

Under $sgn(\delta_\eta) = sgn(\delta_z)$, equation (33) 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-p 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 against cMIV conditions whenever $\mathbb{E}[Y(t)|T = t, Z = z]$ fail to be monotone in the data. In general, the power and size of that test are unclear. 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.

Researchers often impose even stricter versions of this homogeneity assumption. For example, Manski and Pepper (2009) discuss MIV identification under HLR condition: $Y(t) = \beta t + \eta$. Conditions in Proposition 4.2 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.

Proposition 4.2. *Suppose that a): i)* $Y(t) = g(t, \xi) + h(t)\psi(Z, \eta)$, $h(t) \neq 0$ *with* $\xi \perp\!\!\!\perp (T, Z, \eta)$ *and ii) MIV holds, strictly for some* $z, z' \in \mathcal{Z}$ *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)* $\frac{h(t)}{h(d)} > 0$ $\forall t, d \in \mathcal{T}$*; and iii) MIV holds. Then Assumption cMIV-p holds iff* $\mathbb{E}[Y(t)|T = t, Z = z]$ *are all monotone.*

Note that whether or not $h(t) \neq 0$ is observable in the data for case (a) and whether or not $h(t)/h(d) > 0$ is also identified for (b).

Remark 4.2. The monotonicity of observed outcomes has been routinely used in applied work to motivate the use of MIV condition (e.g. De Haan (2017)). We show that, given that MIV holds and under a homogeneity condition, the observed monotonicity is instead equivalent to cMIV-p.

Remark 4.3. 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 $a, b \in \mathbb{R}$.

A test of cMIV-p is thus the test of all $f_t(z) \equiv \mathbb{E}[Y(t)|T = t, Z = z]$ being monotone:

$$\mathcal{H}_0 : f_t(z) - \text{increasing in } z, \ \forall t \in \mathcal{T}$$

To obtain such a test, we may extend the procedure in Chetverikov (2019)[23]. Denote the set of all observations with treatment level t as $\mathcal{I}_t \equiv \{i \in \overline{1,n} : T_i = t\}$ with $n_t \equiv |\mathcal{I}_t|$. 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 $\alpha_t \in (0; 0.5)$. We define the joint test as:

$$\phi_n \equiv \max_{t \in \mathcal{T}} \ \phi^t_{n_t}$$

Denote $\mathcal{P}^C$ to be the set of probability measures, such that for all $P \in \mathcal{P}^C$ and all $t \in \mathcal{T}$ the conditional probability measure given $T = t$ that P generates satisfies the regularity conditions in Theorem 3.1 in Chetverikov (2019). Similarly, let $\mathcal{P}^C_t$ be the set of all the conditional probability measures given $T = t$ that measures from $\mathcal{P}^C$ generate.

Proposition 4.3. *If* $\Pi_{t \in \mathcal{T}}(1 - \alpha_t) \geq 1 - \alpha$, *then:*

$$\inf_{P \in \mathcal{P}^C \cap \mathcal{H}_0} P[\phi_n = 0] \geq 1 - \alpha + o(1) \tag{34}$$

as $n \to \infty$.

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

[23]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 $\mathbb{E}[Y(t)|T = t, Z = z]$ 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.

By the standard optimization argument:

$$\Pi_{t\in\mathcal{T}} \inf_{P\in\mathcal{P}_t^C} P[\phi_{n_t}^t = 0] \leq \inf_{P\in\mathcal{P}^C} \Pi_{t\in\mathcal{T}} P[\phi_{n_t}^t = 0] \tag{35}$$

Theorem 3.1 from Chetverikov (2019) and $\Pi_{t\in\mathcal{T}}(1-\alpha_t) \geq 1-\alpha$ then yield the result. ■

Remark 4.4. 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 $\Pi_{t\in\mathcal{T}}(1-\alpha_t) \geq 1-\alpha$. This may improve the power of the test.

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[24]. 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[25].

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'[26]. Our measure of ability is constructed as a CES aggregator, which is then split into deciles:

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

[24]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.

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

[26]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. $S_i \in [9;11)$ 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.

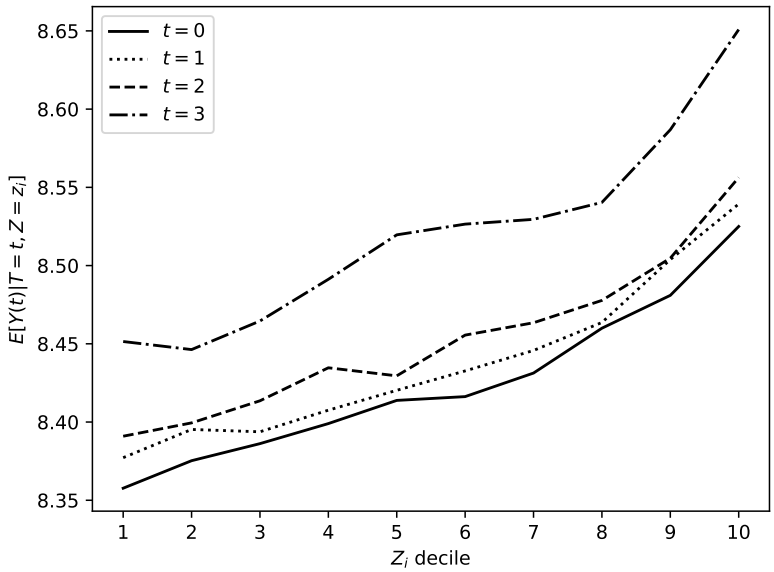


Figure 10: Estimated conditional moments of log-wages given ability and education level.

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 4.2. 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[27]. In line with that, the strongest form of cMIV is not sufficient to provide identification in the absence of further

[27]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.

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) 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).

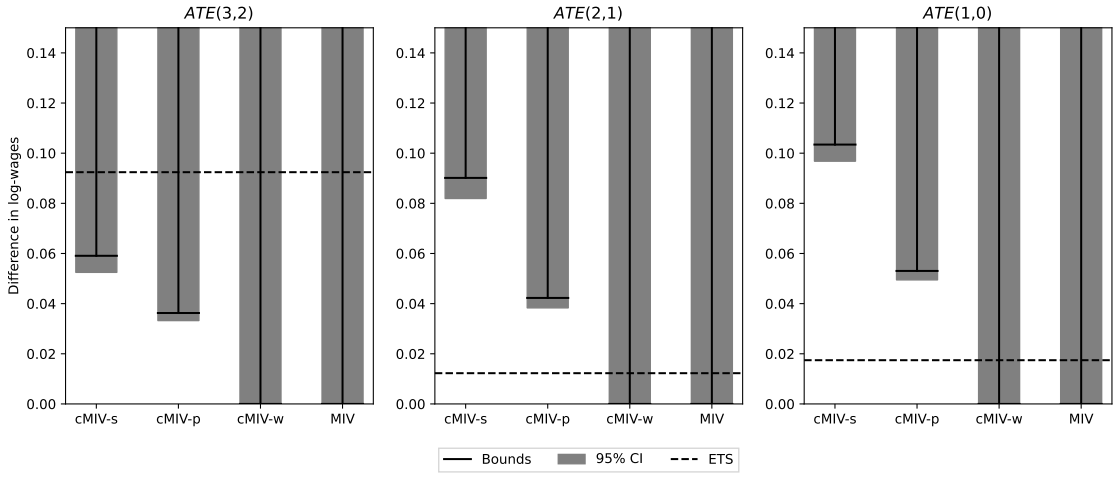


Figure 11: Estimation results for the monotonicity assumptions augmented with MTR. CI constructed according to Proposition 11. The exogenous treatment selection estimates (ETS) are $ATE^{ETS}_{t,d} \equiv \mathbb{E}_n[Y(t)|T=t] - \mathbb{E}_n[Y(d)|T=d]$

We also find significantly positive effects at other education stages, see Figure 11. Further details on data construction and estimation as well as robustness checks are available in the Appendix.

References

Aliprantis, C. and K. Border (2007): *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer.

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

Andrews, D. W. K. (2000): "Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68, 399–405.

Andrews, I., J. Roth, and A. Pakes (2023): "Inference for Linear Conditional Moment Inequalities," *The Review of Economic Studies*, 90, 2763–2791.

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

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

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

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

Bhattacharya, D. (2009): "Inferring optimal peer assignment from experimental data," *Journal of the American Statistical Association*, 104, 486–500.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the distribution of male and female wages accounting for employment composition using bounds," *Econometrica*, 75, 323–363.

Boes, S. (2009): "Bounds on counterfactual distributions under semi-monotonicity constraints," .

Brand, J. E. and Y. Xie (2010): "Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education," *American Sociological Review*, 75, 273–302, pMID: 20454549.

Card, D. (1993): "Using geographic variation in college proximity to estimate the return to schooling," .

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737.

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

Cho, J. and T. M. Russell (2023): "Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments," *Journal of Business & Economic Statistics*, 0, 1–16.

Cygan-Rehm, K., D. Kuehnle, and M. Oberfichtner (2017): "Bounding the causal effect of unemployment on mental health: Nonparametric evidence from four countries," *Health Economics*, 26, 1844–1861.

De Haan, M. (2017): "The Effect of Additional Funds for Low-ability Pupils: A Non-parametric

Bounds Analysis," *The Economic Journal*, 127, 177–198.

Demyanov, V. F. (2009): *Minimax: directional differentiabilityMinimax: Directional Differentiability*, Boston, MA: Springer US, 2075–2079.

Duan, Q., M. Xu, L. Zhang, and S. Zhang (2020): "Hadamard directional differentiability of the optimal value of a linear second-order conic programming problem," *Journal of Industrial and Management Optimization*, 17, 3085–3098.

Fang, Z. and A. Santos (2018): "Inference on Directionally Differentiable Functions," *The Review of Economic Studies*, 86, 377–412.

Gafarov, B. (2024): "Simple subvector inference on sharp identified set in affine models," *arXiv e-prints*, arXiv–1904, conditionally Accepted at Journal of Econometrics, 2024.

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

Gundersen, C., B. Kreider, and J. Pepper (2012): "The impact of the National School Lunch Program on child health: A nonparametric bounds analysis," *Journal of Econometrics*, 166, 79–91, annals Issue on "Identification and Decisions", in Honor of Chuck Manski's 60th Birthday.

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

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

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

Hong, H. and J. Li (2015): "The numerical delta method and bootstrap," Tech. rep., Working paper.

Honoré, B. E. and A. Lleras-Muney (2006): "Bounds in competing risks models and the war on cancer," *Econometrica*, 74, 1675–1698.

Honoré, B. E. and E. Tamer (2006): "Bounds on parameters in panel dynamic discrete choice models," *Econometrica*, 74, 611–629.

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

Imbens, G. W. and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

Kline, B. and E. Tamer (2023): "Recent Developments in Partial Identification," *Annual Review of Economics*, 15, 125–150.

Kline, P. and M. Tartari (2016): "Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach," *American Economic Review*, 106, 972–1014.

Kreider, B., J. V. Pepper, C. Gundersen, and D. 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, 958–975.

Lafférs, L. (2019): "Bounding average treatment effects using linear programming," *Empirical*

economics, 57, 727–767.

Laffέrs, L. (2013): "A note on bounding average treatment effects," *Economics Letters*, 120, 424–428.

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

Manski, C. F. (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

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

——— (2009): "More on monotone instrumental variables," *The Econometrics Journal*, 12, S200–S216.

Masten, M. A. and A. Poirier (2018): "IDENTIFICATION OF TREATMENT EFFECTS UNDER CONDITIONAL PARTIAL INDEPENDENCE," *Econometrica*, 86, 317–351.

Meyer, R. (1979): "Continuity properties of linear programs," Tech. rep., University of Wisconsin-Madison Department of Computer Sciences.

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

Pinar, M. Ç. and B. Chen (1999): "l 1 solution of linear inequalities," *IMA journal of numerical analysis*, 19, 19–37.

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

Rockafellar, R. T. (1970): *Convex Analysis*, Princeton: Princeton University Press.

Semenova, V. (2023): "Adaptive Estimation of Intersection Bounds: a Classification Approach," .

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

Siddique, Z. (2013): "Partially Identified Treatment Effects Under Imperfect Compliance: The Case of Domestic Violence," *Journal of the American Statistical Association*, 108, 504–513.

Syrgkanis, V., E. Tamer, and J. Ziani (2021): "Inference on auctions with weak assumptions on information," *arXiv preprint arXiv:1710.03830*.

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

Van Der Vaart, A. W., J. A. Wellner, A. W. van der Vaart, and J. A. Wellner (1996): *Weak convergence*, Springer.

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

Wainwright, M. J. (2019): *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge university press.

Wright, S. J. (1997): *Primal-dual interior-point methods*, SIAM.

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

6. Appendix

6.1. Proof of Lemma 2.1

Proof. Recall that for any sets A, B, we have $\inf A \cup B = \min\{\inf A, \inf B\}$. Fix any $(\theta, w) \in \mathbb{R}^S \times \mathbb{R}^q$ such that $\Theta_I(\theta) \subseteq \mathcal{X}$ and take $A \equiv L(\Theta_I(\theta); \theta, w)$, $B \equiv L(\mathcal{X} \cap \Theta_I(\theta)'; \theta, w)$. Note that $A \cup B = L(\mathcal{X}; \theta, w)$. Substituting the definitions, it follows that

$$\begin{aligned}\tilde{B}(\theta; w) &= \min\{\inf_{x \in \Theta_I(\theta)} p'x, \inf_{x \in \mathcal{X} \cap \Theta_I(\theta)'} p'x + w'(c - Mx)^+\} \\ &= \min\{B(\theta), \inf_{xx \in \mathcal{X} \cap \Theta_I(\theta)'} p'x + w'(c - Mx)^+\} \leq B(\theta),\end{aligned}$$

which establishes (7).

For the second part, fix any pair (θ_0, w) that satisfies Assumption A0. We write $\theta_0 = (p', c', \text{vec}(M)')'$. Let λ^* be the KKT vector from Assumption A1. The definition of KKT vector (see Section 28 in Rockafellar (1970)) requires that

$$B(\theta_0) = \inf_{x \in \mathbb{R}^d} p'x + \lambda^{*\prime}(c - Mx) \tag{36}$$

Note that, for any $x \in \mathbb{R}^d$,

$$B(\theta_0) \leq p'x + \lambda^{*\prime}(c - Mx) \leq p'x + \lambda^{*\prime}(c - Mx)^+ \leq p'x + w'(c - Mx)^+, \tag{37}$$

where the first inequality follows from (36), the second inequality follows from $\lambda^* \geq 0$ and $(t)^+ \geq t$ for any $t \in \mathbb{R}$, and the third inequality follows by Assumption A1. Taking infinum over $x \in \mathcal{X}$ on both sides of (37) and combining with (7) yields

$$\tilde{B}(\theta_0; w) = B(\theta_0) \tag{38}$$

We now wish to show that $\mathcal{A}(\theta_0) = \tilde{\mathcal{A}}(\theta_0; w)$. From (38), the fact that $L(x; \theta_0, w) = p'x$ for $x \in \Theta_I(\theta_0)$ and $\mathcal{A}(\theta_0) \subseteq \Theta_I(\theta_0)$, it follows that:

$$\mathcal{A}(\theta_0) \subseteq \tilde{\mathcal{A}}(\theta_0; w) \tag{39}$$

To establish the other direction, we proceed by contradiction. Suppose $\exists x^* \in \tilde{\mathcal{A}}(\theta_0; w) \cap \mathcal{A}(\theta_0)'$. Suppose $x^* \in \Theta_I(\theta_0)$. Since $x^* \notin \mathcal{A}(\theta_0)$, it must then be that $p'x^* > B(\theta_0)$, but $\tilde{B}(\theta_0; w) = L(x^*; \theta_0, w) = p'x^*$, which yields a contradiction with (38). So, $x^* \notin \Theta_I(\theta_0)$. Consider

$$\tilde{B}(\theta_0; w) = p'x^* + w'(c - Mx^*)^+ > p'x^* + \lambda^{*\prime}(c - Mx^*)^+ \geq p'x^* + \lambda^{*\prime}(c - Mx^*) \tag{40}$$

where the first inequality follows from Assumption A1 and the fact that $x^* \notin \Theta_I(\theta_0) \implies$

$\exists j : c_j - M_j x^* > 0$. Combining (40) with (36), one gets $\tilde{B}(\theta_0; w) > B(\theta_0)$, which yields a contradiction with (38). So, $\mathcal{A}(\theta_0) \supseteq \tilde{\mathcal{A}}(\theta_0; w)$. Combining with (39) establishes

$$\mathcal{A}(\theta_0) = \tilde{\mathcal{A}}(\theta_0; w).$$

This concludes the proof of the lemma. ■

6.2. Proof of Theorem 2.1

Proof. Fix the true $\theta_0 = (p', c', \text{vec}(M))'$. Recall that for any $g_1, g_2 \in \mathcal{C}(\mathcal{X})$, we can bound

$$|\inf_{\mathcal{X}} g_1 - \inf_{\mathcal{X}} g_2| \leq \sup_{\mathcal{X}} |g_1 - g_2|. \tag{41}$$

Clearly, $L(\cdot; \theta, w) \in \mathcal{C}(\mathcal{X})$ for any $(\theta, w) \in \mathbb{R}^S \times \mathbb{R}_+$. Using this, the bound (41) and the definition of $\tilde{B}(\cdot)$, one gets

$$|\tilde{B}(\hat{\theta}_n; w_n) - \tilde{B}(\theta_0; w_n)| \leq \sup_{x \in \mathcal{X}} |L(x; \hat{\theta}_n, w_n) - L(x; \theta_0, w_n)| \tag{42}$$

Using the definition of $L(\cdot)$, triangle and Cauchy-Schwarz inequalities, for every $x \in \mathcal{X}$

$$|L(x; \hat{\theta}_n, w_n) - L(x; \theta_0, w_n)| \leq ||\hat{p}_n - p|| \cdot ||x|| + w_n \sqrt{q} ||(\hat{c}_n - \hat{M}_n x)^+ - (c - Mx)^+||. \tag{43}$$

It is straightforward to observe that for any $v_1, v_2 \in \mathbb{R}^q$

$$||v_1^+ - v_2^+|| \leq ||v_1 - v_2|| \tag{44}$$

Further, recall that for any $A \in \mathbb{R}^{q \times d}$ and $x \in \mathbb{R}^d$

$$||Ax|| \leq ||x|| \sup_{||y|| \leq 1} ||Ay|| = ||x|| \cdot ||A||_2, \tag{45}$$

where $||A||_2$ is the spectral norm of A. Also recall that if $|| \cdot ||_F$ is the Frobenius norm,

$$||A||_2 \leq ||A||_F = ||\text{vec}(A)||. \tag{46}$$

Combining (44), (45), (46) and using triangle inequality, one gets

$$\begin{aligned} |L(x; \hat{\theta}_n, w_n) - L(x; \theta_0, w_n)| &\leq ||\hat{p}_n - p|| \cdot ||x|| \\ + w_n \sqrt{q} \left(||\hat{c}_n - c|| + ||\text{vec}(\hat{M}_n) - \text{vec}(M)|| \cdot ||x|| \right), \end{aligned}$$

where taking sup on both sides and using (42) yields

$$|\tilde{B}(\hat{\theta}_n; w_n) - \tilde{B}(\theta_0; w_n)| \leq ||\hat{p}_n - p|| \cdot ||x||_\infty \tag{47}$$
$$+w_n\sqrt{q}\left(||\hat{c}_n - c|| + ||\text{vec}(\hat{M}_n) - \text{vec}(M)|| \cdot ||x||_\infty\right) = O_p\left(\frac{w_n}{\sqrt{n}}\right),$$

where $||x||_\infty = \sup_{x\in\mathcal{X}} ||x|| < \infty$ by Assumption A0.ii, and the last equality follows from Assumption A0.iii.

Finally, since θ_0 satisfies A0, there exists some λ^* in $\Lambda(\theta_0)$ with $||\lambda^*||_\infty < \infty$. Let $E_n \equiv \{w_n > ||\lambda^*||_\infty\}$. By Lemma 2.1, $E_n \subseteq \{\tilde{B}(\theta_0; w_n) = B(\theta_0)\}$, so, for any deterministic $\{r_n\}_{n\in\mathbb{N}}$ and any $\varepsilon > 0$,

$$\mathbb{P}\left[r_n|\tilde{B}(\hat{\theta}_n; w_n) - B(\theta_0)| > \varepsilon\right] = \tag{48}$$
$$\mathbb{P}\left[r_n|\tilde{B}(\hat{\theta}_n; w_n) - \tilde{B}(\theta_0; w_n)| > \varepsilon, E_n\right] + \mathbb{P}\left[r_n|\tilde{B}(\hat{\theta}_n; w_n) - B(\theta_0)| > \varepsilon, E_n'\right] \leq$$
$$\mathbb{P}\left[r_n|\tilde{B}(\hat{\theta}_n; w_n) - \tilde{B}(\theta_0; w_n)| > \varepsilon\right] + \mathbb{P}[E_n']$$

By definition, $w_n \to \infty$ w.p.a.1 implies $\mathbb{P}[w_n > ||\lambda^*||_\infty] \to 1$, so $\mathbb{P}[E_n'] \to 0$. Combining this, (47) and (48), one obtains:

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

This concludes the proof of the Theorem.

■

6.2.a. *Set expansions approach.* Proposition 2.1 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[28]. The resulting estimator has the flavor of the approach in Chernozhukov et al. (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.

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

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 $\Theta_I \equiv \{x \in \mathcal{X} | Q(x) = 0\} = \{x \in \mathcal{X} | Mx - c \geq 0\}$.

Lemma 6.1. $||\hat{Q}_n(x) - Q(x)||_\infty \xrightarrow{p} 0$, where $||\cdot||_\infty$ is over Θ_I.

Proof.

$$|\hat{Q}_n(x) - Q(x)| = \left| \sum_j \left([\hat{M}_n x - \hat{c}_n]_j^- \right)^2 - \left([Mx - c]_j^- \right)^2 \right| = \tag{49}$$

$$= \left| \sum_j ([\hat{M}_n x - \hat{c}_n]_j^- - [Mx - c]_j^-)([\hat{M}_n x - \hat{c}_n]_j^- + [Mx - c]_j^-) \right| \leq \tag{50}$$

$$\leq \sum_j |[\hat{M}_n x - \hat{c}_n]_j^- - [Mx - c]_j^-| \cdot |[\hat{M}_n x - \hat{c}_n]_j^- + [Mx - c]_j^-| \leq \tag{51}$$

$$\leq \left(\max_j \, [\hat{M}_n x - \hat{c}_n]_j^- + [Mx - c]_j^- \right) \sum_j \left| \left[(\hat{M}_n - M)x + c - \hat{c}_n \right]_j \right| \tag{52}$$

Where (52) uses the fact that $|(y_0)^- - (y_1)^-| = |\max\{0, -y_0\} - \max\{0, -y_1\}| \leq |y_0 - y_1| \; \forall y_0, y_1 \in \mathbb{R}$. We now show that the last line converges to 0 is supremum over $x \in \mathcal{X}$. 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, $\forall \delta > 0$, we have $(\hat{c}_n', \text{vec}(\hat{M}_n)')' \in B_\delta((c', \text{vec}(M)')')$ w.p. 1 asymptotically.

Since $\mathcal{X}$ 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:[29]:

$$\sup_{x \in \mathcal{X}} \max_j \, [\hat{M}_n x - \hat{c}_n]_j^- + [Mx - c]_j^- \; \leq K + o_p(1) \tag{53}$$

Note that by Cauchy-Schwarz, $\sum_j \left| \left[(\hat{M}_n - M)x + c - \hat{c}_n \right]_j \right| \leq |\mathcal{F}_t| \cdot ||(\hat{M}_n - M)x + c - \hat{c}_n||$. Further using (52), (53) and noting that for nonnegative f, g one has $\sup_A fg \leq \sup_A f \cdot \sup_A g$, we get:

$$||\hat{Q}_n(x) - Q(x)||_\infty \leq (K + o_p) \cdot |\mathcal{F}_t| \cdot \sup_{x \in \mathcal{X}} ||(\hat{M}_n - M)x + c - \hat{c}_n|| \leq \tag{54}$$

$$\leq (\tilde{K} + o_p) \cdot \left(||\hat{M}_n - M|| \cdot ||x||_\infty + ||c - c_n|| \right) = o_p(1) \tag{55}$$

The proof is complete. ■

[29] 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.

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:

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

The plug-in estimator of the identified set, $\{x \in \mathcal{X} | \hat{Q}_n = 0\} = \Theta_I(\hat{\theta}_n)$, may not 'cover' the true asymptotically, as discussed in Chernozhukov et al. (2007) (CHT). To address that, consider the following class of set estimators:

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

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 $\underline{\Theta}_n \equiv \{x | \hat{Q}_n(x) \leq \frac{\kappa_n}{n}\} \subseteq \hat{\Theta}_n$ and $\overline{\Theta}_n \equiv \{x | \hat{Q}_n(x) \leq q\frac{\kappa_n}{n}\} \supseteq \hat{\Theta}_n$.

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

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

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 6.2 hold in our case. The derivation below follows the usual CHT logic. The first equality is established through:

$$P[\sup_{x \in \overline{\Theta}_n} d(\theta, \Theta_I) \leq \varepsilon] = P[\overline{\Theta}_n \subseteq \Theta_I^\varepsilon] = \tag{56}$$
$$P[\overline{\Theta}_n \cap \mathcal{X} \setminus \Theta_I^\varepsilon = \emptyset] \geq P[\sup_{x \in \overline{\Theta}_n} Q(\theta) < \inf_{x \in \mathcal{X} \setminus \Theta_I^\varepsilon} Q(\theta)]$$

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

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

By construction of Θ_I and continuity of $Q(\theta)$, $\exists\delta > 0$: $\inf_{x\in\mathcal{X}\setminus\Theta_I^\varepsilon} Q(\theta) > \delta$. Thus, the RHS of (56) goes to 1. So, $\sup_{x\in\overline{\Theta}_n} d(x,\Theta_I) = o_p(1)$.

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

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

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 6.1. *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 et al. (2007) suggest to select some κ_n that diverges sufficiently slowly with the sample size. We use $\sqrt{\kappa_n} \propto \ln\ln n$ in the simulations in Section 2.6. 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 2.1 without the inequality $x_2 \leq x_1$ and setting $\hat{b}_n = 0$. The minimum is attained at $-1 - \sqrt{\frac{\kappa_n}{n}}$. Our simulation evidence suggests that the set-expansion estimator can be quite conservative in practice, see Section 2.6.

6.3. Proof of Proposition 5

First, notice that the iff result for the following conditions: i) SC holds and ii) $\mathcal{A}(\theta_0)$ and $\Lambda(\theta_0)$ are both singletons follows from Theorem 3.1 in Fang and Santos (2018) combined with Lemma 2.2 and Proposition 2.3. Then, observe that $\mathcal{A}(\theta_0)$ and $\Lambda(\theta_0)$ are both singletons if and only if both LICQ hold and there are no flat faces.

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

6.4.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[30].

Assumption B2 (Slater's condition). $\mathcal{I}nt(\Theta_I) \neq \emptyset$

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:

$$\tilde{B}'_n(\mathbb{Z}^*_n) \equiv \frac{B(\hat{\theta}_n + \epsilon_n \mathbb{Z}^*_n) - B(\hat{\theta}_n)}{\epsilon_n} \tag{57}$$

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

[30] 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.

Proposition 6.2. *If Assumptions B1, B2 hold, and the bootstrapped $\mathbb{Z}_n^*$ satisfies the measurability conditions in Hong and Li (2015):*

$$\sup_{f \in BL_1(\mathbb{R})} |\mathbb{E}[f(\tilde{B}_n'(\mathbb{Z}_n^*))|\{X_i\}_{i=1}^n] - \mathbb{E}[f(B_{\theta_0}'(\mathbb{G}_0))]| = o_p(1) \tag{58}$$

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:

$$\min_{\Theta_I} p'x - ||p|| d_H(\Theta_I, \tilde{\Theta}_I) \leq \min_{\tilde{\Theta}_I} p'x \leq \min_{\Theta_I} p'x \tag{59}$$

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.

6.4.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 map $\tilde{L} : \mathbb{R}^S \to \ell^\infty(\mathcal{X})$, and $\phi : \ell^\infty(\mathcal{X}) \to \mathbb{R}$ is given by:

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

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 6.1. $g(y)(x) \equiv (x+y)^+$ viewed as a map $g : \mathbb{R} \to \ell^\infty(A)$ for $x \in A \equiv [-C; C]$ for some $C > 0$ is **not** Hadamard directionally differentiable for any fixed $y \in [-C/2; C/2]$:

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

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:

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

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[31].

Proposition 6.3. *The penalty function estimator, $\tilde{B}(\theta; w)$ is Hadamard directionally differentiable in θ at θ_0 if either i) $\mathcal{X}$ is a polytope with $Int(\mathcal{X}) \neq \emptyset$, or ii) $\exists x \in \tilde{\mathcal{A}}(\theta_0; w) \cap Int(\mathcal{X})$. The H.d.d. is given by:*

$$\tilde{B}'_{\theta_0}(h; w) = \inf_{x \in \tilde{\mathcal{A}}(\theta_0; w)} \sup_{\lambda \in \tilde{\Lambda}(\theta_0; w)} h'_p x + \sum_{j=1}^{2^q} \lambda_j \sum_{i \in \Pi_j} w_i (h_{c_i} - h'_{M_i} x) \tag{60}$$

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:

$$L(x; \theta, w) = \max_{j \in \overline{1,2^q}} \left\{ p'x + \sum_{i \in \Pi_j} w_i (c_i - M'_i x) \right\}, \tag{61}$$

[31] 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.

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 $g_j(x,\theta) \equiv p'x + \sum_{i\in\Pi_j} w_i(c_i - M_i'x)$ for $j \in \overline{1,2^q}$.

The initial estimator can then be represented as:

$$\tilde{B}(\theta; w) = \min_{x\in\mathcal{X}} \max_{j\in\overline{1,2^q}} g_j(x,\theta) \tag{62}$$

2. Assumptions i) or ii) allow us to impose w.l.g. that the known compact set $\mathcal{X}$ 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 $\mathcal{X}$ to be some compact and non-empty polyhedron that contains the original set. Then, another representation of the considered problem follows:

$$\tilde{B}(\theta; w) = \min_{t,x} t \quad \text{s.t.:} \begin{cases} t \in [\underline{t}; \overline{t}] \\ x \in \mathcal{X} \\ g_j(x,\theta) \leq t, j \in \overline{1,2^q} \end{cases} \tag{63}$$

For some sufficiently wide $[\underline{t}, \overline{t}]$, given θ is close to θ_0 and such that $\tilde{B}(\theta_0; w) \in (\underline{t}, \overline{t})$. 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 (63) 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 $x(\theta_0) \in \tilde{A}(\theta_0; w) \subseteq \mathcal{X}$ 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 $j \in \overline{1,2^q}$, $\exists \delta > 0$ such that $t \geq \max_j g_j(x)$ $\forall t \in (t^* - \frac{\varepsilon}{4}; t^* + \frac{\varepsilon}{4}), \forall x \in \mathbb{B}_\delta(x(\theta_0))$. By either i) or ii) $Int(\mathcal{X}) \neq \emptyset$ and as $x(\theta_0) \in \mathcal{X}$ it follows that $Int(\mathcal{X}) \cap \mathbb{B}_\delta(x(\theta_0))$ is non-empty. It is also open as an intersection of two open sets. Therefore, the open set $\mathcal{O} \equiv (t^* - \frac{\varepsilon}{4}; t^* + \frac{\varepsilon}{4}) \times (\mathbb{B}_\delta(x(\theta_0)) \cap Int(\mathcal{X}))$ 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 (63) 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 $j \in \overline{1,2^q}$. 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 (63) is equal to $\tilde{B}(\theta; w)$ and, moreover, the problems are equivalent, so if $\check{A}(\theta)$ is the arg min of (63),

then $\check{\mathcal{A}}(\theta) = \{\tilde{B}(\theta; w)\} \times \tilde{\mathcal{A}}(\theta; w)$. Using the conclusion of Step 3, direct application of Lemma 6 to (63) yields:

$$\tilde{B}'_{\theta_0}(h; w) = \inf_{x \in \mathcal{A}(\theta_0; w)} \sup_{\lambda \in \tilde{\Lambda}(\theta_0)} \sum_{j=1}^{2^q} \lambda_j \left(h_p' x + \sum_{i \in \Pi_j} w_i (h_{c_i} - h'_{M_i} x) \right), \tag{64}$$

where note that there are no terms corresponding to the objective function and the constraints $t \in [\underline{t}, \overline{t}]$ and $x \in \mathcal{X}$, because there are no corresponding increments. Moreover, differentiating the Lagrangean of (63) and recalling that $t(\theta_0) \in (\underline{t}; \overline{t})$, so the constraints $t \in [\underline{t}, \overline{t}]$ do not bind and the corresponding multipliers are 0, one gets that $\forall \lambda \in \tilde{\Lambda}(\theta_0)$, we have $\sum_{j=1}^{2^q} \lambda_j = 1$, establishing (60).

■

Remark 6.2. By Lemma 2, Assumption A1 ensures ii) in Proposition 11 if $\Theta_I \subseteq Int(\mathcal{X})$.

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 $\mathcal{X}$ 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.

6.5. Proof of Theorem 2.2

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

Lemma 6.3. 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 $\{x_n\} \subseteq A$ such that $f(x_n) \xrightarrow{p} c$, there exists a measurable random sequence $\{x_n^*\} \subseteq B$ 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 $\exists \varepsilon > 0$:

$$\mathbb{P}[||x_n^* - x_n|| > \varepsilon] \not\to 0 \tag{65}$$

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

$$\mathbb{P}[||x_{n_k}^* - x_{n_k}|| > \varepsilon] > \delta \tag{66}$$

Consider the following problem:

$$\min_{x \in A,\ d(x,B) \geq \varepsilon} f(x) \tag{67}$$

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 (66). 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 $k \in \mathbb{N}$:

$$||x_{n_k}^* - x_{n_k}|| > \varepsilon \implies f(x_{n_k}) \geq f(\tilde{x}) = c + K > c \tag{68}$$

So,

$$\mathbb{P}[f(x_{n_k}) - f(x^*) \geq K] \geq \mathbb{P}[||x_{n_k}^* - x_{n_k}|| > \varepsilon] > \delta > 0, \tag{69}$$

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

$$p'x_n + w_n \iota'(\hat{c}_n - \hat{M}_n x_n)^+ - p'x^* = o_p(1) \tag{70}$$

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

$$w_n ||\hat{c}_n - \hat{M}_n x - c + Mx||_\infty = o_p(1) \tag{71}$$

Because the argmin is contained in a compact, $\mathcal{A}(\hat{\theta}_n, w_n) \subseteq \mathcal{X}$, the first term in (70) is bounded in probability: $p'x_n = O_p(1)$, thus, from (70), it also follows that $w_n \iota(\hat{c}_n - \hat{M}_n x_n)^+ = O_p(1)$. By triangle inequality and using with (71), we therefore

conclude:

$$w_n \iota'(c - Mx_n)^+ = O_p(1) \tag{72}$$

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

$$(c - Mx_n)^+ = o_p(1) \tag{73}$$

We shall now consider $\tilde{x}_n$ - a projection of x_n onto $\{x \in \mathbb{R}^d | Mx \geq c\}$. Note that it exists, because distance is a continuous function and the set is a non-empty compact. Note that (73) implies that, for some random $\kappa_n \geq 0$ for all n:

$$c - Mx_n \leq \iota \kappa_n \tag{74}$$

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

$$||x_n - \tilde{x}_n|| = d(x_n, \{x \in \mathbb{R}^d | Mx \geq c\}) \leq \tag{75}$$

$$\leq d_H(\{x \in \mathbb{R}^d | Mx \geq c - \kappa_n\}, \{x \in \mathbb{R}^d | Mx \geq c\}) \leq C\kappa_n, \tag{76}$$

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 (74) as well as:

$$d(x_n, \{x \in \mathbb{R}^d | Mx \geq c\}) \leq \sup_{x \in \{x \in \mathbb{R}^d | Mx \geq c - \kappa_n\}} d(x, \{x \in \mathbb{R}^d | Mx \geq c\})) \tag{77}$$

The final inequality is implied by Lipschitz-continuity of polytopes in Hausdorff distance with respect to RHS expansions (see Li (1993)). Therefore:

$$\tilde{x}_n - x_n \xrightarrow{p} 0 \tag{78}$$

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

$$\mathbb{P}[|p'x_n + w_n \iota'(\hat{c}_n - \hat{M}_n x_n) - p'x^*| > \varepsilon] \geq \tag{79}$$

$$\mathbb{P}[p'x_n > p'x^* + \varepsilon - w_n \iota'(\hat{c}_n - \hat{M}_n x_n)] \geq \tag{80}$$

$$\mathbb{P}[p'x_n > p'x^* + \varepsilon] \tag{81}$$

As the LHS goes to 0 by (70), we have:

$$\mathbb{P}[p'x_n > p'x^* + \varepsilon] \to 0 \tag{82}$$

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,

$$\mathbb{P}[p'x_n < p'x^* - \varepsilon] \leq \mathbb{P}[p'x_n < p'\tilde{x}_n - \varepsilon] \to 0, \tag{83}$$

where the RHS converges to 0 by (78) 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 6.3 applies, where one sets $f(x) = L(x; \theta_0, w)$, $B = \mathcal{A}(\theta_0)$ with $f(x^*) = p'x^*$ for any $x^* \in \mathcal{A}(\theta_0)$. Thus, $\exists x_n^* \in \mathcal{A}(\theta_0)$ such that $||x_n - x_n^*|| \xrightarrow{p} 0$. Therefore, $\exists \delta_n \to 0^+$ such that:

$$\mathbb{P}[||x_n - x_n^*|| < \delta_n] \to 1 \tag{84}$$

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

$$x_n \in \underset{x \in \mathcal{A}(\hat{\theta}_n, w_n)}{\arg\max}\, d(x, \mathcal{A}(\theta_0)) \tag{85}$$

For such x_n, we get:

$$||x_n - x_n^*|| < \delta_n \implies d(x, \mathcal{A}(\theta_0)) < \delta_n \ \forall x \in \mathcal{A}(\hat{\theta}_n, w_n) \tag{86}$$

So:

$$\mathbb{P}[\mathcal{A}(\hat{\theta}_n, w_n) \subseteq \mathcal{A}(\theta_0)^{\delta_n}] \geq \mathbb{P}[||x_n - x_n^*|| < \delta_n] \to 1 \tag{87}$$

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

2. By (87) and using the representation found in Proposition 6.3 we have that:

$$\inf_{x \in \mathcal{X}} L(x; \hat{\theta}_n, w_n) = \inf_{x \in \mathcal{A}(\theta_0)^{\delta_n}} L(x; \hat{\theta}_n, w_n) + o_p(1) \tag{88}$$

$$= \min_{x \in \mathcal{A}(\theta_0)^{\delta_n}} p'x + w_n \max_{j \in \overline{1,2^q}} \left\{ \sum_{i \in \Pi_j} (\hat{c}_{ni} - \hat{M}'_{ni} x) \right\} + o_p(1), \tag{89}$$

where $o_p(1)$ encompasses realizations at which $\mathcal{A}(\hat{\theta}_n, w_n) \notin \mathcal{A}(\theta_0)^{\delta_n}$ or where $\hat{\theta}_n$ is not in a fixed open vicinity of θ_0 that was argued to exist in Proposition 6.3. Suppose that at θ_0 the constraints that do not bind at any $x \in \mathcal{A}(\theta_0)$ are given by $I \subseteq \{1, 2, \ldots, q\}$.

By continuity, it follows that $\exists\ \delta > 0$ and $\varepsilon > 0$ such that:

$$c_i - M_i x < -\varepsilon, \forall i \in I \tag{90}$$

for any $x \in \mathcal{A}(\theta_0)^{\delta}$. From (87) it then also follows that:

$$\inf_{x \in \mathcal{X}} L(x; \hat{\theta}_n, w_n) = \min_{x \in \mathcal{A}(\theta_0)^{\delta_n}} p'x + w_n \max_{\Pi \in 2^{\overline{1,q}} \setminus I} \left\{ \sum_{i \in \Pi_j} (\hat{c}_{ni} - \hat{M}'_{ni} x) \right\} + o_p(1) \tag{91}$$

3. Consider the problem in the linear programming representation found in Proposition 6.3, which it admits w. p. 1 as.:

$$\inf_{x \in \mathcal{X}} L(x; \hat{\theta}_n; w_n) = \min_{t,x} t \quad \text{s.t.:} \quad \begin{cases} t \in [\underline{t}; \overline{t}] \\ x \in \mathcal{X} \\ p'x + \sum_{i \in \Pi_j} w_n (\hat{c}_{ni} - \hat{M}'_{ni} x) \leq t, j \in \overline{1, 2^q} \end{cases} \tag{92}$$

The Lagrangian reads as:

$$\mathcal{L} = t + \sum_{\Pi \in 2^{\overline{1,q}}} \lambda_{\Pi} \left(p'x - t + w_n \sum_{j \in \Pi} \hat{c}_{nj} - \hat{M}'_{nj} x \right), \tag{93}$$

Where the constraints $x \in \mathcal{X}$ and $t \in [\underline{t}; \overline{t}]$ are omitted, as they are not binding with probability 1 as. This holds, as $\mathcal{A}(\theta_0) \subseteq Int(\mathcal{X})$ and $B(\theta_0) \in Int([\underline{t}; \overline{t}])$ by assumption. Because $\mathcal{A}(\theta_0)$ is compact, there further exists[32] a $\overline{\delta} > 0$: $\mathcal{A}(\theta_0)^{\overline{\delta}} \subseteq Int(\mathcal{X})$ and as $\tilde{\mathcal{A}}(\hat{\theta}_n; w_n) \subseteq \mathcal{A}(\theta_0)^{\delta_n}$ w.p. 1 as. for some $\delta_n \to 0^+$, it follows that w.p.a.1 $\mathcal{A}(\hat{\theta}_n; w_n) \subseteq Int(\mathcal{X})$. Similar argument establishes that $t_n^* \in Int([\underline{t}; \overline{t}])$ w.p.a.1. In what follows, we will simply call such optimal pairs *interior*.

Differentiating with respect to t, one notes that:

$$\sum_{\Pi} \lambda_{\Pi} = 1 \tag{94}$$

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

$$t = p'x + w_n \max_{\Pi} \sum_{j \in \Pi} (\hat{c}_{nj} - \hat{M}'_{nj} x) \tag{95}$$

[32] 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. $\exists (b_k, \varepsilon_{b_k}/2)_{k=1}^K$ 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 $\exists y \in A$: $||x - y|| < \delta$. 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 $x \in B_{\varepsilon_{b_k}}(b_k) \subseteq B$.

To see that, note that by contradiction, if:

$$t > p'x + w_n \max_{\Pi} \sum_{j \in \Pi} (\hat{c}_{nj} - \hat{M}'_{nj} x) \tag{96}$$

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:

$$t \geq p'x \tag{97}$$

Furthermore, the problem has a solution w.p.a.1, 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.a.1 any solution is *interior*, any such matrix w.p.a.1 does not feature constraints $x \in \mathcal{X}, t \in [\underline{t}, \overline{t}]$. The only constraints that can be satisfied at such vertex-solution with an equality are of the following type:

$$p'x - t = w_n \sum_{j \in \Pi_k} \hat{c}_{nj} - \hat{M}'_{nj} x, \;\; k \in \tilde{J} \tag{98}$$

for some $\tilde{J} \subseteq 2^{\overline{1,q}} : |\tilde{J}| \geq d + 1$, where the latter inequality holds by definition of a vertex of a linear program[33]. One can write the complete set of the binding constraints (98) as:

$$\hat{R}_{\tilde{J}n} \begin{pmatrix} t \\ x \end{pmatrix} = \hat{r}_n^{\tilde{J}}, \tag{99}$$

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 (98) with $|\tilde{J}| \geq d+1$ and a full-column-rank $\hat{R}_{\tilde{J}n}$ at a given $\hat{\theta}_n$ by $\mathcal{V}^*(\hat{\theta}_n)$. From the previous arguments it follows that $\mathcal{V}^*(\cdot)$ is non-empty w.p.a.1 and finite, because any finite-dimensional polytope has finitely many vertices and therefore the corresponding LP has finitely many optimal vertices. We will write $\mathcal{V}_x^*(\hat{\theta}_n)$ for the projection of that set on the x-coordinates. For any vertex-solution $(t^*, x^*) \in \mathcal{V}^*(\hat{\theta}_n)$, suppose constraints $V^* \subseteq \{1, \dots q\}$ are violated at it, meaning that:

$$V^*(\hat{\theta}_n, x^*) \equiv \{j \in \overline{1,q} | \hat{c}_{nj} - \hat{M}'_{nj} x^* > 0\} \tag{100}$$

[33] 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 $(t \; x')'$.

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

$$t_n^* = p'x_n^* + w_n \max_{\Pi} \sum_{j \in \Pi} (\hat{c}_{nj} - \hat{M}'_{nj} x_n^*) = p'x_n^* + w_n \sum_{j \in V_n^*} (\hat{c}_{nj} - \hat{M}'_{nj} x_n^*) \tag{101}$$

Consider (98) 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.:

$$t_n^* = p'x_n^* + w_n \sum_{j \in \Pi_i} (\hat{c}_{nj} - \hat{M}'_{nj} x_n^*) \ \forall i \in \overline{1,k} \tag{102}$$

It must be that $V_n^* \subseteq \Pi_i \ \forall i \in \tilde{J}_n$, because $j \notin V_n^* \implies (\hat{c}_{nj} - \hat{M}'_{nj} x_n^*) \leq 0$, and so we have:

$$\sum_{j \in V_n^*} (\hat{c}_{nj} - \hat{M}'_{nj} x_n^*) = \sum_{j \in \Pi_i} (\hat{c}_{nj} - \hat{M}'_{nj} x_n^*) \leq \sum_{j \in \Pi_i \cap V_n^*} (\hat{c}_{nj} - \hat{M}'_{nj} x_n^*), \tag{103}$$

where the first equality follows from (102) and (101). We now proceed by contradiction. Suppose that $\exists j : j \in V_n^* \cap \Pi_i'$ (where the complement is taken with respect to $\overline{1,q}$), then:

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

which yields a contradiction with (103), so there can be no such j. In light of (103) it then also follows that $\forall i \in \tilde{J}_n$ and $\forall j \in \Pi_i \cap V_n^{*\prime}$ it must be that:

$$\hat{c}_{nj} - \hat{M}'_{nj} x_n^* = 0 \ \forall j \in \Pi_k \setminus V_n^* \tag{104}$$

Therefore, the complete system described by equation (102), is equivalent to:

$$\begin{cases} \hat{c}_{nj} - \hat{M}'_{nj} x_n^* = 0 \ \forall i \in \tilde{J}_n : \Pi_i \neq V_n^*, \ \forall \ j \in \Pi_i \setminus V_n^* \\ t_n^* = p'x_n^* + w_n \sum_{j \in V_n^*} \hat{c}_{nj} - \hat{M}'_{nj} x_n^* \end{cases} \tag{105}$$

From the representation (99), we know that the matrix corresponding to system (105) 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 $\Pi^*(\hat{\theta}_n, x_n^*) \equiv \{j \in \overline{1,q} | \hat{c}_{nj} - \hat{M}'_{nj} x_n^* = 0\}$, which we shall occasionally write as Π_n^* for brevity. We thus have:

$$|\Pi_n^*| \geq d, \quad \text{rk}(\hat{M}_{\Pi_n^*}) = d \tag{106}$$

5. Consider two collections of sets:

$$\mathcal{E} \equiv \{A \subseteq 2^{[q]} : M_A x \neq c_A \ \forall x \in \mathcal{A}(\theta_0)\} \tag{107}$$

$$\mathcal{F} \equiv \{A \subseteq 2^{[q]} : p \notin \mathcal{R}(M'_A)\} \tag{108}$$

We shall now consider two events E_n and F_n:

$$E_n \equiv \{\Pi_n^* \in \mathcal{E}\}, \quad F_n \equiv \{\Pi_n^* \in \mathcal{F}\} \tag{109}$$

$$\tag{110}$$

We wish to show that $\mathbb{P}[E_n] \to 0$ and $\mathbb{P}[F_n] \to 0$ and therefore $\mathbb{P}[E'_n \cap F'_n] \to 1$.

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

$$\inf_{x \in \mathcal{A}(\theta_0)} ||M_A x - c_A|| > \varepsilon(A) \tag{111}$$

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

$$\min_{A \in \mathcal{E}} \inf_{x \in \mathcal{A}(\theta_0)} ||M_A x - c_A|| > \varepsilon \tag{112}$$

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

$$\min_{A \in \mathcal{E}} \inf_{x \in \mathcal{A}^\kappa(\theta_0)} ||M_A x - c_A|| > \frac{\varepsilon}{2} \tag{113}$$

We now consider:

$$\mathbb{P}[E_n] \leq \mathbb{P}\left[||\hat{M}_{\Pi_n^*} x_n^* - \hat{c}_{\Pi_n^*}|| = 0, \inf_{x \in \mathcal{A}^\kappa(\theta_0)} ||M_{\Pi_n^*} x - c_{\Pi_n^*}|| > \frac{\varepsilon}{2}\right] \tag{114}$$

Observe that for any non-empty $A \subseteq [q]$, by Cauchy-Schwartz and triangle inequali-

ties:

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

We can thus further rewrite:

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

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 $\mathbb{P}[A \cap B'] + \mathbb{P}[A \cap B] = \mathbb{P}[A]$:

$$\begin{aligned} \mathbb{P}\left[\left|\left|M_{\Pi_n^*}x_n^* - c_{\Pi_n^*}\right|\right| \leq \eta_n, \inf_{x \in \mathcal{A}^\kappa(\theta_0)} ||M_{\Pi_n^*}x - c_{\Pi_n^*}|| > \frac{\varepsilon}{2}\right] = \\ \mathbb{P}\left[\left|\left|M_{\Pi_n^*}x_n^* - c_{\Pi_n^*}\right|\right| \leq \eta_n, \inf_{x \in \mathcal{A}^\kappa(\theta_0)} ||M_{\Pi_n^*}x - c_{\Pi_n^*}|| > \frac{\varepsilon}{2}, x_n^* \in \mathcal{A}^\kappa(\theta_0)\right] + \\ + \mathbb{P}\left[\left|\left|M_{\Pi_n^*}x_n^* - c_{\Pi_n^*}\right|\right| \leq \eta_n, \inf_{x \in \mathcal{A}^\kappa(\theta_0)} ||M_{\Pi_n^*}x - c_{\Pi_n^*}|| > \frac{\varepsilon}{2}, x_n^* \notin \mathcal{A}^\kappa(\theta_0)\right], \end{aligned}$$

where the second term is $o(1)$ by Step 1 of the proof. Finally, we note that $x_n^* \in \mathcal{A}^\kappa(\theta_0) \implies \left|\left|M_{\Pi_n^*}x_n^* - c_{\Pi_n^*}\right|\right| > \varepsilon/2$, 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:

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

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 (92) are equivalent to yet another LP:

$$B(\hat{\theta}_n) + o_p(1/\sqrt{n}) = \min_{x,a} p'x + w_n \iota' a \quad \text{s.t.} : \begin{cases} a \geq 0 \\ a \geq \hat{c}_n - \hat{M}_n x \end{cases} \tag{115}$$

Note that we drop the constraints corresponding to $x \in \mathcal{X}$ in (115), 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:

$$p = \hat{M}_n'\mu \tag{116}$$
$$w_n = \omega + \mu \tag{117}$$
$$\omega' a = 0 \tag{118}$$
$$\mu'(\hat{c}_n - \hat{M}_n x - a) = 0 \tag{119}$$
$$a \geq \hat{c}_n - \hat{M}_n x \tag{120}$$
$$a \geq 0, \omega \geq 0, \mu \geq 0 \tag{121}$$

Analyzing the above system, one observes that if at $x_n^* \in \mathcal{V}_x^*(\theta_n)$ a constraint is violated, $j \in V_n^*$, 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 $j \in \Pi_n^*$, then $\mu_j \in [0; w_n]$. Therefore, (116) rewrites as:

$$p = w_n \sum_{j \in V_n^*} \hat{M}_{nj}' + \sum_{j \in \Pi_n^*} \hat{M}_{nj}'\mu_j \tag{122}$$

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

$$p = w_n \sum_{j \in V_n^*} M_j' + \sum_{j \in \Pi_n^*} M_j'\mu_j + O_p(\frac{w_n}{\sqrt{n}}) \tag{123}$$

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

$$p - O_p(\frac{w_n}{\sqrt{n}}) = w_n(I - P_{\Pi_n^*}) \sum_{j \in V^*} M_j' + \underbrace{w_n P_{\Pi_n^*} \sum_{j \in V^*} M_j' + \sum_{j \in \Pi_n^*} M_j'\mu_j}_{T_n \in \mathcal{R}(M_{\Pi_n^*}')} \tag{124}$$

Notice that, if $\sum_{j \in V^*} M_j' \notin \mathcal{R}(M_{\Pi^*}')$, then the RHS of (124) has unbounded norm:

$$\begin{aligned} &\left|\left| w_n(I - P_{\Pi_n^*}) \sum_{j \in V_n^*} M_j' + T_n \right|\right|^2 = \\ &= w_n^2 ||(I - P_{\Pi_n^*}) \sum_{j \in V_n^*} M_j'||^2 + ||T_n||^2 \end{aligned} \tag{125}$$

Since the square norm of the LHS of (124) is bounded from above by $||p||^2 +$

$O_p(\frac{w_n^2}{n}) = ||p||^2 + o_p(1)$, (125) will contradict the equality in (124) w.p.a.1. Suppose, alternatively, that $\exists v : \sum_{j \in V_n^*} M_j' = M_{\Pi_n^*}' v$. Equation (124) rewrites:

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

which implies, for example, that:

$$(I - P_{\Pi_n^*})p + P_{\Pi_n^*}p - M_{\Pi_n^*}'(\mu_{\Pi_n^*} + w_n v) = O_p(\frac{w_n}{\sqrt{n}}) \tag{126}$$

The norm of the LHS of (126) 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.a.1. To complete the proof, one applies the same probabilistic arguments as used in step 5.a above, which we omit here. Thus, $\mathbb{P}[F_n] \to 0$.

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

$$\mathcal{G} \equiv \{A \subseteq [q] : \exists x \in \mathcal{A}(\theta_0) \text{ s.t. } M_A x = c_A,\ p \in \mathcal{R}(M_A')\}$$

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

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

$$B(\theta_0) = p'x = v'M_A x = v'c_A \tag{127}$$

The conclusion then follows from the following chain of equalities:

$$\begin{aligned} G_n \implies p'x_n^* - B(\theta_0) &= v'M_{\Pi_n^*}x_n^* - v'c_{\Pi_n^*} = && (128)\\ &= v'\hat{M}_{\Pi_n^*}x_n^* - v'c_{\Pi_n^*} + v'(M_{\Pi_n^*} - \hat{M}_{\Pi_n^*})x_n^* = && (129)\\ &= v'(\hat{c}_{\Pi_n^*} - c_{\Pi_n^*}) + v'(M_{\Pi_n^*} - \hat{M}_{\Pi_n^*})x_n^* && (130) \end{aligned}$$

Finally, from (130), 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:

$$G_n \implies |p'x_n^* - B(\theta_0)| \leq \varpi_n \equiv \tag{131}$$
$$\max_{A\in\mathcal{G}} \left\{ \left(||\hat{c}_A - c_A|| + ||x||_\infty ||M_A - \hat{M}_A|| \right) \cdot \min_{v\in\mathbb{R}^{|A|}:\ M_A' v = p} ||v|| \right\}$$

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 $\varepsilon > 0$:

$$\mathbb{P}[r_n|p'x_n^* - B(\theta_0)| > \varepsilon] = \mathbb{P}[r_n|p'x_n^* - B(\theta_0)| > \varepsilon, G_n] + o(1) \leq \tag{132}$$
$$\mathbb{P}[r_n\varpi_n > \varepsilon, G_n] + o(1) \leq \mathbb{P}[r_n\varpi_n > \varepsilon] + o(1) \tag{133}$$

and $r_n\varpi_n = O_p(\frac{r_n}{\sqrt{n}})$ for any $r_n \to \infty$, where we used the fact that $\mathbb{P}[G_n' \cap O_n] \leq \mathbb{P}[G_n'] = o(1)$ 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:

$$\sup_{x\in\mathcal{V}_x^*(\hat{\theta}_n)} |p'x - B(\theta_0)| = O_p(1/\sqrt{n}) \tag{134}$$

But because any $x \in \mathcal{A}(\hat{\theta}_n; w_n)$ can be represented as a convex combination of vertices, $\{x_j\}_{j=1}^K \subseteq \mathcal{V}_x^*(\hat{\theta}_n)$, 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 $x \in \mathcal{A}(\hat{\theta}_n; w_n)$:

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

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

■

6.6. Proof of Theorem 2.3

Proof. For $x \in \mathbb{R}^{qd}$, define the inverse-vectorization operator as

$$\text{vec}_{q\times d}^{-1}(x) \equiv (\text{vec}(I_d)' \otimes I_q)\,(I_d \otimes x).$$

Further, define selector matrices C_c and C_M that select the c and M components of θ respectively:

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

Moreover, for a subset of rows $A \subseteq \{1, 2, \ldots, q\}$, define the row-selector $C(A)$ as

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

We first work on $\mathcal{D}_n^{(1)}$. From Step 5.b in the proof of Theorem 2.2, it follows that solving the penalized problem is w.p.a.1 equivalent to solving a relaxed LP, i.e., w.p.a.1,

$$\tilde{\mathcal{A}}(\hat{\theta}_1; w_{n_1}) = \min_{x \in \mathbb{R}^d, a \in \mathbb{R}^q} p'x + w_{n_1} \iota' a, \quad \text{s.t.: } a \geq \hat{c}^{(1)} - \hat{M}^{(1)} x, \ a \geq 0. \tag{135}$$

Denote the set of vertex-solutions of (135) by $\hat{\mathcal{V}}_x$ and define

$$\hat{x} \in \arg\min_{x \in \hat{\mathcal{V}}_x} p'x, \quad \hat{A} = J(\hat{x}; \hat{\theta}^{(1)}).$$

From Step 6 of the proof of Theorem 2.2 it follows that $\hat{A} \in \mathbb{A}$ w.p.a.1.

For a nonempty $A \in 2^{[q]}$ and any $\tilde{M} \in \mathbb{R}^{q \times d}$, define

$$\mathbb{S}(A, \tilde{M}) \equiv \arg \min_{v \in \mathbb{R}^{|A|} : ||v|| \leq \overline{v}} ||p - \tilde{M}_A' v||^2$$

The optimization problem above is continuous in $\tilde{M}$, the constraint correspondence is constant and compact. Hence, for any nonempty $A \in 2^{[q]}$, by Berge's Maximum Theorem, $\mathbb{S}(A, \cdot)$ is compact, nonempty, and upper-hemicontinuous (see Theorem 17.31 in Aliprantis and Border (2007)).

Because $\hat{M}^{(1)} \xrightarrow{p} \hat{M}^{(1)}$, for any nonempty $A \subseteq 2^{[q]}$ it follows by the usual M-estimation argument[34], there exists a deterministic $s_n(A) \downarrow 0$, such that

$$\mathbb{S}(A, \hat{M}^{(1)}) \subseteq \mathbb{S}(A, M)^{s_n(A)}, \quad \text{w.p.a.1}, \tag{136}$$

so that also

$$\mathbb{S}(\hat{A}, \hat{M}^{(1)}) \subseteq \mathbb{S}(\hat{A}, M)^{s_n(\hat{A})} \text{ w.p.a.1} \tag{137}$$

Observing that the objective function is convex, $\mathbb{S}(A, \tilde{M})$ is also convex-valued for any nonempty $A \in 2^{[q]}$ and $\tilde{M} \in \mathbb{R}^{q \times d}$.

Define some measurable $\check{v} \in \mathbb{S}(\hat{A}, \hat{M}^{(1)})$ and denote its projection onto $\mathbb{S}(\hat{A}, M)$ by $\tilde{v}_n$. Both $\check{v}$ is a random sequence, but we suppress the dependence on n for simplicity. $\tilde{v}_n$ is well-defined by the Hilbert Projection Theorem. $\check{v}$ is well-defined, and $\tilde{v}_n$ is measurable by

[34] Theorem 18.19 in Aliprantis and Border (2007) establishes measurability of $\mathbb{S}(\hat{A}, \hat{M}^{(1)})$ and $\mathbb{S}(\hat{A}, M)$)

Theorem 18.19 in Aliprantis and Border (2007). From (137) it follows that

$$||\check{v} - \tilde{v}_n|| = o_p(1) \tag{138}$$

By definition, $A \in \mathbb{A}$ implies that $\exists v^* \in \mathcal{S}_A$ such that $||v^*|| \leq \max_{\tilde{A}\in\mathbb{A}} \min_{v\in\mathcal{S}_{\tilde{A}}} ||v|| \leq \overline{v}$, where the last inequality is by Assumption B3. This implies that, for any $A \in \mathbb{A}$,

$$\inf_{v\in\mathbb{R}^{|A^*|}:||v||\leq\overline{v}} ||p - M'_{A^*}v||^2 = \min_{v\in\mathbb{R}^{|A^*|}:||v||\leq\overline{v}} ||p - M'_{A^*}v||^2 = 0. \tag{139}$$

By (139), we have $\mathbb{S}(A, M) \subseteq \mathcal{S}_A$ if $A \in \mathbb{A}$. Therefore, $\tilde{v}_n \in \mathcal{S}_{\hat{A}}$ w.p.a.1.

The following Lemma justifies our construction:

Lemma 6.4. Suppose $\hat{A} \in \mathbb{A}$. Then,

$$\tilde{v}'_n c_{\hat{A}} = B(\theta_0), \tag{140}$$
$$\tilde{v}'_n M_{\hat{A}} \hat{x} = p'\hat{x}. \tag{141}$$

Proof. If $\hat{A} \in \mathbb{A}$, condition (10) holds for some $x \in \mathcal{A}(\theta_0)$ such that $M_{\hat{A}}x = c_{\hat{A}}$. Since such x is a minimizer, it follows that $p'x = B(\theta_0)$. As $\tilde{v}_n \in \mathcal{S}_{\hat{A}}$, we have $p = M'_{\hat{A}}\tilde{v}_n$. Taking transpose and multiplying by $\hat{x}$ yields (141). To show (140), write:

$$p'x = \tilde{v}'_n M_{\hat{A}} x = \tilde{v}'_n c_{\hat{A}}$$

This concludes the proof of the Lemma. ■

To avoid dealing with changing dimension, we let $\hat{v} \in \mathbb{R}^q$ be such that $\hat{v}_{\hat{A}} = \check{v}$ and $\hat{v}_j = 0$ if $j \notin \hat{A}$. Similarly, define $\dot{v}_n$: $(\dot{v}_n)_{\hat{A}} = \tilde{v}_n$ and $(\dot{v}_n)_j = 0$ if $j \notin \hat{A}$. Note that $||\hat{v} - \dot{v}_n|| = ||\check{v} - \tilde{v}_n||$.

Equipped with $\hat{v}$, $\hat{A}$ and $\hat{x}$, we can now move onto the second fold. For $(A, v, x) \in 2^{\overline{1,q}} \setminus \{\emptyset\} \times \mathbb{R}^q \times \mathcal{X}$, define

$$H_n(A, x, v) \equiv \frac{\sqrt{n_2}}{\sigma_n(A, x, v)} v'_A \left(\hat{c}^{(2)}_A - c_A - (\hat{M}^{(2)}_A - M_A)x\right), \quad H_n \equiv H_n(\hat{A}, \hat{x}, \hat{v}).$$

Let $Z^{(2)}_n \equiv \sqrt{n_2}(\hat{\theta}^{(2)} - \theta_0)$. One can rewrite

$$\sqrt{n_2} v'_A \left(\hat{c}^{(2)}_A - c_A - (\hat{M}^{(2)}_A - M_A)x\right) = v'_A C(A) \left(C_c Z^{(2)}_n - \text{vec}^{-1}_{q\times d}(C_M Z^{(2)}_n)x\right). \tag{142}$$

Applying the definition of $\text{vec}^{-1}_{q\times d}$ and using bilinearity of Kronecker product, one notes that (142) is linear in $Z^{(2)}_n$ and therefore, under Assumption B1, for any $(A, v, x) \in 2^{\overline{1,q}} \setminus \{\emptyset\} \times \mathbb{R}^q \times \mathcal{X}$,

we have

$$\sqrt{n_2}v_A'\left(\hat{c}_A^{(2)} - c_A - (\hat{M}_A^{(2)} - M_A)x\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2(A, x, v, \Sigma)),$$

where $\sigma^2(\cdot)$ is given in Lemma (6.6).

By assumption B4 we then have, for a fixed optimal triplet A, x, v and CMT,

$$\frac{\sqrt{n_2}}{\sigma(A, x, v, \Sigma)}v_A'\left(\hat{c}_A^{(2)} - c_A - (\hat{M}_A^{(2)} - M_A)x\right) \xrightarrow{d} \mathcal{N}(0, 1) \tag{143}$$

We begin by taking the infeasible $\hat{\sigma}_n(A, v, x) = \sigma(A, v, x, \Sigma)$. Consider the set:

$$\aleph(\underline{v}, \underline{\sigma}) \equiv \{(A, v, x) \in 2^{\overline{1,q}} \setminus \{\emptyset\} \times \mathbb{R}^q \times \mathcal{X} : \underline{v} \leq ||v_n|| \leq \overline{v}, \sigma(A, v, x, \Sigma) \geq \underline{\sigma}\} \tag{144}$$

We now fix an aribtrary deterministic sequence $(A_n, v_n, x_n) \in \aleph(\underline{v}, \underline{\sigma})$ for all $n \in \mathbb{N}$ for some small $\underline{v} > 0$ and $\underline{\sigma} > 0$ that we pick below. Consider the limit (integration is with respect to $\mathcal{D}_n^2$ only):

$$\lim_{n\to\infty} \mathbb{P}[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}]$$

The space $2^{\overline{1,q}}\setminus\{\emptyset\}$, to which A_n belongs, is endowed with a discrete metric, and we consider the space $\aleph(\underline{v}, \underline{\sigma})$ as endowed with the maximum product metric ρ_∞. It is straightforward to notice that $\sigma(\cdot)$ is continuous in its first three arguments with respect to ρ_∞ even on the unrestricted space $2^{\overline{1,q}} \setminus \{\emptyset\} \times \mathbb{R}^q \times \mathcal{X}$, and thus $\aleph(\underline{v}, \underline{\sigma})$ 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:

$$H_n(A_n, v_n, x_n) = g(\sqrt{n_2}(\hat{\theta}^{(2)} - \theta_0), A_n, v_n, x_n) \tag{145}$$

for a continuous function g and:

$$\begin{pmatrix} \sqrt{(n_2)_k}(\hat{\theta}^{(2)} - \theta_0) \\ A_{n_k} \\ v_{n_k} \\ x_{n_k} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{N}(0, \Sigma) \\ A \\ v \\ x \end{pmatrix} \tag{146}$$

we conclude that, by continuous mapping theorem, as $k \to \infty$:

$$g(\sqrt{(n_2)_k}(\hat{\theta}_{n_k}^{(2)} - \theta_0), A_{n_k}, v_{n_k}, x_{n_k}) = H_{n_k}(A_{n_k}, v_{n_k}, x_{n_k}) \xrightarrow{d} g(Z, A, v, x), \tag{147}$$

where $Z \sim \mathcal{N}(0, \Sigma)$. By (143), this implies:

$$\lim_{k\to\infty} \mathbb{P}[H_{n_k}(A_{n_k}, v_{n_k}, x_{n_k}) \leq z_{1-\alpha}] = 1 - \alpha \tag{148}$$

We claim that this further implies that:

$$\lim_{n\to\infty} \mathbb{P}[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}] = 1 - \alpha \tag{149}$$

Suppose, by contradiction, $\lim_{n\to\infty} \mathbb{P}[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}] \neq 1 - \alpha$. It means that $\exists \varepsilon > 0$ such that $\forall N \in \mathbb{N}\ \exists n \geq N$ such that:

$$|\mathbb{P}[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}] - (1 - \alpha)| > \varepsilon \tag{150}$$

Thus, we can construct a subsequence n_k such that:

$$|\mathbb{P}[H_{n_k}(A_{n_k}, v_{n_k}, x_{n_k}) \leq z_{1-\alpha}] - (1 - \alpha)| > \varepsilon \tag{151}$$

for all $k \in \mathbb{N}$. 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, (148), yields that:

$$\mathbb{P}[H_{n_{k_j}}(A_{n_{k_j}}, v_{n_{k_j}}, x_{n_{k_j}}) \to (1 - \alpha), \tag{152}$$

which yields a contradiction. Thus, for any (A_n, v_n, x_n) satisfying $x_n \in \mathcal{X}$, $\underline{v} < ||v_n|| \leq \overline{v}$ and $\sigma(A_n, v_n, x_n, \Sigma) \geq \underline{\sigma}$ for all $n \in \mathbb{N}$:

$$\lim_{n\to\infty} \mathbb{P}[H_n(A_n, v_n, x_n) \leq z_{1-\alpha}] = 1 - \alpha \tag{153}$$

Lemma 6.5. There exists a measurable sequence $\breve{x}$, such that $(\hat{A}, v_n, \breve{x})$ is an optimal triplet w.p.a.1 and $\rho_\infty((\hat{A}, \hat{v}, \hat{x}), (\hat{A}, \dot{v}_n, \breve{x})) = o_p(1)$.

Proof. For $A \in 2^{[q]} \setminus \{\emptyset\}$ and $\varepsilon \geq 0$, define

$$\mathbb{X}(A, \varepsilon) \equiv \{x \in \mathbb{R}^d : p'x = B(\theta_0), Mx \geq c, ||M_A x - c_A|| \leq \varepsilon\},$$

From the assumption of Theorem (2.3), that $\mathcal{A}(\theta_0) \subseteq \text{Int}(\mathcal{X})$, it follows that $\mathbb{X}(A, \varepsilon) \subseteq \mathcal{A}(\theta_0)$ if $A \in \mathbb{A}$. Further, define

$$\overline{\mathbb{X}}(A, \varepsilon) \equiv \{x \in \mathbb{R}^d : p'x = B(\theta_0), Mx \geq c, M_A x - c_A \leq \varepsilon \iota_{|A|}\}.$$

Observe that for any $A \in \mathbb{A}$ and $\varepsilon \geq 0$, $\mathbb{X}(A, \varepsilon), \overline{\mathbb{X}}(A, \varepsilon)$ are nonempty, and

$$\mathbb{X}(A, \varepsilon) \subseteq \overline{\mathbb{X}}(A, \varepsilon),$$

so

$$\mathrm{d}_H(\mathbb{X}(A, \varepsilon), \mathbb{X}(A, 0)) = \max\{\sup_{x \in \mathbb{X}(A,\varepsilon)} \mathrm{d}(x, \mathbb{X}(A, 0)), \sup_{x \in \mathbb{X}(A,0)} \mathrm{d}(x, \mathbb{X}(A, \varepsilon))\} =$$
$$\sup_{x \in \mathbb{X}(A,\varepsilon)} \mathrm{d}(x, \mathbb{X}(A, 0)) \leq \sup_{x \in \overline{\mathbb{X}}(A,\varepsilon)} \mathrm{d}(x, \mathbb{X}(A, 0)) = \mathrm{d}_H(\overline{\mathbb{X}}(A, \varepsilon), \mathbb{X}(A, 0)) \leq C|A|\varepsilon,$$

where the last inequality, for some $C > 0$, follows from Lipschitz-continuity of polytopes with respect to the RHS perturbations, see Li (1993). We conclude that

$$\mathrm{d}_H(\mathbb{X}(A, \varepsilon), \mathbb{X}(A, 0)) \leq C|A|\varepsilon \tag{154}$$

From the proof of Therem (2.2) it follows that the projection $\tilde{x}$ of $\hat{x}$ onto $\mathcal{A}(\theta_0)$ is such that

$$||\tilde{x} - \hat{x}|| = o_p(1)$$

By triangle and Cauchy-Schwartz inequalities,

$$\begin{aligned}
||M_{\hat{A}}\tilde{x} - c_{\hat{A}}|| \leq ||M_{\hat{A}}|| \cdot ||\tilde{x} - \hat{x}|| + ||M_{\hat{A}}\hat{x} - c_{\hat{A}}|| = \\
||M_{\hat{A}}|| \cdot ||\tilde{x} - \hat{x}|| + ||M_{\hat{A}}\hat{x} - \hat{M}^{(1)}_{\hat{A}}\hat{x} + \hat{c}^{(1)}_{\hat{A}} - c_{\hat{A}}|| \leq \\
||M_{\hat{A}}|| \cdot ||\tilde{x} - \hat{x}|| + ||M_{\hat{A}} - \hat{M}^{(1)}_{\hat{A}}|| \cdot ||x||_\infty + ||\hat{c}^{(1)}_{\hat{A}} - c_{\hat{A}}|| \leq \\
||M|| \cdot ||\tilde{x} - \hat{x}|| + ||M - \hat{M}^{(1)}|| \cdot ||x||_\infty + ||\hat{c}^{(1)} - c||,
\end{aligned}$$

because the right-hand side vanishes in probability, it follows that for any $\varepsilon > 0$, $\tilde{x} \in \mathbb{X}(\hat{A}, \varepsilon)$ w.p.a.1. Denote the projection of $\tilde{x}$ onto $\mathbb{X}(\hat{A}, 0)$ by $\breve{x}$. It is measurable by the usual arguments, and, by (154),

$$||\tilde{x} - \breve{x}|| \leq C|A| \cdot ||M_{\hat{A}}\tilde{x} - c_{\hat{A}}|| = o_p(1)$$

Finally, by triangle inequality,

$$||\hat{x} - \breve{x}|| \leq ||\hat{x} - \tilde{x}|| + ||\tilde{x} - \breve{x}|| = o_p(1). \tag{155}$$

Observe that $\mathbb{X}(A, 0)$ for $A \in \mathbb{A}$ is the set of $x \in \mathcal{A}(\theta_0)$ that satisfy the respective requirements of an optimal triplet jointly with A. Thus, whenever $\hat{A} \in \mathbb{A}$, $\breve{x}, \hat{A}, v_n$ form an optimal triplet,

which occurs w.p.a.1. Combining (155) and (138),

$$\rho_\infty((\hat{A},\hat{v},\hat{x}),(\hat{A},v_n,\breve{x}))=o_p(1)$$

This concludes the proof of the Lemma. ■

We now show that we can pick $\underline{\sigma}$ and $\underline{v}$ such that the event

$$E_n\equiv\{\sigma(\hat{A},\hat{v},\hat{x},\Sigma)<\overline{\sigma}\}\cup\{||\hat{v}||<\underline{v}\}$$

vanishes asymptotically.

By Lemma 6.5, continuity of $\sigma(\cdot)$ in the first three arguments with respect to the ρ_∞ metric, and Assumption B4 combined with the fact that the set of optimal triplets with the additional requirement that $||v||\leq\overline{v}$ in Assumption B3 is compact, if we consider

$$\underline{\sigma}=0.5\min_{A,x,v-\text{optimal triplet, }||v||\leq\overline{v}}\sigma(A,x,v,\Sigma),$$

then

$$\mathbb{P}[\sigma(\hat{A},\hat{v},\hat{x},\Sigma)<\overline{\sigma}]\to 0.$$

For the second part of E_n, majorize

$$||p||=||M'\dot{v}_n||=||\dot{v}_n||\cdot||M'\frac{\dot{v}_n}{||\dot{v}_n||}||\leq\sigma_1(M)||\dot{v}_n||,$$

so that

$$||\dot{v}_n||\geq\frac{||p||}{\sigma_1(M)}.$$

Set $\underline{v}\equiv 0.5\frac{||p||}{\sigma_1(M)}$. Using (138), triangle inequality and recalling that $||\dot{v}_n||=||\tilde{v}||$, and $||\hat{v}||=||\check{v}||$ establishes

$$\mathbb{P}[||\hat{v}||<\underline{v}]\to 0,$$

so a union bound yields

$$\mathbb{P}[E_n]\to 0.$$

Note that

$$\mathbb{P}[H_n\leq z_{1-\alpha}|\mathcal{D}_n^{(1)}]=\mathbb{P}[H_n\leq z_{1-\alpha}|\hat{A},\hat{v},\hat{x}],\tag{156}$$

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 $\hat{A}, \hat{v}, \hat{x}$.

Observe that:

$$\mathbb{1}_{E'_n} \inf_{A,v,x \in \aleph(\underline{v},\underline{\sigma})} \mathbb{P}[H_n(A,v,x) \leq z_{1-\alpha}] \leq \mathbb{P}[H_n \leq z_{1-\alpha} | \hat{A}, \hat{v}, \hat{x}] \leq \tag{157}$$

$$\leq \sup_{A,v,x \in \aleph(\underline{v},\underline{\sigma})} \mathbb{P}[H_n(A,v,x) \leq z_{1-\alpha}] + \mathbb{1}_{E_n} \tag{158}$$

It follows that:

$$\mathbb{P}[H_n \leq z_{1-\alpha} | \hat{A}, \hat{v}, \hat{x}] = 1 - \alpha + o_p(1) \tag{159}$$

Therefore:

$$\mathbb{P}[H_n \leq z_{1-\alpha} | \hat{A}, \hat{v}, \hat{x}] = 1 - \alpha + o_p(1) \tag{160}$$

By Portmanteau and because probability is bounded, we can integrate (160) over $\mathcal{D}_n^{(1)}$,

$$\mathbb{P}[H_n \leq z_{1-\alpha}] = 1 - \alpha + o(1). \tag{161}$$

Finally, define

$$G_n \equiv \frac{\sqrt{n_2}}{\hat{\sigma}_n(\hat{A}, \hat{v}, \hat{x})} (\check{v} - \tilde{v}_n)'(c_{\hat{A}} - M_{\hat{A}} \hat{x}).$$

Using $||c_{\hat{A}} - M_{\hat{A}} \hat{x}|| = O_p(\frac{1}{\sqrt{n}})$ (see Proof of Lemma (6.5)), (138) and CMT, one concludes that $G_n = o_p(1)$. Applying Lemma 6.4 yields

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

Finally, because $G_n = o_p(1)$, we have, for any $\varepsilon > 0$,

$$o(1) + \mathbb{P}[H_n \leq z_{1-\alpha} - \varepsilon] \leq \mathbb{P}[H_n - G_n \leq z_{1-\alpha}] \leq \mathbb{P}[H_n \leq z_{1-\alpha} + \varepsilon] + o(1). \tag{162}$$

Letting $\alpha^+(\varepsilon) \equiv 1 - \Phi(z_{1-\alpha} - \varepsilon)$ and $\alpha^-(\varepsilon) \equiv 1 - \Phi(z_{1-\alpha} + \varepsilon)$, applying (161), one obtains:

$$o(1) + 1 - \alpha^+(\varepsilon) \leq \mathbb{P}[H_n - G_n \leq z_{1-\alpha}] \leq o(1) + 1 - \alpha^-(\varepsilon) \tag{163}$$

Taking $\varepsilon \to 0$ and using continuity of the normal's cdf, we obtain:

$$\mathbb{P}[H_n - G_n \leq z_{1-\alpha}] = 1 - \alpha + o(1) \tag{164}$$

To extend the proof to other consistent $\hat{\sigma}_n$, refer to CMT. This concludes the proof of the Theorem. ■

6.7. Asymptotic variance

Lemma 6.6. At fixed A, x, v,

$$\sigma^2(A, x, v, \Sigma) = J_1 \Sigma J_1' - 2J_2(I_d \otimes C_M \Sigma J_1')x + J_2 \left(xx' \otimes C_M \Sigma C_M'\right) J_2',$$

where

$$J_1 \equiv \check{v}' C(\hat{A}) C_c, \quad J_2 \equiv \check{v}' C(\hat{A})(\text{vec}(I_d)' \otimes I_q).$$

Proof.

$$\text{Var}\left(\check{v}' C(\hat{A}) \left(C_c Z - \text{vec}_{q\times d}^{-1}(C_M Z)\hat{x}\right)\right) = \tag{165}$$

$$= \text{Var}\left(\check{v}' C(\hat{A}) C_c Z\right) - 2\text{Cov}\left(\check{v}' C(\hat{A}) C_c Z, \check{v}' C(\hat{A}) \text{vec}_{q\times d}^{-1}(C_M Z)\hat{x}\right) + \tag{166}$$

$$+\text{Var}\left(\check{v}' C(\hat{A}) \text{vec}_{q\times d}^{-1}(C_M Z)\hat{x}\right) \tag{167}$$

where $Z \sim \mathcal{N}(0, \Sigma)$ has the asymptotic distribution of $Z_n^{(2)}$. The first term rewrites as:

$$\text{Var}\left(\check{v}' C(\hat{A}) C_c Z\right) = J_1 \Sigma J_1' \tag{168}$$

To deal with the last term, rewrite:

$$\text{Var}\left(\check{v}' C(\hat{A}) \text{vec}_{q\times d}^{-1}(C_M Z)\hat{x}\right) = J_2 \text{Var}\left(\left(I_d \otimes C_M Z\right)\hat{x}\right) J_2' \tag{169}$$

Direct computation yields:

$$(I_d \otimes C_M Z)\hat{x} = \begin{pmatrix} C_M Z \hat{x}_1 \\ C_M Z \hat{x}_2 \\ \cdots \\ C_M Z \hat{x}_d \end{pmatrix} \tag{170}$$

So:

$$\text{Var}\left(\left(I_d \otimes C_M Z\right)\hat{x}\right) = \hat{x}\hat{x}' \otimes C_M \Sigma C_M' \tag{171}$$

Consider:

$$\text{Cov}\left(\check{v}'C(\hat{A})C_cZ, \check{v}'C(\hat{A})\text{vec}^{-1}_{q\times d}(C_MZ)\hat{x}\right) = \mathbb{E}[J_1ZJ_2(I_d\otimes C_MZ)\hat{x}] = \tag{172}$$

$$= J_2\mathbb{E}[(I_d\otimes C_MZZ'J_1')]\hat{x} = J_2(I_d\otimes C_M\Sigma J_1')\hat{x} \tag{173}$$

Combining everything, we get:

$$\sigma(\hat{A},\hat{x},\hat{v},\Sigma) = J_1\Sigma J_1' - 2J_2(I_d\otimes C_M\Sigma J_1')\hat{x} + J_2\left(\hat{x}\hat{x}'\otimes C_M\Sigma C_M'\right)J_2' \tag{174}$$

We thus have, for fixed $\hat{A},\hat{v},\hat{x}$ with $\hat{v}\neq 0$. ■

6.8. Proof of Lemma 2.4

Proof. Let $\delta > 0$ be a jump at $\mathbb{P}_0$. Construct a sequence $\{\mathbb{P}_n\}\subset\mathcal{P}$ such that for some $0<\vartheta<1$:

$$||\mathbb{P}_0 - \mathbb{P}_n||_{TV} < \vartheta n^{-1} \tag{175}$$

While $||V(\mathbb{P}_0) - V(\mathbb{P}_n)|| > \delta$. Recall that:

$$||\mathbb{P}_0^n - \mathbb{P}_n^n||_{TV} \leq n||\mathbb{P}_0 - \mathbb{P}_n||_{TV} \tag{176}$$

It follows that:

$$||\mathbb{P}_0^n - \mathbb{P}_n^n||_{TV} \leq \vartheta \tag{177}$$

Using the binary Le Cam's method[35], one obtains $\forall n$:

$$\inf_{\hat{V}_n}\sup_{\mathbb{P}\in\mathcal{P}}\mathbb{E}_{\mathbb{P}}[||V(\mathbb{P}) - \hat{V}_n(X(\mathbb{P}^n))||] \geq \frac{\delta(1-\vartheta)}{2} \tag{178}$$

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. ■

6.9. Proof of Proposition 2.5

Proof. Consider the problem and its associated Lagrangean:

$$(P): \min_x\ p'x \quad \text{s.t.}: Mx\geq c, \quad \mathcal{L}\equiv p'x + \lambda'(c-Mx)$$

[35] See Chapter 15 of Wainwright (2019)

FOCs:

$$\begin{aligned}[x] &: p - M'\lambda = 0 \\ [\lambda] &: c - Mx \leq 0 \\ [\text{CS}] &: \lambda'(c - Mx) = 0 \\ [\text{POS}] &: \lambda \geq 0\end{aligned}$$

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

$$M_J x^* = c_J,$$

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

$$M_I x^* > c_I$$

It follows that $\lambda_I^* = 0$. Notice that the KKT condition that:

$$p = M_J' \lambda_J$$

for some $\lambda_J \geq 0$ means that $p \in \text{Cone}(M_J')$. By the conical hull version of Caratheodory's Theorem, it follows that $\exists J^* \subseteq J$ such that $|J^*| = r \leq d$ and $p \in \text{Cone}(M_J^{*\prime})$ and, moreover, the columns of 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 $\text{rk}(M_{J^*}) = d$, setting the appropriate λ_i^* to 0. By necessity and sufficiency of KKT for LP problems, this constitues a solution. ■

6.10. Proof of Theorem 2.5

Proof. We first establish a well-known Lemma.

Lemma 6.7. For any $A \in \mathbb{R}^{l \times m}$ and $b \in \mathbb{R}^m$ the following inequality holds:

$$||Ab||_\infty \leq ||A||_2 ||b|| = \sigma_1(A)||b||$$

Proof. Suppose $a_i,\ i \in [l]$ are rows of A . Then,

$$||Ab||_\infty = \max_i \{|(Ab)_i|\} = \max_i \{|a_i' b|\} \leq ||b|| \max_i ||a_i|| \tag{179}$$

Recall that the operator norm is transpose-invariant, and can be written as:

$$||A||_2 = ||A'||_2 = \sup_{||y||\leq 1} ||A'y|| \geq \max_i ||A'e_i|| = \max_i ||a_i|| \tag{180}$$

Combining (179) and (180) yields the result. ■

We now prove the Theorem. We write $M(\mathbb{P}), c(\mathbb{P})$ for components of $\theta_0(\mathbb{P})$. Fix $\delta > 0$. By definition of $\mathcal{P}^\delta$ and using Proposition 2.5, for any $\mathbb{P} \in \mathcal{P}^\delta$ there exists $J^* = J^*(\mathbb{P}, \delta) \subseteq [q]$ and the associated KKT vector $\lambda^* = \lambda^*(\mathbb{P}, \delta) \in \Lambda(\theta_0(\mathbb{P}))$, such that $M_{J^*} = M(\mathbb{P})_{J^*(\mathbb{P},\delta)}$ is invertible, and

$$\lambda^*_{J^*} = M_{J^*}^{-1\prime} p, \quad \sigma_1(M_{J^*}^{-1\prime}) = \sigma_d^{-1}(M_{J^*}) < \delta^{-1}.$$

Using Lemma 10, one observes that

$$||\lambda^*||_\infty \leq \delta^{-1}||p||.$$

One concludes that for any $\mathbb{P} \in \mathcal{P}^\delta$,

$$E_n \equiv \{w_n > \delta^{-1}||p||\} \subseteq \{\tilde{B}(\theta_0(\mathbb{P}), w_n) = B(\theta_0(\mathbb{P}))\}. \tag{181}$$

In what follows, we denote $B = B(\theta_0(\mathbb{P}))$, $\tilde{B} = \tilde{B}(\theta_0(\mathbb{P}); w_m(\mathbb{P}))$ and $\tilde{B}_m = \tilde{B}(\hat{\theta}_m(\mathbb{P}); w_m(\mathbb{P}))$, and $\tilde{r}_m \equiv \frac{r_m}{w_m}$. Furthermore, let $F_n \equiv \{\inf_{m\geq n} 1_{E_m} = 1\}$. Consider

$$\begin{aligned}
\mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left|\tilde{B}_m - B\right| > \varepsilon\right] = \mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left(\mathbb{1}_{E_m}\left|\tilde{B}_m - \tilde{B}\right| + \mathbb{1}_{E_m'}|\tilde{B}_m - B|\right) > \varepsilon\right] = \\
\mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left|\tilde{B}_m - \tilde{B}\right| > \varepsilon, F_n\right] + \\
\mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left(\mathbb{1}_{E_m}\left|\tilde{B}_m - \tilde{B}\right| + \mathbb{1}_{E_m'}|\tilde{B}_m - B|\right) > \varepsilon, F_n'\right] \leq \\
\mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left|\tilde{B}_m - \tilde{B}\right| > \varepsilon\right] + \mathbb{P}\left[F_n'\right].
\end{aligned} \tag{182}$$

Using (182), the fact that for any sequences g_k, h_k, $\sup_k\ g_k + h_k \leq \sup_k g_k + \sup_k h_k$ and the fact that $\sup A \leq \sup B$ whenever $A \subseteq B$, we get

$$\sup_{\mathbb{P}\in\mathcal{P}^\delta} \mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left|\tilde{B}_m - B\right| > \varepsilon\right] \leq \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left|\tilde{B}_m - \tilde{B}\right| > \varepsilon\right] + 1 - \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left[F_n\right] \tag{183}$$

Observe that

$$\mathbb{P}[F_n] = \mathbb{P}[\cap_{m\geq n} E_m] \geq \mathbb{P}[\inf_{m\geq n} w_m > \delta^{-1}||p||].$$

Using this and condition ii), taking limits in (183) yields

$$\lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}^\delta} \mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left|\tilde{B}_m - B\right| > \varepsilon\right] \leq \lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left|\tilde{B}_m - \tilde{B}\right| > \varepsilon\right]. \tag{184}$$

From the bound in (47) and condition i) it follows that

$$\lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left|\tilde{B}_m - \tilde{B}\right| > \varepsilon\right] = 0 \tag{185}$$

Combining (184), (185) and recalling that $\delta > 0$ was arbitrary, one can take suprema on both sides, as

$$\sup_{\delta>0} \lim_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}^\delta} \mathbb{P}\left[\sup_{m\geq n} \tilde{r}_m \left|\tilde{B}_m - B\right| > \varepsilon\right] = 0$$

This concludes the proof of the Theorem.

■

6.11. Proof of Lemma 2.5

Proof. If $x \in \Theta$, 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:

$$\min_{y\in\Theta} \frac{1}{2}(y - x)'(y - x) \tag{186}$$

Construct Lagrangean:

$$\mathcal{L} = (y - x)'(y - x) + \lambda'(c - My) \tag{187}$$

FOCs:

$$y - x - M'\lambda = 0 \tag{188}$$

$$\lambda_j(M_j'y - c_j) = 0 \tag{189}$$

$$My \geq c \tag{190}$$

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:

$$J \equiv \{j \in \overline{1,q} | M_j' y^* = c_j\} \tag{191}$$

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

$$f^* = \bigcap_{f-\text{face of } \Theta_I:\ y \in f} f, \tag{192}$$

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

$$y - x \in \text{Cone}(M_J'), \tag{193}$$

Therefore, by Caratheodory's Conical Hull theorem, there exists a subset $J^* \subseteq J$ such that $|J^*| = r \leq d - k^*$ and a corresponding $\lambda_{J^*}^* > 0$:

$$y - x = M_{J^*}' \lambda_{J^*}^* \tag{194}$$

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 $J^{**} \supseteq J^*$, such that: $|J^{**}| = \text{rk}(M_{J^{**}}) = d - k^*$. Finally, setting $\lambda^{**} \equiv \lambda_{J^{**}}^*$, we get:

$$y^* - x = M_{J^{**}}' \lambda^{**} \tag{195}$$

From where it follows that:

$$\lambda^{**} = (M_{J^{**}} M_{J^{**}}')^{-1} M_{J^{**}} (y^* - x) \tag{196}$$

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

$$||(M_{J^{**}} M_{J^{**}}')^{-1} M_{J^{**}}|| \leq \sigma_{d-k^*}^{-1}(M_{J^{**}}) \leq \kappa^{-1}(\Theta_I) \tag{197}$$

By Cauchy-Schwarz, we then obtain:

$$||\lambda^{**}|| \leq \varepsilon \kappa^{-1}(\Theta_I) \tag{198}$$

Since $||\lambda^{**}||_\infty \leq ||\lambda^{**}||$, it also follows that:

$$||\lambda^{**}||_\infty \leq \varepsilon\kappa^{-1}(\Theta_I) \tag{199}$$

Further, since $M_{J^{**}}y^* = c_{J^{**}}$ by construction, multiplying both sides of (195) by $M_{J^{**}}$ yields:

$$c_{J^{**}} - M_{J^{**}}x = M_{J^{**}}M'_{J^{**}}\lambda^{**} \tag{200}$$

And plugging (195) into the value function, one gets:

$$\varepsilon^2 = \lambda'_{J^{**}}M_{J^{**}}M'_{J^{**}}\lambda^{**} = \lambda'_{J^{**}}(c_{J^{**}} - M_{J^{**}}x) \tag{201}$$

Combining (201), the fact that at least one of the components of $\lambda^*_{J^{**}}$ is positive from $y^* - x \neq 0$ and (195), 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:

$$\exists j \in J^{**} : (c - M_{J^{**}}x)_j \geq \frac{\kappa(\Theta_I)\varepsilon}{d - k^*} \tag{202}$$

It then follows that:

$$\iota'(c - Mx)^+ \geq \frac{\kappa(\Theta_I)\varepsilon}{d - k^*} \tag{203}$$

Taking the minimum over k^* yields the claim of the proposition. ■

6.12. Proof of Theorem 2.6

Proof. Note that:

$$p'x^*_n - B(\theta_0) \geq \min_{x \in \Theta_I^{d(x^*_n, \Theta_I)}} p'x - \min_{x \in \Theta_I} p'x =$$

$$||p|| \left(s\left(\frac{p}{||p||}, \Theta_I^{d(x^*_n, \Theta_I)} \right) - s\left(\frac{p}{||p||}, \Theta_I \right) \right) \geq$$

$$-||p|| \max_{||y|| \leq 1} \left| s\left(y, \Theta_I^{d(x^*_n, \Theta_I)} \right) - s\left(y, \Theta_I \right) \right| =$$

$$-||p|| d_H\left(\Theta_I^{d(x^*_n, \Theta_I)}, \Theta_I \right) \geq -\frac{||p|| d\iota'(c - Mx^*_n)^+}{\kappa(\Theta_I)}$$

Thus:

$$O_p(1) = \frac{\sqrt{n}}{w_n}\left(p'x_n^* - B(\theta_0) + w_n\iota'(\hat{c}_n - \hat{M}_n x_n^*)\right) =$$
$$\frac{\sqrt{n}}{w_n}\left(p'x_n^* - B(\theta_0) + w_n\iota'(c - Mx_n^*)\right) + O_p(1) \geq$$
$$\sqrt{n}(1 - \frac{1}{w_n}\frac{||p||d}{\kappa(\Theta_I)})\iota'(c - Mx_n^*)^+ + O_p(1)$$

From where it follows that:

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

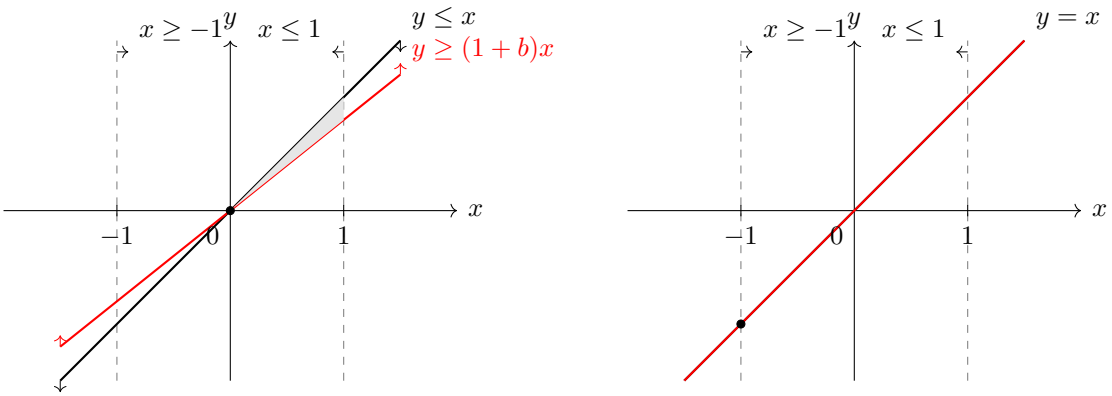
Using:

$$\frac{\sqrt{n}}{w_n}\left(p'x_n^* - B(\theta_0) + w_n\iota'(\hat{c}_n - \hat{M}_n x_n^*)^+\right) \geq \frac{\sqrt{n}}{w_n}(p'x_n^* - B(\theta_0)) \geq$$
$$\frac{-1}{w_n}\frac{\sqrt{n}||p||d\iota'(c - Mx_n^*)^+}{\kappa(\Theta_I)}$$

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 $\delta > 0$. ■

6.13. 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:



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

Figure 12: $B(b) = \min_{x,y} x \quad \text{s.t.} : y \geq (1+b)x,\ y \leq x,\ x \in [-1; 1]$

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 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 $-\hat{b}_n^{-1}$. 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:

$$s^2 \equiv \frac{1}{Qd-1} \sum_{i,j} (\hat{M}_{ij} - \overline{\hat{M}_{..}})^2, \tag{204}$$

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 6.1 (Tao and Vu (2010))**.** Let Ξ_n be a sequence of $n \times n$ matrices with $[\Xi_n]_{ij} \sim \xi_{ij}$, independently across i, j where ξ_{ij} are such that $\mathbb{E}[\xi] = 0$, $Var(\xi) = 1$ and $\mathbb{E}[|\xi|^{C_0}] < \infty$ for some sufficiently large C_0, then:

$$\sqrt{n}\sigma_n(\Xi_n) \xrightarrow{d} \Sigma \tag{205}$$

with the cdf of Σ given by:

$$\mathbb{P}[\Sigma \leq t] = 1 - e^{-t/2-\sqrt{t}} \tag{206}$$

Remark 6.3. 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:

$$w_n = ||\hat{p}_n||\delta^{-1}d_n \tag{207}$$

$$\delta = \frac{\left(\sqrt{1 - 2\ln(1-\alpha)} - 1\right)^2}{\sqrt{d}} \tag{208}$$

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 **??**.

6.14. Proof of Theorem ??

Proof. We first show that:

$$\Theta^* = \{\beta \in \mathbb{R} | \exists x \in \Theta_I : \beta = p'x + \overline{p}'\overline{x}\} \tag{209}$$

Fix $x \in \Theta_I$. It follows that the quantity $m = P_m x + \overline{P}_m \overline{x}$ satisfies (19) by construction. To see that there exists at least one $P \in \mathcal{P}$ 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:

$$Y(t) = \mathbb{I}\{t \in \mathcal{O}\}(\mathbb{I}\{T \neq t\}f(t, T, Z) + \mathbb{I}\{T = t\}\eta(t)) + \mathbb{I}\{t \in \mathcal{U}\}f(t, T, Z), \tag{210}$$

where $f : \mathcal{T}^2 \times \mathcal{Z} \to \mathbb{R}$ 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: $\mathbb{E}[Y(t)|T = d, Z = z]$ 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: $F_{\eta(t)|T=t,Z=z}(y) = F_{Y(t)|T=t,Z=z}(y),\ \forall y \in \mathbb{R}$ and $\forall t \in \mathcal{O}$, $\forall z \in \mathcal{Z}$. By construction, this DGP generates $m(P) = m$ and delivers the required identified distribution across observed dimensions, $F_{Y|T=t,Z}(\cdot)$ for $t \in O$. Therefore:

$$x \in \Theta_I \implies \mu^{*\prime}(P_m x + \overline{P}_m \overline{x}) = p'x + \overline{p}'\overline{x} \in \Theta^* \tag{211}$$

The other direction holds by construction: $\forall \beta \in \Theta^*\ \exists x \in \Theta_I : p'x + \overline{p}'\overline{x} = \beta$.
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 $\beta_0, \beta_1 \in \Theta^*$ with $\beta_0 < \beta_1$, then $\exists x_0, x_1 \in \Theta_I$ such that $\beta_i = p'x_i + \overline{p}'\overline{x}$ for $i = 0, 1$. Because Θ_I is convex, for arbitrary $\beta \in [\beta_0, \beta_1]$ setting $\alpha = \frac{\beta_1 - \beta}{\beta_1 - \beta_0}$, one obtains $\alpha x_0 + (1-\alpha)x_1 \in \Theta_I \implies \beta \in \Theta^*$. ■

6.15. Proof of Theorem ??

Lemma 6.8. Fix $K_0, \mu_v, \mu_w, K_1 \in \mathbb{R}$: $K_0 \leq \mu_v \leq \mu_w \leq K_1$ and $F_w(\cdot)$ that is a valid c.d.f. with expectation μ_w. Suppose the probability space $(P, \Omega, \mathcal{S})$ 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 $\mathbb{E}[V] = \mu_v$.

Proof. Proof. Suppose $\mu_w > K_0$ as otherwise the statement is trivial. W can be represented as:

$$W = F_w^{-1}(U) \tag{212}$$

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:

$$\int x dG(x) = K_0 \tag{213}$$

Moreover, by linearity of the Lebesgue integral $\forall \alpha \in [0;1]$ we have:

$$\int x d(\alpha G(x) + (1-\alpha)F_w(x)) = \alpha K_0 + (1-\alpha)\mu_w \tag{214}$$

Let $F_v(x) \equiv \alpha^* G(x) + (1-\alpha^*)F_w(x)$ where $\alpha^* \equiv \frac{\mu_w - \mu_v}{\mu_w - K_0}$. Then, notice that:

$$V = F_v^{-1}(U) \tag{215}$$

Yields the required random variable. ■

To prove the inverse inclusion in (**??**) for some $\tilde{M}, \tilde{b}$, note that from Theorem 1:

$$\{\beta \in \mathbb{R} | \exists P \in \mathcal{P} : \beta = \mu^{*\prime} m(P) \wedge b^{**} + M^{**} m(P) \geq 0\} = \{\beta \in \mathbb{R} | x : Mx \geq b : \beta = p'x + \overline{p}'\overline{x}\} \tag{216}$$

Where:

$$\overline{p} \equiv \overline{P}'_m \mu^*, \quad p \equiv P'_m \mu^* \tag{217}$$

$$M \equiv M^{**} P_m \quad b \equiv -b^{**} - M^{**} \overline{P}_m \overline{x} \tag{218}$$

Therefore proving the inclusion consists in finding such data-consistent $\mathbb{Y}$ (or, equivalently, the measure $P \in \mathcal{P}$) for any given $x : Mx \geq b$ that it generates $m(P) = p'x + \overline{p}'\overline{x}$ with $M^{**} m(P) + b^{**} \geq 0$ and $\tilde{M}\mathbb{Y} \geq \tilde{b}$ 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:

$$Y(t) = \mathbb{I}\{t \in \mathcal{O}\}(\mathbb{I}\{T \neq t\} f(t, T, Z) + \mathbb{I}\{T = t\}\eta(t)) + \mathbb{I}\{t \in \mathcal{U}\} f(t, T, Z), \tag{219}$$

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, $b^{**} + M^{**} m(P) \geq 0$ and $P \in \mathcal{P}$ for this P holds by construction, and: $Y(t) \in [K_0; K_1]$ $\forall t \in \mathcal{T}$ a.s., therefore $\tilde{M}\mathbb{Y} \geq \tilde{b}$ a.s. by construction.

2) MTR In this case it is clear that (219) fails, because it does not necessarily satisfy monotonicity almost surely. Consider:

$$\begin{aligned} \mathbb{Y} = (\mathbb{I}\{t \in \mathcal{O}\}(\mathbb{I}\{T \neq t\} f(t, T, Z) + \mathbb{I}\{T = t\}\eta(t)) + \mathbb{I}\{t \in \mathcal{U}\} f(t, T, Z))_{t \in \mathcal{T}} + \\ + \sum_{t \in \mathcal{O}} (\iota_{N_T} - e_t)\mathbb{I}\{T = t\}(\eta(t) - \mathbb{E}[Y(t)|T = t, Z]) \end{aligned} \tag{220}$$

Where e_t is the standard basis vector with 1 in the position of the potential outcome corresponding to t in $\mathbb{Y}$. Notice that the process in (220) has the same conditional means as the deterministic process of form (219), and therefore the corresponding $m(P)$ satisfies $M^{**} m(P) + b^{**} \geq 0$. Furthermore, by construction of M^{**} it must be that $\forall t \in \mathcal{O}$ and $\forall d \in \mathcal{T} : d \neq t$, we have:

$$\mathbb{E}[Y(d)|T = t, Z] = f(d, t, Z) \leq \mathbb{E}[Y(t)|T = t, Z] \text{ iff } d < t \tag{221}$$

and for $d_0, d_1 \in \mathcal{T} \setminus \{t\} : d_0 < d_1$:

$$\mathbb{E}[Y(d_0)|T = t, Z] = f(d_0, t, Z) \leq f(d_1, t, Z) = \mathbb{E}[Y(d_1)|T = t, Z] \tag{222}$$

Consider $\mathbb{Y}$ constructed in (220) over some element of the partition of Ω induced by T, where $T = t$.

i) If $t \in \mathcal{U}$, it is simply:

$$\mathbb{Y} = \begin{pmatrix} f(1,t,Z) \\ f(2,t,Z) \\ \cdots \\ f(N_T,t,Z) \end{pmatrix} \tag{223}$$

Which satisfies $\tilde{M}\mathbb{Y} + \tilde{b} \geq 0$ over this element of the partition a.s., by construction of f.

ii) If $t \in \mathcal{O}$:

$$\mathbb{Y} = \begin{pmatrix} f(1,t,z) + \eta(t) - \mathbb{E}[Y(t)|T=t,Z] \\ \cdots \\ f(t-1,t,z) + \eta(t) - \mathbb{E}[Y(t)|T=t,Z] \\ \mathbb{E}[Y(t)|T=t,Z] + \eta(t) - \mathbb{E}[Y(t)|T=t,Z] \\ f(t+1,t,z) + \eta(t) - \mathbb{E}[Y(t)|T=t,Z] \\ \cdots \\ f(N_T,t,z) + \eta(t) - \mathbb{E}[Y(t)|T=t,Z] \end{pmatrix} \tag{224}$$

Notice that by (221) and (222) the MTR is then satisfied, i.e. $\tilde{M}\mathbb{Y} + \tilde{b} \geq 0$.

3) MTR + Bounds It is clear that the process given in (220) 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 $t \in \mathcal{U}$ we can again set $\mathbb{Y}$ as in (223). Because each $f(d,t,Z)$ satisfies MTR and boundedness by construction, we have $\tilde{M}\mathbb{Y} + \tilde{b} \geq 0$ over this element of the T-partition.
Suppose $t \in \mathcal{O}$. The solution of the linear programming results in some moments that are given by our map $f(d,t,Z)$ that satisfies (221) and (222). Observe that constructing $\mathbb{Y}$ over the considered element of partition consists in constructing the counterfactual $Y(d)$ s.t. $d \in \mathcal{T} : d \neq t$ such that:

$$\mathbb{E}[Y(d)|T=t,Z] = f(d,t,Z) \ \forall d \in \mathcal{T} \setminus \{t\} \tag{225}$$

$$Y(1) \leq Y(2) \leq \cdots \leq Y(t) \leq \cdots \leq Y(N_T) \ a.s. \tag{226}$$

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 (225)-(226) hold jointly a.s. This concludes the proof of the Theorem. ■

6.16. Failure of the converse inclusion for almost sure inequalities

Consider a binary treatment $T \in \{0, 1\}$ and suppose we estimate the sharp lower bound for $\mathbb{E}[Y(1)|T = 0]$. Suppose that conditional on $T = 0$, $Y(0)$ is 1 and -1 with equal probability. Assume that there is the only conditional restriction that $\mathbb{E}[Y(1)|T = 0] \geq 0$. Further suppose that there is an almost sure restriction:

$$\begin{pmatrix} 1 & 1 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} Y(0) \\ Y(1) \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{227}$$

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 $\mathbb{E}[Y(1)|T = 0] \geq 1.5$. Taking the expectation of this system conditional on $T = 0$, however, yields that $\mathbb{E}[Y(1)|T = 0] = 0$ is a solution. Therefore, although 0 is a lower bound, it is not sharp.

6.17. 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 $\mathbb{R}^{n\times m}$ 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.

6.17.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, $\mathbb{E}[Y(t)|Z = z_i]$. Similarly, let $l_i^{-t}(t)$, $u_i^{-t}(t)$ be the sharp upper and lower bounds for the counterfactual subset, $\mathbb{E}[Y(t)|T \neq t, Z = z_i]$. 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 6.4. *If i) cMIV-w holds or ii) treatment is binary and cMIV-s or cMIV-p hold, the sharp bounds for $\mathbb{E}[Y(t)|Z = z_j]$ are obtained through the following recursion for $j \geq 2$:*

$$l_j = l_{j-1} + \Delta_j \tag{228}$$

$$l_j^{-t} = l_{j-1}^{-t} + \Delta_j^{-t} \tag{229}$$

Where $\Delta_j, \Delta_j^{-t} \geq 0$ are defined as follows:

$$\Delta_j \equiv \left(\underbrace{\frac{\Delta P[T \neq t|Z = z_j]}{P[T \neq t|Z = z_{j-1}]} \left(l_{j-1} - P[T = t|Z = z_{j-1}]\mathbb{E}[Y(t)|T = t, Z = z_{j-1}]\right)}_{\Delta P[T \neq t|Z = z_j] l_{j-1}^{-t}} + \delta_j \right)^+ \tag{230}$$

$$\Delta_j^{-t} \equiv \frac{1}{P[T \neq t|Z = z_j]} \left(-\Delta P[T \neq t|Z = z_j] l_{j-1}^{-t} - \delta_j\right)^+ \tag{231}$$

$$\delta_j \equiv \Delta(P[T = t|Z = z_j]\mathbb{E}[Y(t)|T = t, Z = z_j]) \tag{232}$$

Sharp upper bounds u_i, u_i^{-t} are obtained analogously. Moreover,

$$\sum_{i=1}^{N} P[Z = z_i] l_i(t) \leq \mathbb{E}[Y(t)] \leq \sum_{i=1}^{N} P[Z = z_i] u_i(t) \tag{233}$$

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, $l_1 = \mathbb{P}[T = t|Z = z_1]\mathbb{E}[Y(t)|T = t, Z = z_1] + \mathbb{P}[T \neq t|Z = z_1]K_0$, $u_N = \mathbb{P}[T = t|Z = z_N]\mathbb{E}[Y(t)|T = t, Z = z_N] + \mathbb{P}[T \neq t|Z = z_N]K_1$. First, we note that the equations above may be rearranged to yield:

$$l_j^{-t} = max\left\{ \frac{1}{P[T \neq t|Z = z_j]} \left(l_{j-1} - \mathbb{E}[Y(t)|T = t, Z = z_j]P[T = t|Z = z_j]\right),\ l_{j-1}^{-t} \right\} \tag{234}$$

$$l_j = \mathbb{E}[Y(t)|T = t, Z = z_j]P[T = t|Z = z_j] + l_j^{-t} P[T \neq t|Z = z_j] \tag{235}$$

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[36]:

$$\begin{cases} \mathbb{E}[Y(t)|Z = z_2] \geq \mathbb{E}[Y(t)|Z = z_1] \\ \mathbb{E}[Y(t)|T \neq t, Z = z_2] \geq \mathbb{E}[Y(t)|T \neq t, Z = z_1] \end{cases}$$

Which can be rewritten as a single condition on $\mathbb{E}[Y(t)|T \neq t, Z = z_2]$:

$$\begin{aligned} &\mathbb{E}[Y(t)|T \neq t, Z = z_2] \geq max\Big\{\mathbb{E}[Y(t)|T \neq t, Z = z_1], \\ &P[T \neq t|Z = z_2]^{-1}\Big(\mathbb{E}[Y(t)|Z = z_1] - P[T = t|Z = z_2]\mathbb{E}[Y(t)|T = t, Z = z_2]\Big)\Big\} \end{aligned}$$

Because l_1^{-t} is a sharp lower bound on $\mathbb{E}[Y(t)|T \neq t, Z = z_1]$, we get:

$$\begin{aligned} l_2^{-t} &= max\Big\{l_1^{-t},\ P[T \neq t|Z = z_2]^{-1}\Big(l_1 - P[T = t|Z = z_2]\mathbb{E}[Y(t)|T = t, Z = z_2]\Big)\Big\} \\ l_2 &= P[T = t|Z = z_2]\mathbb{E}[Y(t)|T = t, Z = z_2] + P[T \neq t|Z = z_2]l_2^{-t} \end{aligned}$$

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:

$$\begin{cases} \mathbb{E}[Y(t)|Z = z_j] \geq \mathbb{E}[Y(t)|Z = z],\ z < z_j \\ \mathbb{E}[Y(t)|T \neq t, Z = z_j] \geq \mathbb{E}[Y(t)|T \neq t, Z = z],\ z < z_j \end{cases}$$

Or, equivalently,

$$\begin{aligned} &\mathbb{E}[Y(t)|T \neq t, Z = z_j] \geq max\Big\{\max_{i<j}\left\{\mathbb{E}[Y(t)|T \neq t, Z = z_i]\right\}, \\ &P[T \neq t|Z = z_j]^{-1}\Big(\max_{i<j}\left\{\mathbb{E}[Y(t)|Z = z_i]\right\} - P[T = t|Z = z_j]\mathbb{E}[Y(t)|T = t, Z = z_j]\Big)\Big\} \end{aligned}$$

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:

$$\begin{aligned} l_j^{-t} &= max\Big\{l_{j-1}^{-t}, P[T \neq t|Z = z_j]^{-1}\Big(l_{j-1} - P[T = t|Z = z_j]\mathbb{E}[Y(t)|T = t, Z = z_j]\Big)\Big\} \\ l_j &= \mathbb{E}[Y(t)|T = t, Z = z_j]P[T = t|Z = z_j] + l_j^{-t}P[T \neq t|Z = z_j] \end{aligned}$$

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 $\mathbb{E}[Y(t)|T = d, Z = z_j] = l_j^{-t}, d \neq t, j \in \overline{1, N}$. For such process cMIV-w will hold by construction and l_j

[36]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

and l_j^{-t} are both attained for all j. An example of such process is given by:

$$Y(w) = \sum_t \mathbb{I}\{t = w\} \left(\sum_j \left\{ \mathbb{I}\{Z = z_j, T = t\}\, \eta(t) + \sum_{d \neq t} \mathbb{I}\{Z = z_j, T = d\}\, l_j^{-t} \right\} \right) \tag{236}$$

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 $\mathbb{E}[Y(t)|Z = z]$, which can be seen in equation (230). cMIV-w additionally 'irons' the counterfactual moments $\mathbb{E}[Y(t)|T \neq t, Z = z]$, as evident from (231). Figure 1 plots the derived sharp bounds as well as the benchmark MIV sharp bounds for a simulation exercise.

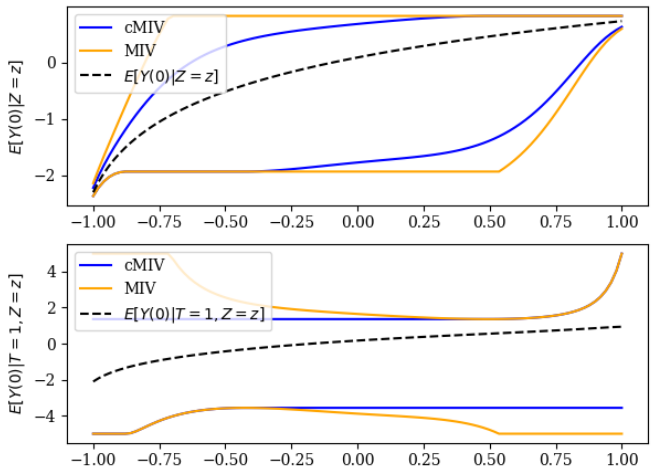


Figure 13: Bounds for the d.g.p. in Appendix 6.19.

6.17.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, $Q \equiv |\mathcal{F}| = 2^{N_T} - 2$. 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:

$$\mathbb{E}[Y(t)|T \in A^k, Z = z_j] \geq \mathbb{E}[Y(t)|T \in A^k, Z = z_{j-1}],\ k = 1, \dots, Q,\ j = 2, \dots N_Z \tag{237}$$

$$\mathbb{E}[Y(t)|T = d, Z = z_N] \leq K_1, d \in \mathcal{T} \setminus \{t\} \tag{238}$$

$$\mathbb{E}[Y(t)|T = d, Z = z_1] \geq K_0, d \in \mathcal{T} \setminus \{t\} \tag{239}$$

Where notice that the LHS of (238) is the largest marginal moment due to monotonicity in Z, while the LHS of (239) is the smallest marginal moment. Therefore, once almost sure bounds for these two moments are imposed $\forall d \in \mathcal{T} \setminus \{t\}$, these are also implied for all other moments through equation (237) and the law of total probability.

We now rewrite the expectations in (237) in terms of pointwise conditional moments. Let the vector of unobserved treatment responses be $x^j \equiv (\mathbb{E}[Y(t)|T = d, Z = z_j])'_{d\neq t}$ and $p^j \equiv (P[T = d|Z = z_j])'_{d\neq t}$ be the vector of respective probabilities at $Z = z_j$. Denote the element of x^j corresponding to $T = d$ as $x^j_d = \mathbb{E}[Y(t)|T = d, Z = z_j]$.

For $k = 1, \ldots, Q$ and $j = 2, \ldots, N_Z$, we can rewrite inequality (237) as follows:

$$\begin{aligned}
&\sum_{d\neq t} \mathbb{I}\left\{d \in A^k\right\} \frac{P[T = d|Z = z_j]}{P[T \in A^k|Z = z_j]} x^j_d + \\
+\mathbb{I}\left\{t \in A^k\right\} &\frac{P[T = t|Z = z_j]}{P[T \in A^k|Z = z_j]} \mathbb{E}[Y(t)|T = t, Z = z_j] \geq \\
&\geq \sum_{d\neq t} \mathbb{I}\left\{d \in A^k\right\} \frac{P[T = d|Z = z_{j-1}]}{P[T \in A^k|Z = z_{j-1}]} x^{j-1}_d + \\
+\mathbb{I}\left\{t \in A^k\right\} &\frac{P[T = t|Z = z_{j-1}]}{P[T \in A^k|Z = z_{j-1}]} \mathbb{E}[Y(t)|T = t, Z = z_{j-1}]
\end{aligned}$$

Inequalities (238)-(239) are just $x^N_d \leq K_1, d \neq t$ and $x^1_d \geq K_0, d \neq t$. This can be written succinctly in matrix notation. Introdude the following:

$$G_j \equiv \left(\mathbb{I}\left\{d \in A^k\right\} \frac{P[T = d|Z = z_j]}{P[T \in A^k|Z = z_j]}\right)_{k\in\overline{1,Q}, d\neq t} \in \mathbb{R}^{Q\times N_T - 1} \tag{240}$$

$$c_j \equiv \left(\mathbb{I}\left\{t \in A^k\right\} \frac{P[T = t|Z = z_j]}{P[T \in A^k|Z = z_j]} \mathbb{E}[Y(t)|T = t, Z = z_j]\right)_{k\in\overline{1,Q}} \in \mathbb{R}^Q \tag{241}$$

The whole set of information given by cMIV-s can be represented as follows:

$$G_j x^j - G_{j-1} x^{j-1} \geq -\Delta c_j, j = 2, \ldots, N_Z \tag{242}$$

$$x^N \leq K_1 \iota \tag{243}$$

$$x^1 \geq K_0 \iota \tag{244}$$

The procedure for cMIV-p is similar. First, we note that all the information under it is given

by:

$$\mathbb{E}[Y(t)|Z=z_j] \geq \mathbb{E}[Y(t)|Z=z_{j-1}],\ j=2,\dots N_Z \tag{245}$$

$$\mathbb{E}[Y(t)|T=d,Z=z_j] \geq \mathbb{E}[Y(t)|T=d,Z=z_{j-1}],\ d\in\mathcal{T}\setminus\{t\},\ j=2,\dots N_Z \tag{246}$$

$$\mathbb{E}[Y(t)|T=d,Z=z_N] \leq K_1, d\in\mathcal{T}\setminus\{t\} \tag{247}$$

$$\mathbb{E}[Y(t)|T=d,Z=z_1] \geq K_0, d\in\mathcal{T}\setminus\{t\} \tag{248}$$

Where (245) is just MIV and (246) is the monotonicity of the pointwise conditional moments. In this case, we can once again represent all information in the matrix form (242)-(244) with the following matrices:

$$G_j \equiv \begin{pmatrix} p^{j\prime} \\ I_{N_T-1} \end{pmatrix} \in \mathbb{R}^{N_T\times N_T-1} \tag{249}$$

$$c_j \equiv \begin{pmatrix} P[T=t|Z=z_j]\mathbb{E}[Y(t)|T=t,Z=z_j] \\ 0_{N_T-1} \end{pmatrix} \in \mathbb{R}^{N_T-1} \tag{250}$$

Corollary 2. Under cMIV-s and cMIV-p, sharp bounds on $\mathbb{E}[Y(t)]$ take the form:

$$\min_{Mx\geq c}\left\{\sum_{j=1}^{N} P[Z=z_j]\cdot p^{j\prime}x^j\right\} + \sum_{j=1}^{N} P[T=t,Z=z_j]\mathbb{E}[Y(t)|T=t,Z=z_j] \leq \mathbb{E}[Y(t)] \leq$$
$$\leq \max_{Mx\geq c}\left\{\sum_{j=1}^{N} P[Z=z_j]\cdot p^{j\prime}x^j\right\} + \sum_{j=1}^{N} P[T=t,Z=z_j]\mathbb{E}[Y(t)|T=t,Z=z_j],$$

where:

$$M \equiv \begin{bmatrix} -I_{N_T-1} & \dots & 0 & 0 \\ G_N & -G_{N-1} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & G_2 & -G_1 \\ 0 & \dots & 0 & I_{N_T-1} \end{bmatrix}, \quad c \equiv \begin{pmatrix} -K_1\cdot\iota_{N_T-1} \\ -\Delta c_N \\ \vdots \\ -\Delta c_2 \\ K_0\cdot\iota_{N_T-1} \end{pmatrix}, \quad x = \begin{pmatrix} x^N \\ \vdots \\ x^1 \end{pmatrix}, \tag{251}$$

and G_j and c_j are given by (240) and (241) for cMIV-s and by (249) and (250) for cMIV-p respectively.

6.18. Proof of Proposition 4.2

Let $\Gamma(z) \equiv \sum_{d\in\mathcal{T}} P[T=d|Z=z]\mathbb{E}[\psi(z,\eta)|Z=z]$.

a) Let $\tilde{g}(t) \equiv \mathbb{E}[g(t,\xi)|T=d,Z=z] = \mathbb{E}[g(t,\xi)|Z=z]$, where we use independence of ξ

and T, Z.
MIV implies:

$$\mathbb{E}[Y(t)|Z=z] = \tilde{g}(t) + h(t)\Gamma(z) \ - \ increasing \tag{252}$$

Since inequality is strict for some z, z', it follows that $h(t) \neq 0$ and $h(t)/h(d) > 0$. Note that:

$$\mathbb{E}[Y(t)|T=d, Z=z] - \tilde{g}(t) = \frac{h(t)}{h(d)} \left(\mathbb{E}[Y(d)|T=d, Z=z] - \tilde{g}(d)\right) \tag{253}$$

Therefore, cMIV-p holds iff all observed moments are monotone.

b) Let $\tilde{g}(t,d) \equiv \mathbb{E}[g(t,\xi)|T=d, Z=z]$, where we use independence of ξ and T, Z. We can write:

$$\mathbb{E}[Y(t)|T=d, Z=z] - \tilde{g}(t,d) = \frac{h(t)}{h(d)} \left(\mathbb{E}[Y(d)|T=d, Z=z] - \tilde{g}(d,d)\right) \tag{254}$$

Using b): ii) yields the result.

6.19. Simulation exercise

We now consider the following parametric example:

$$Y(t) = c + \alpha t + \beta\eta + Z \tag{255}$$
$$T = \mathbb{I}\{\varepsilon + f(Z) \geq 0\} \tag{256}$$
$$\eta = min\{u, max\{\varepsilon, l\}\} \tag{257}$$
$$\varepsilon \sim \mathcal{N}(0,1) \tag{258}$$

Where $u, l, \alpha, \beta, c \in \mathbb{R}$ and $u > l$. Moreover, ε is independent of all other variables. Also suppose for simplicity that $Z \in [l; u]$ a.s. Consider:

$$\mathbb{E}[Y(t)|T=1, Z=z] = c + \alpha t + z + \beta\mathbb{E}[min\{u, \varepsilon\}|\varepsilon > -f(z)] = \tag{259}$$
$$= c + \alpha t + z + \beta\left(\frac{1-\Phi(u)}{\Phi(f(z))}u + \frac{\phi(f(z)) - \phi(u)}{\Phi(f(z))}\right) \tag{260}$$
$$\mathbb{E}[Y(t)|T=0, Z=z] = c + \alpha t + z + \beta\mathbb{E}[max\{u, \varepsilon\}|\varepsilon \leq -f(z)] = \tag{261}$$
$$= c + \alpha t + z + \beta\left(\frac{\Phi(l)}{\Phi(-f(z))}l + \frac{\phi(l) - \phi(f(z))}{\Phi(-f(z))}\right) \tag{262}$$

For the Figure, suppose:

$$\begin{aligned} t &= 0 \\ [l, u] &= [-4, 2] \\ Z &\sim U[-1, 1] \\ f(z) &= -2z^4 \\ g(z) &= ln(z + 1.1) \\ \beta &= 0.1 \end{aligned}$$

6.20. 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.

6.21. 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 $\tilde{\mathcal{A}}(\theta; w)$ 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:

$$B(a,b,c,d) \equiv \min_{x,y} y - (1+a)x, \quad \text{s.t.:} \begin{cases} y \leq (1+b)x + d \\ y \geq (1+c)x \\ x \in [-1;1] \end{cases}, \tag{263}$$

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 $a \in [-0.1, 0.1]$.

2. For any fixed n, we take the grid of 9 points:

$$\mathcal{G}_n \equiv \{-0.1, 0, 0.1\} \cup \{-0.1C_1 n^{-1/2}, 0.1C_1 n^{-1/2}\} \cup \\ \{-0.1C_2 w_n n^{-1/2}, 0.1C_2 w_n n^{-1/2}\} \cup \{-0.1C_3 w_n^{-1}, 0.1C_3 w_n^{-1}\},$$

   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:

$$\sup_{a \in \mathcal{G}_n} |\tilde{B}(a, b_n, c_n, d_n; w_n) - B(a, 0, 0, 0)| \tag{264}$$

4. We then compute the standard deviation of (264) 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 14, it appears that standard deviations multiplied by $\sqrt{n}$ are indeed exploding, 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.

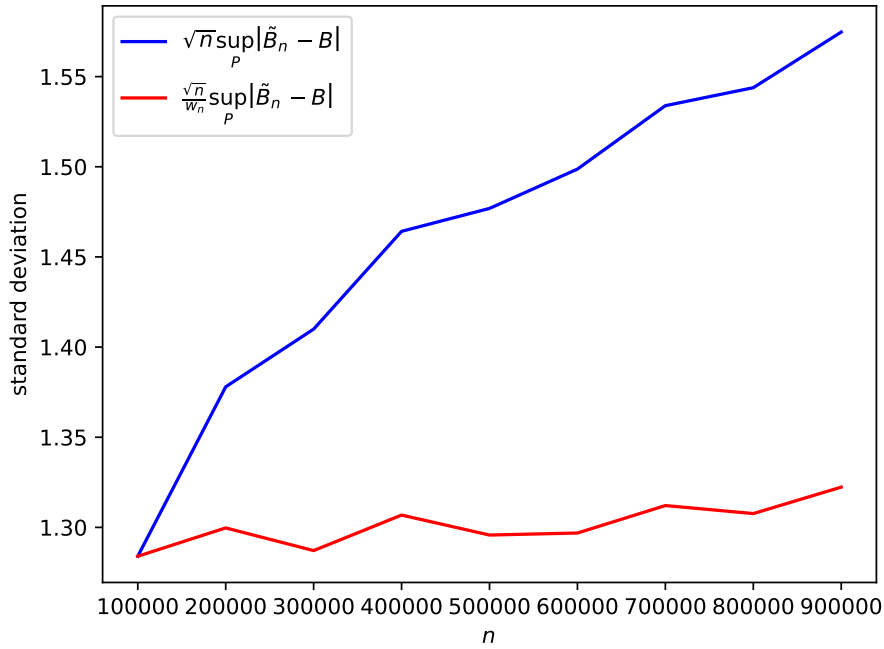


Figure 14: Uniformity of the penalized estimator: continuous vicinitiy of a flat face

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:

$$\mathcal{G}_n \equiv \{-0.1, 0, 0.1\} \cup \{-0.05(1 + C_1 n^{-1/2}), 0.05(1 + C_1 n^{-1/2})\} \cup$$
$$\{-0.05(1 + C_2 w_n n^{-1/2}), 0.05(1 + C_2 w_n n^{-1/2})\} \cup \{-0.05(1 + C_3 w_n^{-1}), 0.05(1 + C_3 w_n^{-1})\}$$

In this case, Figure 15 suggests that uniform $\sqrt{n}$-consistency is achieved.

Finally, we return to the original grid $\mathcal{G}_n$, 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.

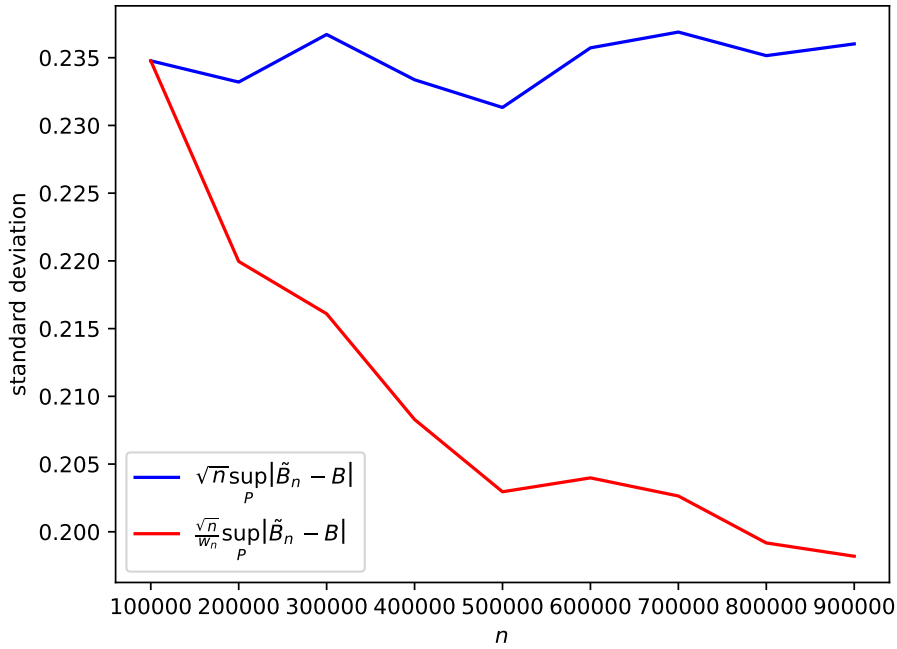


Figure 15: Uniformity of the penalized estimator: restricted vicinitiy of a flat face, flat face included.

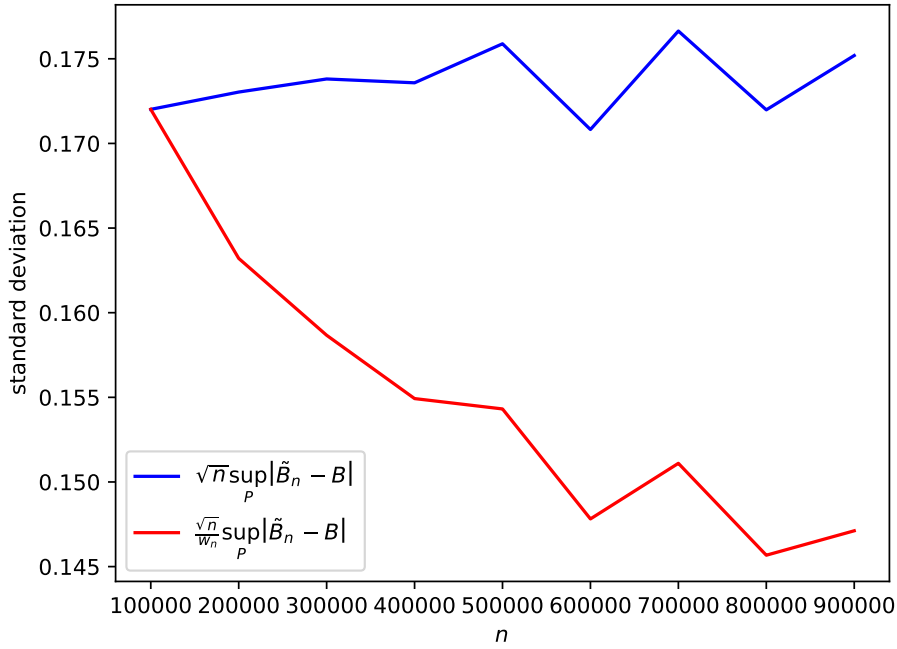


Figure 16: Uniformity of the penalized estimator: continuous vicinitiy of a flat face, Slater's condition holds.

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 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.