Confidence Intervals for Projections of Partially Identified Parameters *

Hiroaki Kaido† Francesca Molinari‡ Jörg Stoye§

January 4, 2016

Abstract

This paper proposes a bootstrap-based procedure to build confidence intervals for single components of a partially identified parameter vector, and for smooth functions of such components, in moment (in)equality models. The extreme points of our confidence interval are obtained by maximizing/minimizing the value of the component (or function) of interest subject to the sample analog of the moment (in)equality conditions properly relaxed. The novelty is that the amount of relaxation, or critical level, is computed so that the component (or function) of θ, instead of θ itself, is uniformly asymptotically covered with prespecified probability. Calibration of the critical level is based on repeatedly checking feasibility of linear programming problems, rendering it computationally attractive. Computation of the extreme points of the confidence interval is based on a novel application of the response surface method for global optimization, which may prove of independent interest also for applications of other methods of inference in the moment (in)equalities literature.

The critical level is by construction smaller (in finite sample) than the one used if projecting confidence regions designed to cover the entire parameter vector θ. Hence, our confidence interval is weakly shorter than the projection of established confidence sets (Andrews and Soares, 2010), if one holds the choice of tuning parameters constant. We provide simple conditions under which the comparison is strict. Our inference method controls asymptotic coverage uniformly over a large class of data generating processes. Our assumptions and those used in the leading alternative approach (a profiling based method) are not nested. We explain why we employ some restrictions that are not required by other methods and provide examples of models for which our method is uniformly valid but profiling based methods are not.

Keywords: Partial identification; Inference on projections; Moment inequalities; Uniform inference.

*We are grateful to Ivan Canay and seminar and conference participants at Bonn, BC/BU joint workshop, Brown, Cambridge, Chicago, Columbia, Cornell, Maryland, Michigan, Michigan State, NYU, Penn State, Syracuse, Toronto, UCSD, UPenn, Vanderbilt, Yale, Wisconsin, CEME, the Econometric Society Winter Meeting 2015, the Frontiers of Theoretical Econometrics Conference, and the Econometric Society World Congress 2015 for comments. We are grateful to Zhonghao Fu, Debi Prasad Mohapatra, and Sida Peng for excellent research assistance.

†Department of Economics, Boston University, hkaido@bu.edu. Financial support from NSF grant SES-1230071 is gratefully acknowledged.

‡Department of Economics, Cornell University, fm72@cornell.edu. Financial support from NSF grant SES-0922330 is gratefully acknowledged.

§Department of Economics, Cornell University, stoye@cornell.edu. Financial support from NSF grant SES-1260980 is gratefully acknowledged.

1 Introduction

A growing body of literature in econometric theory focuses on estimation and inference in partially identified models. For a given d-dimensional parameter vector θ characterizing the model, much work has been devoted to develop testing procedures and associated confidence sets in $\mathbb{R}^d$ that satisfy various desirable properties. These include coverage of each element of the d-dimensional identification region, denoted Θ_I, or coverage of the entire set Θ_I, with a prespecified –possibly uniform– asymptotic probability. From the perspective of researchers familiar with inference in point identified models, this effort is akin to building confidence ellipsoids for the entire parameter vector θ. However, applied researchers are frequently interested in conducting inference for each component of a partially identified vector, or for linear combinations of components of the partially identified vector, similarly to what is typically done in multiple linear regression.

The goal of this paper is to provide researchers with a novel procedure to conduct such inference in partially identified models. Our method yields confidence intervals whose coverage is *uniformly correct* in a sense made precise below. It is computationally relatively attractive because to compute critical levels, we check feasibility of a set of linear constraints rather than solving a linear or even nonlinear optimization problem.

Given the abundance of inference procedures for the entire parameter vector θ, one might be tempted to just report the projection of one of them as confidence interval for the projections of Θ_I (e.g., for the bounds on each component of θ). Such a confidence interval is asymptotically valid but typically conservative. The extent of the conservatism increases with the dimension of θ and is easily appreciated in the case of a point identified parameter. Consider, for example, a linear regression in $\mathbb{R}^{10}$, and suppose for simplicity that the limiting covariance matrix of the estimator is the identity matrix. Then a 95% confidence interval for each component of θ is obtained by adding and subtracting 1.96 to that component's estimate. In contrast, projection of a 95% Wald confidence ellipsoid on each component amounts to adding and subtracting 4.28 to that component's estimate. We refer to this problem as *projection conservatism*.

The key observation behind our approach is that projection conservatism can be anticipated. In the point identified case, this is straightforward. Returning to the example of multiple linear regression, if we are interested in a confidence interval with a certain asymptotic coverage for a component of θ, we can determine the level of a confidence ellipsoid whose projection yields just that confidence interval. When the limiting covariance matrix of the estimator is the identity matrix and $d = 2$, projection of a confidence ellipsoid with asymptotic coverage of 85.4% yields an interval equal to the component's estimate plus/minus 1.96, and therefore asymptotic coverage of 95% for that component; when $d = 5$, the required ellipsoid's coverage is 42.8%; when $d = 10$, the required ellipsoid's coverage is 4.6%.[1]

[1]The fast decrease in the required coverage level can be explained observing that the volume of a ball of

The main contribution of this paper is to show how this insight can be generalized to models that are partially identified through moment (in)equalities, while preserving computational feasibility and desirable coverage properties. The main alternative procedure in the literature, introduced in Romano and Shaikh (2008) and significantly advanced in Bugni, Canay, and Shi (2014, BCS henceforth), is based on profiling out a test statistic.[2] The classes of data generating processes (DGPs) over which our procedure and profiling-based methods are (pointwise or uniformly) valid are non-nested. The method proposed by Pakes, Porter, Ho, and Ishii (2011, PPHI henceforth) is based on bootstrapping the sample distribution of the projection.[3] This controls asymptotic coverage over a significantly smaller class of models than our approach.

Our approach ensures that asymptotic approximations to coverage are uniformly valid over a large class of models that we describe below. The importance of such uniformity in settings of partial identification was first pointed out by Imbens and Manski (2004), further clarified in Stoye (2009), and fully developed for moment (in)equalities models by Romano and Shaikh (2008), Andrews and Guggenberger (2009) and Romano and Shaikh (2010).[4] These authors show that poor finite sample properties may result otherwise. For example, consider an interval identified (scalar) parameter whose upper and lower bounds can be estimated. Then a confidence interval that expands each of the estimated bounds by a one-sided critical value controls the asymptotic coverage probability pointwise for any DGP at which the length of the identified set is positive. This is because the sampling variation becomes asymptotically negligible relative to the (fixed) length of the interval, making the inference problem essentially one-sided. However, this approximation is misleading in finite sample settings where sampling variation and the length of the interval are of comparable order. In such settings, coverage of the true parameter can fail when the true parameter falls below the lower bound of the confidence interval or above its upper bound; hence, a uniformly valid procedure must take into account the two-sided nature of the problem. More generally, uniformly valid inference methods need to account for inequalities that are close to be binding if not perfectly binding at the parameter of interest (Andrews and Guggenberger, 2009; Andrews and Soares, 2010; Bugni, 2009; Canay, 2010).

In our problem, uniformity is furthermore desirable along a novel dimension. Across DGPs, there can be substantial variation in the shape of the parameter set formed by the

radius r in $\mathbb{R}^d$ decreases geometrically in d.

[2]The profiling method provides uniformly valid confidence intervals also for nonlinear functions of θ. The corresponding extension of our method is addressed in Section 6 with our concluding remarks.

[3]This is the working paper version of Pakes, Porter, Ho, and Ishii (2015). We reference it because the published version does not contain the inference part.

[4]Universal uniformity is obviously unattainable (Bahadur and Savage, 1956). Other example of recent literatures where uniformity over broad, though not universal, classes of models is a point of emphasis include inference close to unit roots (Mikusheva, 2007), weak identification (Andrews and Cheng, 2012), and post-model selection inference (see Leeb and Pötscher 2005 for a negative take). See also the discussion, with more examples, in Andrews and Guggenberger (2009).

moment (in)equalities around each point in the identification region. Our analysis reveals that validity of inference and degree of projection conservatism depend crucially on the shape of the constraints in relation to the projection direction of interest, which we call the *local geometry* of the identification region. This is a novel dimension of uniformity which does not arise when one's interest is in the entire vector. We address this challenge by developing an inference method that is uniformly valid across various shapes formed by the constraints. To our knowledge, this is the first such effort.

This is also useful for achieving another desirable uniformity property. That is, holding one (reasonably well-behaved) model fixed, confidence regions should be equally valid for different directions of projection. It is surprisingly easy to fail this criterion. For example, if one does not properly account for flat faces which are orthogonal to the direction of projection, the resulting confidence interval will not be valid uniformly over directions of projection if the true identified set is a polyhedron. A polyhedron is not only a simple shape but also practically relevant: It arises for best linear prediction (BLP) with interval outcome data and discrete regressors, as shown in Beresteanu and Molinari (2008). In this example, a method that does not apply at (or near) flat faces is not equally applicable to all linear hypotheses that one might want to test. This stands in stark contrast to point identified BLP estimation: Barring collinearity, an F-test is applicable uniformly over simple linear hypotheses. Under this latter condition and some others, our method too applies uniformly over linear hypotheses, while other methods do not (PPHI assume away all flat faces that are near orthogonal to the direction of projection; BCS assume away many such cases).

Overview of the method. We consider models for which the identified set can be written as the set of parameter values that satisfy a finite number of moment equalities and inequalities, $\Theta_I = \{\theta : E(m(X_i, \theta)) \leq 0\}$.[5] Here X_i is a $d_X \times 1$ vector of random variables with distribution P and $m = (m_1, \ldots, m_J) : \mathbb{R}^{d_X} \times \Theta \to \mathbb{R}^J$ is a known measurable function of the finite dimensional parameter vector $\theta \in \Theta \subset \mathbb{R}^d$. We are interested in the projection $p'\theta$ of θ. We propose to report as confidence interval

$$CI_n = \left[\inf_{\theta \in \mathcal{C}_n(\hat{c}_n)} p'\theta, \sup_{\theta \in \mathcal{C}_n(\hat{c}_n)} p'\theta \right], \tag{1.1}$$

where

$$\mathcal{C}_n(\hat{c}_n) \equiv \left\{ \theta \in \Theta : n^{-1} \sum_{i=1}^{n} m_j(X_i, \theta)/\hat{\sigma}_{n,j}(\theta) \leq \hat{c}_n(\theta),\ j = 1, \ldots, J \right\}, \tag{1.2}$$

where $\hat{\sigma}_{n,j}$ is a suitable estimator of the asymptotic standard deviation of $n^{-1/2} \sum_i m_j(X_i, \theta)$.[6]

[5] We write equalities as two opposing inequalities in what follows. See section 2.1 for further elaboration.

[6] Our confidence region is by construction an interval. Conceptually, our method is easily adapted so as to capture gaps in the projection of the identified set. We recommend this only if one is genuinely interested in those gaps. Also, CI_n can be empty. We briefly discuss both matters in Section 6 with our concluding

Here, $\hat{c}_n(\theta)$ is loosely analogous to a critical value, though the reader should keep in mind that our confidence interval does not invert a hypothesis test. That said, one could use in the above construction, e.g., critical values $\hat{c}_n^{CHT}(\theta)$ or $\hat{c}_n^{AS}(\theta)$ from the existing literature (Chernozhukov, Hong, and Tamer, 2007; Andrews and Soares, 2010, respectively). These are calibrated so that $\mathcal{C}_n$ covers the entire vector θ and therefore any linear projection of it. Clearly, this is more than needed, and so projecting $\mathcal{C}_n(c)$ with $c = \hat{c}_n^{CHT}(\theta)$ or $c = \hat{c}_n^{AS}(\theta)$ is conservative. As we show below, this conservatism is severe in relevant examples. We (mostly) avoid it because we anticipate projection conservatism when calibrating $\hat{c}_n(\theta)$. In particular, for each candidate θ, we calibrate $\hat{c}_n(\theta)$ so that across bootstrap repetitions, the projection of θ is covered with at least some pre-specified probability. Computationally, this bootstrap is relatively attractive for two reasons: We linearize all constraints around θ, so that coverage corresponds to the projection of a stochastic linear constraint set covering zero.[7] We furthermore verify this coverage event without solving the linear program, but simply checking that a properly constructed linear constraint set is feasible.

The end points of our confidence interval can be obtained by solving constrained optimization problems for each direction of projection. The constraints of these problems involve $\hat{c}_n(\cdot)$, which in general is an unknown function of θ and, therefore, gradients of constraints are not available in closed form. When the dimension of the parameter is large, solving optimziation problems with such a component can be relatively expensive even if evaluating $\hat{c}_n(\cdot)$ at each point is computationally cheap. This is because commonly used optimization algorithms repeatedly evaluate the constraints and their (numerical) gradients. To overcome this challenge, we propose an algorithm that is a contribution to the moment (in)equalities literature in its own right and should also be helpful for implementing other approaches. Our algorithm is based on the response surface method (Jones, 2001) and computes the confidence interval as follows. First, it evaluates $\hat{c}_n(\cdot)$ on a coarse set of parameter values. Then, it fits a flexible auxiliary model (response surface) to the map $\theta \mapsto \hat{c}_n(\theta)$ to obtain surrogate constraint functions whose values and gradients are provided in closed form. Finally, it solves the optimization problems using the surrogate constraints. The algorithm then iterates these steps until the optimal values converge, while adding evaluation points to the set that contains parameter values that nearly attain the maximum (or minimum) and refining the surrogate constraints in each iteration. Computational savings come from the fact that the proposed method controls the number of evaluation points and the optimization problems only involve functions that are cheap to evaluate. Our Monte Carlo experiments show that this algorithm performs well even in a model with a moderately high number of parameters.

DGPs for which the method is uniformly valid. We establish uniform asymptotic validity of our procedure over a large class of DGPs that can be related to the existing

remarks.

[7] Previously, Pakes, Porter, Ho, and Ishii (2011) had also proposed local linear approximation to the moment inequalities.

literature as follows. We start from the same assumptions as Andrews and Soares (2010, AS henceforth), and similarly to the related literature, we ensure uniform validity in the presence of drifting-to-binding inequalities by adopting Generalized Moment Selection (AS, Bugni (2009), Canay (2010)). In addition, we impose some restrictions on the correlation matrix of the sample moment (in)equalities. A simple sufficient condition is that this matrix has eigenvalues uniformly bounded from below, an assumption that was considered in AS (for a specific criterion function) but eliminated by Andrews and Barwick (2012). It can be weakened substantially because we can allow for perfect or near perfect correlation of moment inequalities that are known not to cross; this case is relevant as it naturally occurs with missing-data bounds and static, simultaneous move, finite games with multiple equilibria. That said, profiling-based methods do not require any such assumption. We also assume that each individual constraint uniformly admits a local linear approximation that can be uniformly consistently estimated.

However, and in contrast to the leading alternative approaches, we do not impose further conditions that jointly restrict the inequality constraints, for example by restricting the local geometry of Θ_I. This is important because such assumptions, which are akin to constraint qualifications in nonlinear programming, can be extremely challenging to verify. Moreover, and again in contrast to leading alternative approaches, we do not impose restrictions on the limit distribution of a test statistic, e.g. continuity at the quantile of interest, which again can be challenging to verify. Our ability to dispense with such assumptions comes at the price of an additional, non-drifting tuning parameter. In Section 4, we explain why this additional parameter is needed and provide a heuristic for choosing it.

Going back to AS, our method can be directly compared to projection of their confidence region if one uses comparable tuning parameters. By construction, our confidence intervals are (weakly) shorter in any finite sample. They are asymptotically strictly shorter whenever at least one of the binding constraints is not locally orthogonal to the direction of projection.

Other related papers that explicitly consider inference on projections include Andrews, Berry, and Jia (2004), Beresteanu and Molinari (2008), Bontemps, Magnac, and Maurin (2012), Chen, Tamer, and Torgovitsky (2011), Kaido (2012), Kitagawa (2012), Kline and Tamer (2015) and Wan (2013). However, some are Bayesian, as opposed to our frequentist approach, and none of them establish uniform validity of confidence sets.

Structure of the paper. Section 2 sets up notation and describes our approach in detail, including computational implementation. Section 3 lays out our assumptions and presents our main theoretical results, namely uniform validity and a formal comparison to projection of the AS confidence region. Section 4 discusses the challenges posed by the local geometry of Θ_I for uniform inference and why we resolve them. In doing so, it further elucidates the relation between our method and the existing literature. Section 5 reports the results of Monte Carlo simulations. Section 6 offers concluding remarks and discusses a number of extensions that are of interest in applications. All proofs are collected in the Appendix.

2 Detailed Explanation of the Method

2.1 Setup and Definition of CI_n

We start by introducing some basic notation. Let $X_i \in \mathcal{X} \subseteq \mathbb{R}^{dx}$ be a random vector with distribution P, let $\Theta \subseteq \mathbb{R}^d$ denote the parameter space, and let $m_j : \mathcal{X} \times \Theta \to \mathbb{R}$ for $j = 1, \ldots, J_1 + J_2$ denote measurable functions characterizing the model, known up to parameter vector $\theta \in \Theta$. The true parameter value θ is assumed to satisfy the moment inequality and equality restrictions:

$$\begin{aligned} &E_P[m_j(X_i, \theta)] \leq 0, \ j = 1, \cdots, J_1, \\ &E_P[m_j(X_i, \theta)] = 0, \ j = J_1 + 1, \cdots, J_1 + J_2. \end{aligned} \tag{2.1}$$

The *identification region* $\Theta_I(P)$ is the set of parameter values in Θ that satisfy these moment restrictions. In what follows, we simply write Θ_I whenever its dependence on P is obvious. For a random sample $\{X_i, i = 1, \cdots, n\}$ of observations drawn from P, we let $\bar{m}_{n,j}(\theta) \equiv n^{-1} \sum_{i=1}^n m_j(X_i, \theta), j = 1, \cdots, J_1 + J_2$ denote the sample moments. Also, the population moment conditions have standard deviations $\sigma_{P,j}$ with estimators (e.g., sample analogs) $\hat{\sigma}_{n,j}$.

A key tool for our inference procedure is the support function of a set. We denote the unit sphere in $\mathbb{R}^d$ by $\mathbb{S}^{d-1} \equiv \{p \in \mathbb{R}^d : \|p\| = 1\}$, an inner product between two vectors $x, y \in \mathbb{R}^d$ by $x'y$, and use the following standard definition of support function and support set:

DEFINITION 2.1: *Given a closed set $A \subset \mathbb{R}^d$, its* support function *is*

$$s(p, A) = \sup\{p'a, \ a \in A\}, \ p \in \mathbb{S}^{d-1},$$

and its support set *is*

$$H(p, A) = \{a \in \mathbb{R}^d : p'a = s(p, A)\} \cap A, \ p \in \mathbb{S}^{d-1}.$$

It is useful to think of $p'a$ as a projection of $a \in \mathbb{R}^d$ to a one-dimensional subspace spanned by the direction p. For example, when p is a vector whose j-th coordinate is 1 and other coordinates are 0s, $p'a = a_j$ is the projection of a to the j-th coordinate. The support function of a set A gives the supremum of the projections of points belonging to this set.

The support function of the set $\mathcal{C}_n(\hat{c}_n)$ in equation (1.2) is, then, the optimal value of the following nonlinear program (NLP):

$$\begin{aligned} s(p, \mathcal{C}_n(\hat{c}_n)) = \sup_{\theta \in \Theta} \ & p'\theta \\ s.t. \quad & \sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta) \leq \hat{c}_n(\theta), \ j = 1, \cdots, J, \end{aligned} \tag{2.2}$$

where $J = J_1 + 2J_2$ and we define the last J_2 moments as $\bar{m}_{n,J_1+J_2+k}(\theta) = -\bar{m}_{J_1+k}(\theta)$ for

$k = 1, \cdots, J_2$. That is, we split moment equality constraints into two opposing inequality constraints relaxed by $\hat{c}_n(\theta)$ and impose them in addition to the first J_1 inequalities relaxed by the same amount. For a simple analogy, consider the point identified model defined by the single moment equality $E_P(m_1(X_i, \theta)) = E_P(X_i) - \theta = 0$, where θ is a scalar. In this case, $\mathcal{C}_n(\hat{c}_n) = \bar{X} \pm \hat{c}_n \hat{\sigma_n}/\sqrt{n}$. The upper endpoint of the confidence interval can be written as $\sup_\theta \{p'\theta \text{ s.t. } -\hat{c}_n \leq \sqrt{n}(\bar{X} - \theta)/\hat{\sigma_n} \leq \hat{c}_n\}$, with $p = 1$, and similarly for the lower endpoint.

Define the asymptotic size of the confidence interval by

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(p'\theta \in CI_n), \tag{2.3}$$

with $\mathcal{P}$ a class of distributions that we specify below. Our two-sided confidence interval is

$$CI_n \equiv [-s(-p, \mathcal{C}_n(\hat{c}_n)), s(p, \mathcal{C}_n(\hat{c}_n))], \tag{2.4}$$

and our goal is to calibrate $\hat{c}_n$ so that (2.3) is at least equal to a prespecified level while projection conservatism is anticipated. Unlike the simple adjustment of the confidence level for the Wald ellipsoid proposed in the introduction, however, the calculation of such a critical level in the moment (in)equalities setting is nontrivial, and it requires a careful analysis of the local behavior of the moment restrictions at each point in the identification region. This is because calibration of $\hat{c}_n(\theta)$ depends on (i) the asymptotic behavior of the sample moments entering the inequality restrictions, which can change discontinuously depending on whether they bind at θ or not; and (ii) the local geometry of the identification region at θ. Here, by local geometry, we mean the shape of the constraint set formed by the moment restrictions and its relation to the level set of the objective function $p'\theta$. These features can be quite different at different points in $\Theta_I(P)$, which in turn makes uniform inference for the projection challenging. In particular, the second issue does not arise if one only considers inference for the entire parameter vector, and hence this new challenge requires a new methodology. The core innovation of this paper is to provide a novel and computationally attractive procedure to construct a critical level that overcomes these challenges.

To build intuition, fix (θ, P) s.t. $\theta \in \Theta_I(P), P \in \mathcal{P}$. The projection of θ is covered if

$$\begin{aligned}
&-s(-p, \mathcal{C}_n(\hat{c}_n)) \leq p'\theta \leq s(p, \mathcal{C}_n(\hat{c}_n)) \\
&\Leftrightarrow \left\{ \begin{array}{ll} \inf_\vartheta p'\vartheta & \\ s.t.\, \vartheta \in \Theta, & \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leq \hat{c}_n(\vartheta), \forall j \end{array} \right\} \leq p'\theta \leq \left\{ \begin{array}{ll} \sup_\vartheta p'\vartheta & \\ s.t.\, \vartheta \in \Theta, & \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leq \hat{c}_n(\vartheta), \forall j \end{array} \right\} \\
&\Leftrightarrow \left\{ \begin{array}{ll} \inf_\lambda p'\lambda & \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta), & \frac{\sqrt{n}\bar{m}_{n,j}(\theta+\lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta+\lambda/\sqrt{n})} \leq \hat{c}_n(\theta + \lambda/\sqrt{n}), \forall j \end{array} \right\} \leq 0 \\
&\qquad \leq \left\{ \begin{array}{ll} \sup_\lambda p'\lambda & \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta), & \frac{\sqrt{n}\bar{m}_{n,j}(\theta+\lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta+\lambda/\sqrt{n})} \leq \hat{c}_n(\theta + \lambda/\sqrt{n}), \forall j \end{array} \right\}
\end{aligned} \tag{2.5}$$

where the second equivalence follows from rewriting the problem which maximizes $p'\vartheta$ with respect to ϑ localized as $\vartheta = \theta + \lambda/\sqrt{n}$ by another problem which maximizes the same objective function with respect to the localization parameter λ. One could then control asymptotic size by finding $\hat{c}_n$ such that 0 asymptotically lies within the optimal values of the NLPs in (2.5) with probability $1-\alpha$.

To reduce the computational cost of calibrating $\hat{c}_n$, we approximate the probability of the event in equation (2.5) by taking a linear expansion in λ of the constraint set. In particular, for the j-th constraint, adding and subtracting $E_P[m_j(X_i, \theta + \lambda/\sqrt{n})]$ yields

$$\begin{aligned}
&\frac{\sqrt{n}\bar{m}_{n,j}(\theta + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta + \lambda/\sqrt{n})} \\
&= \sqrt{n}\frac{(\bar{m}_{n,j}(\theta + \lambda/\sqrt{n}) - E_P[m_j(X_i, \theta + \lambda/\sqrt{n})])}{\hat{\sigma}_{n,j}(\theta + \lambda/\sqrt{n})} + \sqrt{n}\frac{E_P[m_j(X_i, \theta + \lambda/\sqrt{n})]}{\hat{\sigma}_{n,j}(\theta + \lambda/\sqrt{n})} \\
&= \{\mathbb{G}_{n,j}(\theta + \lambda/\sqrt{n}) + D_{P,j}(\bar{\theta})\lambda + \sqrt{n}\gamma_{1,P,j}(\theta)\}(1 + \eta_{n,j}(\theta_n)), \qquad (2.6)
\end{aligned}$$

where $\mathbb{G}_{n,j}(\cdot) \equiv \sqrt{n}(\bar{m}_{n,j}(\cdot) - E_P[m_j(X_i, \cdot)])/\sigma_{P,j}(\cdot)$ is a normalized empirical process indexed by $\theta \in \Theta$, $D_{P,j}(\cdot) \equiv \nabla_\theta\{E_P[m_j(X_i, \cdot)]/\sigma_{P,j}(\cdot)\}$ is the gradient of the normalized moment (a $1 \times d$ vector), $\gamma_{1,P,j}(\cdot) \equiv E_P[m_j(X_i, \cdot)]/\sigma_{P,j}(\cdot)$ is the studentized population moment, and $\eta_{n,j}(\cdot) \equiv \sigma_{P,j}(\cdot)/\hat{\sigma}_{n,j}(\cdot) - 1$. The second equality follows from the mean value theorem, where $\bar{\theta}$ represents a mean value that lies componentwise between θ and $\theta + \lambda/\sqrt{n}$.

Under suitable regularity conditions set forth in Section 3.1 (which include differentiability of $E_P[m_j(X_i, \theta)]/\sigma_{P,j}(\theta)$ in θ for each j), we show that the probability that 0 asymptotically lies within the optimal values of the NLPs in equation (2.5) is approximated by the probability that 0 asymptotically lies within the optimal values of a program linear in λ. The constraint set of this linear program is given by the sum of (i) an empirical process $\mathbb{G}_{P,j}(\theta)$ evaluated at θ (that we can approximate using the bootstrap) (ii) a rescaled gradient times λ, $D_{P,j}(\theta)\lambda$ (that we can uniformly consistently estimate on compact sets), and (iii) the parameter $\gamma_{1,P,j}(\theta)$ that measures the extent to which each moment inequality is binding and that we can conservatively estimate using insights from AS. This suggests a computationally attractive bootstrap procedure based on linear programs. We further show that introducing an additional linear constraint allows us to simply check feasibility of a linear program, without having to compute optimal values.

Our use of linearization to obtain a first-order approximation to the statistic of interest can be related to standard techniques in the analysis of nonlinear models. In our setting, the object of interest is the support function of the relaxed nonlinear constraint set. Calculating this support function subject to the moment (in)equality constraints is similar to calculating a nonlinear GMM estimator in the sense that both search for a particular parameter value which "solves" a system of sample moment restrictions. The difference is that we search for a parameter value satisfying suitably relaxed moment (in)equalities whose projection is

maximal, whereas GMM searches for a parameter value that minimizes the norm of sample moments, or necessarily a value that solves its first-order conditions. Hence, the solution concepts are different. However, the methodology for obtaining approximations is common. Recall that one may obtain an influence function of the GMM estimator by linearizing the moment restrictions in the first-order conditions around the true parameter value and by solving for the estimator. In analogy to this example, calculating the optimal value of the linear program discussed above can be interpreted as applying a particular solution concept (the maximum value of the linear projections) to a system of moment (in)equality constraints linearized around the parameter value of interest.

2.2 Computation of Critical Level

For a given $\theta \in \Theta$, we calibrate $\hat{c}_n(\theta)$ through a bootstrap procedure that iterates over linear programs (LP). Define

$$\Lambda_n^b(\theta, \rho, c) \quad = \quad \{\lambda \in \rho B^d : \mathbb{G}_{n,j}^b(\theta) + \hat{D}_{n,j}(\theta)\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta)) \leq c,\ j = 1, \ldots, J\}, \quad (2.7)$$

where $\mathbb{G}_{n,j}^b(\cdot) = n^{-1/2} \sum_{i=1}^n (m_j(X_i^b, \cdot) - \bar{m}_{n,j}(\cdot))/\hat{\sigma}_{n,j}(\cdot)$ is a normalized bootstrap empirical process indexed by $\theta \in \Theta$,[8] $\hat{D}_{n,j}(\cdot)$ is a consistent estimator of $D_{P,j}(\cdot)$, $\rho > 0$ is a constant chosen by the researcher (see Section 4 for suggestions on how to choose it), $B^d = \{x \in \mathbb{R}^d : |x_j| \leq 1, \forall j\}$ is a unit box in $\mathbb{R}^d$, and $\hat{\xi}_{n,j}$ is defined by

$$\hat{\xi}_{n,j}(\theta) \equiv \begin{cases} \kappa_n^{-1} \sqrt{n} \bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta) & j = 1, \ldots, J_1 \\ 0 & j = J_1 + 1, \ldots, J, \end{cases} \quad (2.8)$$

where κ_n is a user-specified thresholding sequence such that $\kappa_n \to \infty$, and $\varphi : \mathbb{R}^J_{[\pm\infty]} \to \mathbb{R}^J_{[\pm\infty]}$ is one of the generalized moment selection (GMS) functions proposed by AS, and where $\mathbb{R}_{[\pm\infty]} = \mathbb{R} \cup \{\pm\infty\}$. A common choice is given componentwise by

$$\varphi_j(x) = \begin{cases} 0 & \text{if } \ x \geq -1 \\ -\infty & \text{if } \ x < -1. \end{cases} \quad (2.9)$$

Restrictions on φ and the rate at which κ_n diverges are imposed in Assumption 3.2.

Remark 2.1: For concreteness, in (2.9) we write out the "hard thresholding" GMS function; we also remark that this function simplifies computation as it completely removes non-local-to-binding constraints. Under Assumption 3.3 below, our results apply to all but one

[8] Bugni, Canay, and Shi (2014) propose a different approximation to the stochastic process $\mathbb{G}_{P,j}$, namely $n^{-1/2} \sum_{i=1}^n [(m_j(X_i, \cdot) - \bar{m}_{n,j}(\cdot))/\hat{\sigma}_{n,j}(\cdot)]\chi_i$ with $\{\chi_i \sim N(0,1)\}_{i=1}^n$ i.i.d. This approximation is equally valid in our approach, and can be computationally faster as it avoids repeated evaluation of $m_j(X_i^b, \cdot)$ across bootstrap replications.

of the GMS functions in AS, see Lemma B.3.[9] Under Assumption 3.3′, our method requires the use of hard thresholding GMS.

Heuristically, the random set $\Lambda_n^b(\theta, \rho, c)$ in (2.7) is a local (to θ), linearized bootstrap approximation to the random constraint set in (2.5). To see this, note first that the bootstrapped empirical process and the estimator of the gradient approximate the first two terms in the constraint in (2.5). Next, for $\theta \in \Theta_I$, the GMS function conservatively approximates the local slackness parameter $\sqrt{n}\gamma_{1,P,j}(\theta)$. This is needed because $\sqrt{n}\gamma_{1,P,j}(\theta)$ cannot be consistently estimated due to its scaling. GMS resolves this by shrinking estimated intercepts toward zero, thereby tightening constraints and hence increasing $\hat{c}_n(\theta)$. As with other uses of GMS, the resulting conservative distortion vanishes pointwise but not uniformly. Finally, restricting λ to the "ρ-box" ρB^d has a strong regularizing effect: It ensures uniform validity in challenging situations, including several that are assumed away in most of the literature. We discuss this point in detail in Section 4.

The critical level $\hat{c}_n(\theta)$ to be used in (2.2) is the smallest value of c that makes the bootstrap probability of the event

$$\min_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \leq 0 \leq \max_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \tag{2.10}$$

at least $1-\alpha$. Furthermore, Lemma C.1 in the Appendix establishes that

$$\min_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \leq 0 \leq \max_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \iff \Lambda_n^b(\theta,\rho,c) \cap \{p'\lambda = 0\} \neq \emptyset.$$

The intuition for this is simple: $\Lambda_n^b(\theta, \rho, c)$ is a polyhedron, therefore it contains some λ with $p'\lambda \geq 0$ but also some λ with $p'\lambda \leq 0$ if and only if it contains some λ with $p'\lambda = 0$. Our bootstrap critical level is, therefore, defined as

$$\hat{c}_n(\theta) \equiv \inf\{c \in \mathbb{R}_+ : P^*(\Lambda_n^b(\theta,\rho,c) \cap \{p'\lambda = 0\} \neq \emptyset) \geq 1-\alpha\}, \tag{2.11}$$

where P^* denotes the probability distribution induced by the bootstrap sampling process.

For a given $\theta \in \Theta$, coverage increases in c, and so $\hat{c}_n(\theta)$ can be quickly computed through a bisection algorithm. To do so, let $\bar{c}_n(\theta)$ be an upper bound on $\hat{c}_n(\theta)$. For example, the asymptotic Bonferroni bound $\bar{c}_n(\theta) = \Phi^{-1}(1-\alpha/J)$ is trivial to compute and would be too small only in very contrived cases which the algorithm would furthermore detect. Alternatively, in view of Theorem 3.2 below, the critical value proposed by AS is a valid upper bound in finite sample and typically much smaller, though harder to compute. By construction, $\hat{c}_n(\theta) \geq 0$. Hence, one can quickly find $\hat{c}_n(\theta)$ by initially guessing $\bar{c}_n(\theta)/2$, checking coverage, and then

[9] These are $\varphi^1 - \varphi^4$ in AS, all of which depend on $\kappa_n^{-1}\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)$. We do not consider GMS function φ^5 in AS, which depends also on the covariance matrix of the moment functions.

moving up or down by $\bar{c}_n(\theta)/2^{\mathtt{t}+1}$ in the $\mathtt{t}$'th step of the algorithm. More formally, define

$$\psi_b(c) \equiv \mathbf{1}(\Lambda_n^b(\theta, \rho, c) \cap \{p'\lambda = 0\} \neq \emptyset), \tag{2.12}$$

so that the bootstrap probability to be calibrated is $P^*(\psi_b(c) = 1)$. We propose the following algorithm:

Step 0

Set `Tol` equal to a chosen tolerance value or fix the number of iterations `T`.
Initialize $C(0) = 0$.
Initialize $\mathtt{t} = 1$.
Initialize $c = \bar{c}_n(\theta)/2$.
Initialize $\varphi_j(\hat{\xi}_{n,j}(\theta)) = 0,\ j = 1, \ldots, J$.
Compute $\varphi_j(\hat{\xi}_{n,j}(\theta)),\ j = 1, \ldots, J_1$.
Compute $\hat{D}_{P,n}(\theta)$.
Compute $\mathbb{G}_{n,j}^b(\theta)$ for $b = 1, ..., B$.
Compute $\psi_b(c)$ for $b = 1, ..., B$.

Step 1

Compute $C(\mathtt{t}) = n^{-1} \sum_{b=1}^B \psi_b(c)$.

Step 2

If $C(\mathtt{t}) > 1 - \alpha$, set $c \leftarrow c - \frac{\bar{c}_n(\theta)}{2^{\mathtt{t}+1}}$ and recompute $\psi_b(c)$ for each b such that $\psi_b(c) = 1$.
If $C(\mathtt{t}) < 1 - \alpha$, set $c \leftarrow c + \frac{\bar{c}_n(\theta)}{2^{\mathtt{t}+1}}$ and recompute $\psi_b(c)$ for each b such that $\psi_b(c) = 0$.

Step 3

If $|C(\mathtt{t}) - C(\mathtt{t}-1)| >$ `Tol`, set $\mathtt{t} = \mathtt{t} + 1$ and return to Step 1.
If $|C(\mathtt{t}) - C(\mathtt{t}-1)| <$ `Tol` or $\mathtt{t} =$ `T`, set $\hat{c}_n(\theta) = c$ and exit.

Execution of this is further simplified by the following observation: W.l.o.g. let $p = (1, 0, \ldots, 0)'$, implying that $p'\lambda = 0$ if and only if $\lambda_1 = 0$. Evaluation of $\psi_b(c)$ thus entails determining whether a constraint set comprised of $J+2d-1$ linear inequalities in $d-1$ variables is feasible. This can be accomplished efficiently employing commonly used software.[10] Also, note that the B bootstrap draws remain fixed across iterations, and we know that for any given bootstrap sample, coverage will obtain if and only if c is above some threshold. Hence, one needs to recompute $\psi_b(c)$ in Step 2 only for a subset of bootstrap draws that decreases in $\mathtt{t}$. Our algorithm reflects this insight.

[10] Examples of high-speed solvers for linear programs include CVXGEN, available from http://cvxgen.com, and Gurobi, available from http://www.gurobi.com

2.3 Computation of Outer Maximization Problem

The constrained optimization problem in (2.2) has nonlinear constraints involving a component $\hat{c}_n(\theta)$ which in general is an unknown function of θ. Moreover, in all methods, including ours and AS, the gradients of constraints are not available in closed form. When the dimension of the parameter is large, directly solving optimization problems with such a component can be relatively expensive even if evaluating $\hat{c}_n(\theta)$ at each θ is computationally cheap. This is because commonly used optimization algorithms repeatedly evaluate the constraints and their (numerical) gradients.

To mitigate the computational cost, we suggest an algorithm that is a contribution to the moment (in)equalities literature in its own right and should also be helpful for implementing other approaches. The algorithm consists of three steps called E, A, and M below, and is based on the response surface method used in the optimization literature (see e.g. Jones, 2001; Jones, Schonlau, and Welch, 1998, and references therein). In what follows, we assume that computing the sample moments is less expensive than computing $\hat{c}_n(\theta)$.

E-step: (Evaluation) Evaluate $\hat{c}_n(\theta^{(\ell)})$ for $\ell = 1, \cdots, L$. Set $\Upsilon^{(\ell)} = \hat{c}_n(\theta^{(\ell)}), \ell = 1, \cdots, L$. We suggest setting $L = 20d+1$, so L grows linearly with the dimensionality of parameter space.

A-step: (Approximation) Approximate $\theta \mapsto \hat{c}_n(\theta)$ by a flexible auxiliary model. For example, a Gaussian-process regression model (or kriging) is

$$\Upsilon^{(\ell)} = \mu + \epsilon(\theta^{(\ell)}), \ \ell = 1, \cdots, L, \tag{2.13}$$

where $\epsilon(\cdot)$ is a mean-zero Gaussian process indexed by θ with a constant variance σ^2 whose correlation functional is $Corr(\epsilon(\theta), \epsilon(\theta')) = \exp(-\delta(\theta, \theta'))$ for some distance measure δ, e.g. $\delta(\theta, \theta') = \sum_{k=1}^{d} \beta_k |\theta_k - \theta'_k|^{\gamma_k}, \beta_k \geq 0, \gamma_k \in [1, 2]$. The unknown parameters (μ, σ^2) can be estimated by running a GLS regression of $\boldsymbol{\Upsilon} = (\Upsilon^{(1)}, \cdots, \Upsilon^{(L)})'$ on a constant with the given correlation matrix. The unknown parameters in the correlation matrix can be estimated by a (concentrated) MLE. The (best linear) predictor of the critical value and its gradient at an arbitrary point are then given by

$$\hat{c}_n^A(\theta) = \hat{\mu} + \mathbf{r}(\theta)' \mathbf{R}^{-1} (\boldsymbol{\Upsilon} - \hat{\mu} \mathbf{1}), \tag{2.14}$$

$$\nabla_\theta \hat{c}_n^A(\theta) = \hat{\mu} + \mathbf{Q}(\theta) \mathbf{R}^{-1} (\boldsymbol{\Upsilon} - \hat{\mu} \mathbf{1}), \tag{2.15}$$

where $\mathbf{r}(\theta)$ is a vector whose ℓ-th component is $Corr(\epsilon(\theta), \epsilon(\theta^{(\ell)}))$ as given above with estimated parameters, $\mathbf{Q}(\theta) = \nabla_\theta \mathbf{r}(\theta)'$, and $\mathbf{R}$ is an L-by-L matrix whose (ℓ, ℓ') entry is $Corr(\epsilon(\theta^{(\ell)}), \epsilon(\theta^{(\ell')}))$ with estimated parameters. This approximation (or surrogate) model has the property that its predictor satisfies $\hat{c}_n^A(\theta^{(\ell)}) = \hat{c}_n(\theta^{(\ell)}), \ell = 1, \cdots, L$. Hence, it provides an analytical interpolation to the evaluated critical values together

with an analytical gradient.[11]

M-step: (Maximization or Minimization): Solve the optimization problem

$$\begin{aligned} &\max / \min_{\theta \in \Theta} \quad p'\theta \\ &s.t. \quad \sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta) \leq \hat{c}_n^A(\theta), \end{aligned} \tag{2.16}$$

while using p and $\sqrt{n}\hat{D}_{n,j}(\theta) - \nabla_\theta \hat{c}_n^A(\theta), j = 1, \cdots, J$ as the gradients of the objective function and constraint functions respectively. This step can be implemented by standard nonlinear optimization solvers (e.g. Matlab's `fmincon` or `KNITRO`).

Once the optimal value from the M-step is obtained, draw L_1 additional points in a subset of the parameter space that contains parameter values that nearly attain the maximum. Add them to the previously used evaluation points and update the total number of evaluation points as $L + L_1$. Iterate the E-A-M-steps until the maximized value converges.[12] Report the maximum and minimum values of the optimization problem as the endpoints of the confidence interval.

REMARK 2.2: The advantages of the proposed algorithm are twofold. First, we control the number of points at which we evaluate $\hat{c}_n(\cdot)$. Since the evaluation of the critical value is the relatively expensive step, controlling the number of evaluations is important. One should also note that this step can easily be parallelized. Second, the proposed algorithm makes the maximization step computationally cheap by providing constraints and their gradients in closed form. It is well known that gradient-based algorithms solve optimization problems more efficiently than those that do not use gradients. The price to pay is the additional approximation step. According to our numerical exercises, this additional step is not costly.

3 Asymptotic Validity of Inference

In this section, we justify our procedure by establishing uniform (over an interesting class of DGPs) asymptotic validity. Subsection 3.1 states and motivates our assumptions; subsection 3.2 states and discusses our main results.

3.1 Assumptions

Our first assumption is on the parameter space and the criterion function. Below, ϵ and M are used to denote generic constants which may be different in different appearances.

ASSUMPTION 3.1: *$\Theta \subseteq \mathbb{R}^d$ is compact and convex with a nonempty interior.*

[11] See details in Jones, Schonlau, and Welch (1998). We use the DACE Matlab kriging toolbox (http://www2.imm.dtu.dk/projects/dace/) for this step in the Monte Carlo experiments based on the entry game.

[12] One can make the subset to which one adds evaluation points smaller as one iterates.

Compactness is a standard assumption on Θ for extremum estimation. In addition, we require convexity as we use mean value expansions of $E_P[m_j(X_i,\theta)]$ in θ as shown in equation (2.6).

The next assumption defines our moment (in)equalities model. It is based on AS, and most of it is standard in the literature.[13]

ASSUMPTION 3.2: *The function φ_j is continuous at all $x \geq 0$ and $\varphi_j(0) = 0$; $\kappa_n \to \infty$ and $\kappa_n^{-1} n^{1/2} \to \infty$. The model $\mathcal{P}$ for P satisfies the following conditions:*

(i) $E_P[m_j(X_i,\theta)] \leq 0,\ j = 1,\ldots,J_1$ *and* $E_P[m_j(X_i,\theta)] = 0,\ j = J_1+1,\cdots,J_1+J_2$ *for some* $\theta \in \Theta$*;*

(ii) $\{X_i, i \geq 1\}$ *are i.i.d. under* P*;*

(iii) $\sigma^2_{P,j}(\theta) \in (0,\infty)$ *for* $j = 1,\cdots,J$ *for all* $\theta \in \Theta$*;*

(iv) For some $\delta > 0$ *and* $M \in (0,\infty)$ *and for all* j, $E_P[\sup_{\theta\in\Theta} |m_j(X_i,\theta)/\sigma_{P,j}(\theta)|^{2+\delta}] \leq M$.

In what follows, for any sequence of random variables $\{X_n\}$ and a positive sequence a_n, we write $X_n = o_{\mathcal{P}}(a_n)$ if for any $\epsilon,\eta > 0$, there is $N \in \mathbb{N}$ such that $\sup_{P\in\mathcal{P}} P(|X_n/a_n| > \epsilon) < \eta, \forall n \geq N$. We write $X_n = O_{\mathcal{P}}(a_n)$ if for any $\eta > 0$, there is a $M \in \mathbb{R}_+$ and $N \in \mathbb{N}$ such that $\sup_{P\in\mathcal{P}} P(|X_n/a_n| > M) < \eta, \forall n \geq N$. Given a square matrix A, we write $\text{eig}(A)$ for its smallest eigenvalue.

Next, and unlike some other papers in the literature, we restrict the correlation matrix of the moment conditions. Because our method is based on replacing a nonlinear program with a linear one, it is intuitive that a Karush-Kuhn-Tucker condition (with uniformly bounded Lagrange multipliers) is needed. Imposing this condition directly, however, would yield an assumption that can be very hard to verify in a given application – as constraint qualification conditions often are.[14] On the other hand, we are able to show that restrictions on the correlation matrix of the moments, together with imposition of the ρ-box constraints, yield such constraint qualification conditions on the set $\Lambda_n^b(\theta,\rho,c)$ defined in (2.7) with arbitrarily high probability for n large enough. We provide additional details in Section 4.3, see in particular footnote 27 for an illustration. Here we begin with an easy sufficient condition, and then discuss an alternative condition that holds for some cases in which the first one does not. For a reader interested in alternative assumptions, we note that Assumption 3.3

[13] The requirement that φ_j is continuous for $x \geq 0$ is restrictive only for GMS function $\varphi^{(2)}$ in AS. We also remark that one specific result, namely Lemma C.2 below, requires $\varphi_j(x) \leq 0$ for all x. To keep the treatment general, we do not impose this restriction throughout, but we only recommend functions φ_j with this feature anyway. It is easy to see that for any φ_j that can take strictly positive values, substituting $\min\{\varphi_j(x), 0\}$ attains the same asymptotic size but generates CIs that are weakly shorter for all and strictly shorter for some sample realizations.

[14] Restrictions of this type are imposed both in PPHI and Chernozhukov, Hong, and Tamer (2007), as we explain in Section 4

(or Assumption 3.3′ below) is used exclusively to obtain the conclusions of Lemma B.6 and Lemma B.7, hence any alternative assumption that delivers such results can be used.

ASSUMPTION 3.3: *The model $\mathcal{P}$ for P satisfies the following additional conditions:*

(i) There is a positive constant ϵ such that $\Theta_I(P) \subset \Theta^{-\epsilon} \equiv \{\theta \in \Theta : d(\theta, \mathbb{R}^d \setminus \Theta) \geq \epsilon\}$, where d denotes Euclidean point-set distance.

(ii) For all $\theta \in \Theta$, $\eta_{n,j}(\theta) \equiv \sigma_{P,j}(\theta)/\hat{\sigma}_{n,j}(\theta) - 1 = o_{\mathcal{P}}(\kappa_n/\sqrt{n})$.

(iii) Let $\tilde{m}(X_i, \theta) \equiv (m_1(X_i, \theta), \cdots, m_{J_1+J_2}(X_i, \theta))'$. Let $\tilde{\Omega}_P(\theta) = Corr_P(\tilde{m}(X_i, \theta))$. Then $\inf_{\theta \in \Theta_I(P)} \text{eig}(\tilde{\Omega}_P(\theta)) \geq \omega$ for some constant $\omega > 0$.

Assumption 3.3 (i) requires that the identified set is in an ϵ-contraction of the parameter space. This implies that the behavior of the support function of $\mathcal{C}_n(\hat{c}_n)$ is determined only by the moment restrictions asymptotically under any $P \in \mathcal{P}$. This assumption could be dropped if the parameter space can be defined through a finite list of smooth nonstochastic inequality constraints, e.g. if $\Theta = [0, 1]^d$.

Assumption 3.3 (ii) is a weak regularity condition requiring that each moment's standard deviation can be estimated at a rate faster than $\kappa_n/\sqrt{n}$.

The crucial part of Assumption 3.3 is (iii), which requires that the correlation matrix of the sample moments has eigenvalues uniformly bounded from below. While it holds in many applications of interest, we are aware of two examples in which it may fail. One are missing data scenarios when the unconditional or some conditional proportion of missing data vanishes. This is easiest to see for the scalar mean with missing data, where sample analogs of upper and lower bound approach perfect correlation as the population probability of missing data vanishes. The observation also applies to higher dimensional examples, e.g. best linear prediction with missing outcome data. The other example is the Ciliberto and Tamer (2009) entry game model when the solution concept is pure strategy Nash equilibrium, as illustrated in the following example.

EXAMPLE 3.1 (Two player entry game): Consider the simple case of a static two player entry game of complete information with pure strategy Nash equilibrium as solution concept. Suppose each player $k = 1, 2$ in market $i = 1, \ldots, n$ can choose to enter ($X_{ik} = 1$) or to stay out of the market ($X_{ik} = 0$). Let $\varepsilon_{i1}, \varepsilon_{i2}$ be two random variables representing unobservable payoff shifters, and for simplicity assume they are distributed i.i.d. $U(0, 1)$. Let players' payoffs be

$$u_{ik} = X_{ik}(-\theta_k X_{i,3-k} + \varepsilon_{ik}), \quad k = 1, 2,$$

with $\theta \in \Theta = [0, 1]^2$ the parameter vector of interest. Each player enters the game if and only if $u_{ik} \geq 0$. This game admits multiple equilibria, and one can show that $\Theta_I(P)$ is defined by

the following (in)equalities:

$$E_P(m_1(X_i,\theta)) = E_P[X_{i1}(1-X_{i2}) - \theta_2] \leq 0,$$
$$E_P(m_2(X_i,\theta)) = E_P[\theta_2(1-\theta_1) - X_{i1}(1-X_{i2})] \leq 0,$$
$$E_P(m_3(X_i,\theta)) = E_P[X_{i1}X_{i2} - (1-\theta_1)(1-\theta_2)] = 0.$$

Then moment functions 1 and 2 violate Assumption 3.3 (iii) because they are perfectly correlated.[15]

These examples and more complex ones are covered by our next assumption. If it is invoked, our procedure requires the use of the specific GMS function in equation (2.9).

Assumption 3.3′. *The function φ used to obtain $\hat{c}_n(\theta)$ in (2.11) is given in equation (2.9); $\kappa_n = o(n^{1/4})$. The model $\mathcal{P}$ for P satisfies Assumption 3.3(i)-(ii), and in addition:*

(iii-1) *The first $2J_{11}$ moment functions, $0 \leq 2J_{11} \leq J_1$, are related as follows:*
$m_{j+J_{11}}(X_i,\theta) = -m_j(X_i,\theta) - t_j(X_i,\theta),\ j = 1,\ldots,J_{11}$
where for each $\theta \in \Theta$ and $j = 1,\ldots,J_{11}$, $t_j : \mathcal{X} \times \Theta \to \mathbb{R}_+$ is a measurable function such that $0 \leq t_j(X,\theta) \leq M$ a.s., $j = 1,\ldots,J_{11}$.

(iii-2) *Let $\check{m}(X_i,\theta)$ be a J_{11}-vector that selects exactly one of each pair of moment functions $\{m_j(X_i,\theta), m_{j+J_{11}}(X_i,\theta)\}$, $j = 1,...,J_{11}$. Let $\tilde{m}(X_i,\theta) \equiv (\check{m}(X_i,\theta), m_{2J_{11}+1}(X_i,\theta), \cdots, m_{J_1+J_2}(X_i,\theta))'$. Denote $\tilde{\Omega}_P(\theta) = Corr_P(\tilde{m}(X_i,\theta))$. Then $\inf_{\theta\in\Theta_I(P)} \text{eig}(\tilde{\Omega}_P(\theta)) \geq \omega$ for some constant $\omega > 0$, uniformly over all $2^{J_{11}}$ possible vectors $\check{m}(X_i,\theta)$.*

(iii-3) $\inf_{\theta\in\Theta_I(P)} \sigma_{P,j}(\theta) > \underline{\sigma}$ *for* $j = 1,\ldots,J_{11}$.

(iii-4) *For $\theta \in \Theta$,* $\lim_{n\to\infty} P\left(\frac{\bar{m}_{j+J_{11},n}(\theta)}{\hat{\sigma}_{n,j+J_{11}}(\theta)} \leq -\frac{\bar{m}_{n,j}(\theta)}{\hat{\sigma}_{n,j}(\theta)}\right) \to 1.$

In words, Assumption 3.3′ allows for (drifting to) perfect correlation among moment inequalities that cannot cross. Again, the scalar mean with missing data is perhaps the easiest example. In the generalization of this example in Imbens and Manski (2004) and Stoye (2009), parts (iii-1)-(iii-2) of Assumption 3.3′ are satisfied by construction, part (iii-3) is directly assumed, and part (iii-4) can be verified to hold.

Regarding Ciliberto and Tamer (2009), inspection of Example 3.1 reveals that part (iii-1) of the assumption is satisfied with $t_j(\cdot,\theta) = t_j(\theta)$ for each $j = 1,\ldots,J$; in more general instances of the model, this follows because any pair of moment conditions that involve the same outcome of the game differ by model predicted probabilities of regions of multiplicity. Part (iii-2) of the assumption holds in the example provided that $|Corr(X_{i1}(1-X_{i2}), X_{i1}X_{i2})| < 1-\epsilon$ for some $\epsilon > 0$; in more general instances, it follows if the multinomial distribution of outcomes

[15] One can show, however, that under a different solution concept, e.g. rationality of level 1, the resulting moment inequalities would satisfy Assumption 3.3 (iii).

of the game (reduced by one element) has a correlation matrix with eigenvalues uniformly bounded away from zero.[16] To see that part (iii-3) of the assumption also holds, note that Assumption 3.2 (iv) yields that $P(X_{i1} = 1, X_{i2} = 0)$ is uniformly bounded away from 0 and 1, thereby implying that $\sigma_1 \geq \underline{\sigma} > 0$ and similarly for σ_2; the same holds for $P(X_{i1} = 1, X_{i2} = 1)$ and so $\sigma_3 \geq \underline{\sigma} > 0$. An analogous reasoning holds in more general instances of the Ciliberto and Tamer model. Finally, part (iii-4) of the assumption requires that the studentized sample moments are ordered with probability approaching one. This condition is immediately implied by condition (iii-1) in any model in which for each $j = 1, \ldots, J_1 + J_2$ the function $m_j(X_i, \theta)$ can be written as the sum of a function that depends on X_i only, and a function that depends on θ only. General instances of the Ciliberto and Tamer (2009) model (and of course Example 3.1) belong to this class of models.

In what follows, we refer to a pair of inequality constraints indexed by $\{j, j + J_{11}\}$ as described in Assumption 3.3′ as "paired inequalities". The presence of paired inequalities requires that we modify our bootstrap procedure. All modifications are carried out within Step 0 of the Algorithm in Section 2.2. If

$$\varphi_j(\hat{\xi}_{n,j}(\theta)) = 0 = \varphi_j(\hat{\xi}_{n,j+J_{11}}(\theta)),$$

with φ_j as defined in (2.9), we replace $\mathbb{G}^b_{P,j+J_{11},n}(\theta)$ with $-\mathbb{G}^b_{n,j}(\theta)$, and $\hat{D}_{P,j+J_{11},n}(\theta)$ with $-\hat{D}_{n,j}(\theta)$, so that inequality

$$\mathbb{G}^b_{P,j+J_{11},n}(\theta) + \hat{D}_{P,j+J_{11},n}(\theta)\lambda \leq c$$

is replaced with

$$-\mathbb{G}^b_{n,j}(\theta) - \hat{D}_{n,j}(\theta)\lambda \leq c$$

in equation (2.7). In words, when hard threshold GMS indicates that both paired inequalities bind, we pick one of them, treat it as an equality, and drop the other one. This tightens the stochastic program because by Assumption 3.3′, each inequality if interpreted as equality implies the other one. The rest of the procedure is unchanged.

Finally, we informally remark that if hard thresholding, i.e. expression (2.9), is used for GMS, then two inequalities that are far from each other in the sense that GMS only picks at most one of them at any given θ may be arbitrarily correlated. This condition could be used to further weaken Assumption 3.3 or 3.3′ and is easy to pre-test for; we omit an elaboration.

We next lay out regularity conditions on the gradients of the moments.

ASSUMPTION 3.4: *The model $\mathcal{P}$ for P satisfies the following additional conditions:*

[16] In the Ciliberto and Tamer (2009) framework there is a single vector of moment functions $\tilde{m}(X_i, \theta)$ to consider instead of $2^{J_{11}}$. If the game admits K possible outcome, the vector $\tilde{m}(X_i, \theta)$ includes one moment function for each of $K - 1$ possible outcomes of the game.

(i) For each j, there exist $D_{P,j}(\cdot) \equiv \nabla_\theta\{E_P[m_j(X,\cdot)]/\sigma_{P,j}(\cdot)\}$ and its estimator $\hat{D}_{n,j}(\cdot)$ such that $\sup_{\theta\in\Theta}\|\hat{D}_{n,j}(\theta) - D_{P,j}(\theta)\| = o_{\mathcal{P}}(1)$. Further, there exists $\bar{M} > 0$ such that $\|D_{P,j}(\theta)\| \leq \bar{M}$ for all $\theta \in \Theta_I$ and $j = 1, \cdots, J$;

(ii) There exists $M > 0$ such that $\max_{j=1,\cdots,J}\sup_{\theta,\tilde{\theta}\in\Theta}\|D_{P,j}(\theta) - D_{P,j}(\tilde{\theta})\| \leq M\|\theta - \tilde{\theta}\|$.

Assumption 3.4 requires that the normalized population moment is differentiable, that its derivative is Lipschitz continuous, and that this derivative can be consistently estimated uniformly in θ and P. We require these conditions because we use a linear expansion of the population moments to obtain a first-order approximation to the support function of $\mathcal{C}_n$ and our bootstrap procedure requires an estimator of the population gradient. We do not assume that a criterion function that aggregates moment violations (e.g., $T_n(\theta)$ in equation (3.7) below) is bounded from below by a polynomial function of θ outside a neighborhood of the identification region. This is assumed in related work (see e.g. Chernozhukov, Hong, and Tamer, 2007) but fails in relevant examples, e.g. when two moment inequalities form an extremely acute corner of the identified set. We return to such examples in Section 4.

A final set of assumptions is on the normalized empirical process. For this, define the variance semimetric ϱ_P by

$$\varrho_P(\theta,\tilde{\theta}) \equiv \left\|\left\{Var_P\left(\sigma_{P,j}^{-1}(\theta)m_j(X,\theta) - \sigma_{P,j}^{-1}(\tilde{\theta})m_j(X,\tilde{\theta})\right)^{1/2}\right\}_{j=1}^{J}\right\|. \tag{3.1}$$

For each $\theta,\tilde{\theta} \in \Theta$ and P, let $Q_P(\theta,\tilde{\theta})$ denote a J-by-J matrix whose (j,k)-th element is the covariance between $m_j(X_i,\theta)/\sigma_{P,j}(\theta)$ and $m_k(X_i,\tilde{\theta}))/\sigma_{P,k}(\tilde{\theta})$ under P.

Assumption 3.5: *(i) For every $P \in \mathcal{P}$, and $j = 1, \cdots, J$, $\{\sigma_{P,j}^{-1}(\theta)m_j(\cdot,\theta) : \mathcal{X} \to \mathbb{R}, \theta \in \Theta\}$ is a measurable class of functions; (ii) The empirical process $\mathbb{G}_n$ with j-th component $\mathbb{G}_{n,j}$ is asymptotically ϱ_P-equicontinuous uniformly in $P \in \mathcal{P}$. That is, for any $\epsilon > 0$,*

$$\lim_{\delta\downarrow 0}\limsup_{n\to\infty}\sup_{P\in\mathcal{P}} P^*\left(\sup_{\varrho_P(\theta,\tilde{\theta})<\delta}\|\mathbb{G}_n(\theta) - \mathbb{G}_n(\tilde{\theta})\| > \epsilon\right) = 0; \tag{3.2}$$

(iii) Q_P satisfies

$$\lim_{\delta\downarrow 0}\sup_{\|(\theta_1,\tilde{\theta}_1)-(\theta_2,\tilde{\theta}_2)\|<\delta}\sup_{P\in\mathcal{P}}\|Q_P(\theta_1,\tilde{\theta}_1) - Q_P(\theta_2,\tilde{\theta}_2)\| = 0. \tag{3.3}$$

Under this assumption, the class of normalized moment functions is uniformly Donsker (Bugni, Canay, and Shi, 2015). This allows us to show that the first-order linear approximation to $s(p,\mathcal{C}_n(\hat{c}_n))$ is valid and further establish the validity of our bootstrap procedure.

3.2 Theoretical Results

Result 1: Uniform asymptotic validity.

The following theorem establishes the asymptotic validity of the proposed confidence interval $CI_n \equiv [-s(-p, \mathcal{C}_n(\hat{c}_n)), s(p, \mathcal{C}_n(\hat{c}_n))]$, where $\hat{c}_n$ was defined in equation (2.11).

THEOREM 3.1: *Suppose Assumptions 3.1, 3.2, 3.3 or 3.3′, 3.4, and 3.5 hold. Let* $0 < \alpha < 1/2$. *Then,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(p'\theta \in CI_n) \geq 1-\alpha. \tag{3.4}$$

Some brief remarks on proof strategy are as follows. Using equations (2.5) and (2.6) and recalling that adding constraints can only make the coverage probability lower, we show that asymptotic size control is ensured if we choose the function c to (asymptotically and uniformly over $\mathcal{P}$ and $\Theta_I(P)$) guarantee that

$$P\left(\Lambda_n^{NL}(\theta, \rho, c(\theta)) \cap \{p'\lambda = 0\} \neq \emptyset\right) \geq 1-\alpha, \tag{3.5}$$

where

$$\Lambda_n^{NL}(\theta, \rho, c(\theta)) = \left\{\lambda \in \rho B^d : \left(\mathbb{G}_{n,j}(\theta + \lambda/\sqrt{n}) + D_{P,j}(\bar{\theta}_j)\lambda + \sqrt{n}\gamma_{1,P,j}(\theta)\right)(1 + \eta_{n,j}(\theta)) \leq c(\theta)\right\}$$

and $\bar{\theta}_j$ lies component-wise between θ and $\theta + \lambda/\sqrt{n}$. Our bootstrap procedure is based on the feasible polyhedral set

$$\Lambda_n^b(\theta, \rho, \hat{c}_n(\theta)) = \left\{\lambda \in \rho B^d : \mathbb{G}_{n,j}^b(\theta) + \hat{D}_{n,j}(\theta)\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta)) \leq \hat{c}_n(\theta)\right\},$$

yielding as a bootstrap analog of equation (3.5),

$$P\left(\Lambda_n^b(\theta, \rho, \hat{c}_n(\theta)) \cap \{p'\lambda = 0\} \neq \emptyset\right) \geq 1-\alpha. \tag{3.6}$$

We do *not* establish that our bootstrap based critical level $\hat{c}_n(\theta)$ consistently estimates an oracle level $c(\theta)$. Indeed, we allow that $\hat{c}_n(\theta)$ might not (uniformly) converge anywhere. This is why, unlike the related literature, we avoid assumptions on limit distributions of test statistics. What we do show is that, for n large enough, the probability in (3.5) weakly exceeds (up to $o_{\mathcal{P}}(1)$) the one in (3.6) uniformly over arguments c and therefore, in particular, for $\hat{c}_n(\theta)$. It is also worth noting that the proof contains a novel (to the best of our knowledge) use of a fixed point theorem in the moment (in)equalities literature. This occurs in our argument that, if the linear program Λ_n^b is feasible, then the nonlinear program Λ_n^{NL} is very likely to be feasible as well. In the presence of equality constraints, showing this requires to show that a certain nonlinear system of equations can be solved, which is where the fixed

point theorem comes in.

REMARK 3.1: By replacing the constraint $p'\lambda = 0$ with $p'\lambda \geq 0$ in calibrating $\hat{c}_n$:

$$\hat{c}_n(\theta) = \inf\{c \in \mathbb{R}_+ : P^*(\Lambda_n^b(\theta, \rho, c) \cap \{p'\lambda \geq 0\} \neq \emptyset) \geq 1 - \alpha\},$$

one obtains a critical level that yields a valid one-sided confidence interval $(-\infty, s(p, \mathcal{C}_n(\hat{c}_n))]$ (or $[-s(-p, \mathcal{C}_n(\hat{c}_n)), \infty)$ if one uses $p'\lambda \leq 0$ in the calibration of $\hat{c}_n$). This differentiates our method from profiling methods and also from projection of AS, where the analogous adaptation is not obvious.

Result 2: Improvement over projection of AS.

Our second set of results establish that for each $n \in \mathbb{N}$, CI_n is a subset of a confidence interval obtained by projecting an AS confidence set.[17] Moreover, we derive simple conditions under which our confidence interval is a proper subset of the projection of AS's confidence set. Below we let $\hat{c}_n^{AS}$ denote the critical value obtained applying AS with criterion function

$$T_n(\theta) = \max\left\{\max_{j=1,\cdots,J_1} \sqrt{n}[\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)]_+, \max_{j=J_1+1,\cdots,J_1+J_2} \sqrt{n}|\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)|\right\}, \quad (3.7)$$

and with the same choice as for $\hat{c}_n$ of GMS function φ and tuning parameter κ_n. We also note that for given function c, one can express $\mathcal{C}_n(c)$ in (1.2) as

$$\mathcal{C}_n(c) = \{\theta \in \Theta : T_n(\theta) \leq c(\theta)\}.$$

THEOREM 3.2: *Suppose Assumptions 3.1, 3.2, 3.3 or 3.3′, 3.4, and 3.5 hold. Let* $0 < \alpha < 1/2$. *Then for each* $n \in \mathbb{N}$

$$CI_n \subseteq [-s(-p, \mathcal{C}_n(\hat{c}_n^{AS})), s(p, \mathcal{C}_n(\hat{c}_n^{AS}))]. \quad (3.8)$$

The result in Theorem 3.2 is due to the following fact. Recall that AS's confidence region calibrates its critical value so that, at each θ, the following event occurs with probability at least $1 - \alpha$:

$$\max_{j=1,\ldots,J}\left\{\mathbb{G}_{n,j}^b(\theta) + \varphi_j(\hat{\xi}_{n,j}(\theta))\right\} \leq c. \quad (3.9)$$

On the other hand, we determine $\hat{c}_n$ using the event (2.10). If c satisfies (3.9), it also satisfies (2.10) because in that case $\lambda = 0$ is in the feasibility set $\Lambda_n^b(\theta, \lambda, c)$ defined in (2.7).[18] Therefore, by construction, our critical level $\hat{c}_n$ is weakly dominated by $\hat{c}_n^{AS}$, and hence our CI_n is a subset of the projection of AS's confidence region that uses the same statistic and GMS function.

[17] Of course, AS designed their confidence set to uniformly cover each vector in Θ_I with prespecified asymptotic probability, a different inferential problem than the one considered here.

[18] Indeed, $\hat{c}_n^{AS}$ can be seen as the special case of $\hat{c}_n$ where ρ was set to 0.

A natural question is, then, whether there are conditions under which CI_n is strictly shorter than the projection of AS's confidence region. Heuristically, this is the case with probability approaching 1 when $\hat{c}_n(\theta)$ is strictly less than $\hat{c}_n^{AS}(\theta)$ at each θ that is relevant for projection. For this, restrict $\varphi(\cdot)$ to satisfy $\varphi_j(x) \leq 0$ for all x, fix θ and consider the pointwise limit of (3.9):

$$\mathbb{G}_{P,j}(\theta) + \zeta_{P,j}(\theta) \leq c, \ \ j = 1, \cdots, J, \tag{3.10}$$

where $\{\mathbb{G}_{P,j}(\theta), j = 1, \cdots, J\}$ follows a multivariate normal distribution, and $\zeta_{P,j}(\theta) \equiv (-\infty)\mathbf{1}(\sqrt{n}\gamma_{1,P,j}(\theta) < 0)$ is the pointwise limit of $\varphi_j(\hat{\xi}_{n,j}(\theta))$ (with the convention that $(-\infty)0 = 0$). Under mild regularity conditions, $\hat{c}_n^{AS}(\theta)$ then converges in probability to a critical value $c = c^{AS}(\theta)$ such that (3.10) holds with probability $1-\alpha$. Similarly, the limiting event that corresponds to our problem (2.10) is

$$\Lambda(\theta, \rho, c) \cap \{p'\lambda = 0\} \neq \emptyset, \tag{3.11}$$

where the limiting feasibility set $\Lambda(\theta, \rho, c)$ is given by

$$\Lambda(\theta, \rho, c) = \{\lambda \in \rho B^d : \mathbb{G}_{P,j}(\theta) + D_{P,j}(\theta)\lambda + \zeta_{P,j}(\theta) \leq c, j = 1, \cdots, J\}. \tag{3.12}$$

Note that if the gradient $D_{P,j}(\theta)$ is a scalar multiple of p, i.e. $D_{P,j}(\theta)/\|D_{P,j}(\theta)\| \in \{p, -p\}$, for all j such that $\zeta_{P,j}(\theta) = 0$, the two problems are equivalent because (3.10) implies (3.11) (again by arguing that $\lambda = 0$ is in $\Lambda(\theta, \rho, c)$), and for the converse implication, whenever (3.11) holds, there is λ such that $\mathbb{G}_{P,j}(\theta) + D_{P,j}(\theta)\lambda + \zeta_{P,j}(\theta) \leq c$ and $p'\lambda = 0$. Since $D_{P,j}(\theta)\lambda = 0$ for all j such that $\zeta_{P,j}(\theta) = 0$, one has $\mathbb{G}_{P,j}(\theta) + \zeta_{P,j}(\theta) \leq c$ for all j.[19] In this special case, the limits of the two critical values coincide asymptotically, but any other case is characterized by projection conservatism. Lemma C.2 in the Appendix formalizes this insight. Specifically, for fixed θ, the limit of $\hat{c}_n(\theta)$ is strictly less than the limit of $\hat{c}_n^{AS}(\theta)$ if and only if there is a constraint that binds or is violated at θ and has a gradient that is not a scalar multiple of p.[20]

The parameter values that are relevant for the lengths of the confidence intervals are the ones whose projections are in a neighborhood of the projection of the identified set. Therefore, a leading case in which our confidence interval is strictly shorter than the projection of AS asymptotically is that in which at any θ (in that neighborhood of the projection of the identified set) at least one local-to-binding or violated constraint has a gradient that is not parallel to p. We illustrate this case with an example based on Manski and Tamer (2002).

[19] The gradients of the non-binding moment inequalities do not matter here because $\mathbb{G}_{P,j}(\theta) + \zeta_{P,j}(\theta) \leq c$ holds due to $\zeta_{P,j}(\theta) = -\infty$ for such constraints.

[20] The condition that all binding moment inequalities have gradient collinear with p is not as exotic as one might think. An important case where it obtains is the "smooth maximum," i.e. the support set is a point of differentiability of the boundary of Θ_I.

EXAMPLE 3.2 (Linear regression with an interval valued outcome): Consider a linear regression model:

$$E[Y|Z] = Z'\theta, \tag{3.13}$$

where Y is an unobserved outcome variable, which takes values in the interval $[Y_L, Y_U]$ with probability one, and Y_L, Y_U are observed. The vector Z collects regressors taking values in a finite set $S_Z \equiv \{z_1, \cdots, z_K\}, K \in \mathbb{N}$. We then obtain the following conditional moment inequalities:

$$E_P[Y_L|Z = z_j] \leq z_j'\theta \leq E_P[Y_U|Z = z_j],\ j = 1, \cdots, K, \tag{3.14}$$

which can be converted into unconditional moment inequalities with $J_1 = 2K$ and

$$m_j(X, \theta) = \begin{cases} Y_L 1\{Z = z_j\}/g(z_j) - z_j'\theta, & j = 1, \cdots, K \\ z_{j-K}'\theta - Y_U 1\{Z = z_{j-K}\}/g(z_{j-K}) & j = K+1, \cdots, 2K, \end{cases} \tag{3.15}$$

where g denotes the marginal distribution of Z, which is assumed known for simplicity. Consider making inference for the value of the regression function evaluated at a counterfactual value $\tilde{z} \notin S_Z$. Then, the projection of interest is $\tilde{z}'\theta$. Note that the identified set is a polyhedron whose gradients are given by $D_{P,j}(\theta) = -z_j/\sigma_j, j = 1, \cdots, K$ and $D_{P,j}(\theta) = z_{j-K}/\sigma_{j-K}, j = K+1, \cdots, 2K$. This and $\tilde{z} \notin S_Z$ imply that for any θ not in the interior of the identified set, there exists a binding or violated constraint whose gradient is not a scalar multiple of p. Hence, for all such θ, our critical value is strictly smaller than $c_n^{AS}(\theta)$ asymptotically. In this case, our confidence interval becomes strictly shorter than that of AS asymptotically. We also note that the same argument applies even if the marginal distribution of Z is unknown. In such a setting, one needs to work with a sample constraint of the form $n^{-1}\sum_{i=1}^n Y_{L,i} 1\{Z_i = z_j\}/n^{-1}\sum_{i=1}^n 1\{Z_i = z_j\} - z_j\theta$ (and similarly for the upper bound). This change only alters the (co)variance of the Gaussian process in our limiting approximation but does not affect any other term.

We conclude this section with a numerical illustration. Assume that $p = (d^{-1/2}, ..., d^{-1/2}) \in \mathbb{R}^d$ and that there are d binding moment inequalities whose gradients are known and correspond to rows of the identity matrix. Assume furthermore that $\mathbb{G}$ is known to be exactly d-dimensional multivariate standard Normal. (Thus, Θ_I is the negative quadrant. Its unboundedness from below is strictly for simplicity.) Also ignore the ρ-box; if our heuristic for choosing ρ were followed, the influence of the ρ-box in this example would remain small as d grows.

Under these simplifying assumptions (which can, of course, be thought of as asymptotic

Table 3.1: Conservatism from projection in a one-sided testing problem as a function of d

d	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**100**	∞
$\hat{c}_n$	1.64	1.16	0.95	0.82	0.74	0.67	0.62	0.58	0.55	0.52	0.16	0
$\hat{c}_n^{AS}$	1.64	1.95	2.12	2.23	2.32	2.39	2.44	2.49	2.53	2.57	3.28	∞
$1-\alpha^*$	.95	.77	.57	.40	.27	.18	.11	.07	.04	.03	10^{-25}	0

approximations), it is easy to calculate in closed form that

$$\begin{aligned} \hat{c}_n &= d^{-1/2}\Phi^{-1}(1-\alpha), \\ \hat{c}_n^{AS} &= \Phi^{-1}\left((1-\alpha)^{1/d}\right). \end{aligned}$$

Furthermore, for any $\alpha < 1/2$, one can compute α^* s.t. applying $\hat{c}_n$ with target coverage $(1-\alpha)$ yields the same confidence interval as using $\hat{c}_n^{AS}$ with target coverage $(1-\alpha^*)$.[21] Some numerical values are provided in Table 3.1 (with $\alpha = 0.05$).

So, to cover $p'\theta$ in $\mathbb{R}^{10}$ with probability 95%, it suffices to project an AS-confidence region of size 3%. The example is designed to make a point; our Monte Carlo analyses below showcase less extreme cases. We note, however, that the core defining feature of the example – namely, the identified set has a thick interior, and the support set is the intersection of d moment inequalities – frequently occurs in practice, and all such examples will qualitatively resemble this one as d grows large.

4 Local Geometry of $\Theta_I(P)$ and Uniform Inference

As we discussed in the introduction, the main alternative to our method is based on a profiled test statistic as introduced in Romano and Shaikh (2008) and significantly advanced in BCS. We now explain in more detail how the class of DGPs over which our procedure and theirs are asymptotically uniformly valid are non-nested. We also compare our method with that of PPHI, which is based on directly bootstrapping the support function of a sample analog of the identified set.

As explained in Section 3.1, our method imposes Assumption 3.3 (or 3.3′), which is not imposed by either BCS nor PPHI, and uses an additional (non-drifting) tuning parameter ρ. From this, we reap several important benefits. We are able to establish validity of our method even when the local geometry of the set $\Theta_I(P)$ poses challenges to uniform inference as described below, and without imposing restrictions on the limit distribution of a test

[21] Equivalently, $(1-\alpha^*)$ is the probability that $\mathcal{C}_n(\hat{c}_n^{AS})$ contains $\{0\}$, the true support set in direction p which furthermore, in this example, minimizes coverage within $\Theta_I(P)$. The closed-form expression is $1-\alpha^* = \Phi(d^{-1/2}\Phi^{-1}(1-\alpha))^d$. AS prove validity of their method only for $\alpha < 1/2$, but this is not important for the point made here.

statistic, e.g. that it is continuous at the quantile of interest.

In particular, we allow for an extreme point of Θ_I in direction of projection to be (i) a point of differentiability of the boundary of Θ_I, (ii) a point on a flat face that is orthogonal to the direction of projection, or (iii) a point on a flat face that is drifting-to-orthogonal to the direction of projection. Case (iii) is excluded by Romano and Shaikh (2008) and BCS, and all three cases are excluded by PPHI. As already discussed in the introduction, drifting-to-flat faces occur, for example, in best linear prediction with interval outcome data and discrete regressors. They may also occur when Θ_I is drifting to be lower dimensional in the direction of projection, i.e. when the component of interest is drifting to being point identified. Our method remains valid also when Θ_I locally exhibits corners with extremely acute angles, meaning that the interior of Θ_I locally vanishes and that the joint linear approximation of constraints is not a good approximation to the local geometry of Θ_I. This case is again excluded by PPHI and also by Chernozhukov, Hong, and Tamer (2007).

We further illustrate these observations through a sequence of examples illustrating some key challenges faced by the existing alternative methods and how our approach handles them.

4.1 A Simple Example to Set the Stage

We begin with a one-sided testing problem similar to the one explored in Table 3.1.

EXAMPLE 4.1: Let $\Theta = [-K, K]^2$ for some $K > 0$ and moment functions be given by

$$m_1(x, \theta) = x^{(1)}(\theta_1 - 1)^2 + \theta_2 - x^{(2)} \tag{4.1}$$

$$m_2(x, \theta) = x^{(3)}(\theta_1 + 1)^2 + \theta_2 - x^{(4)}, \tag{4.2}$$

where we assume $X^{(l)}, l = 1, \cdots, 4$ are i.i.d. random variables with mean $\mu_x \geq 0$ and variance σ_x^2. The parameter of interest is θ_2. So, we let $p = (0, 1)'$.

The projection of $\theta \in \Theta_I$ is maximized at a unique point $\theta^* = 0$. For simplicity, consider constructing a one-sided confidence interval $CI_n = (-\infty, s(p, \mathcal{C}_n(\hat{c}_n))]$, where $s(p, \mathcal{C}_n(\hat{c}_n))$ is defined as in (2.2) with $J_1 = 2$ inequality restrictions with the moment functions in (4.1)-(4.2) and no equality restrictions. Then, θ^* represents the least favorable case for coverage by this one-sided confidence interval.

Now consider the linear program

$$\begin{aligned} &\sup_{\lambda \in \mathbb{R}^2} p'\lambda \\ &s.t.\ \mathbb{G}_n(\theta^*) + D_P(\theta^*)\lambda + \sqrt{n}\gamma_{1,P,n}(\theta^*) \leq c, \end{aligned} \tag{4.3}$$

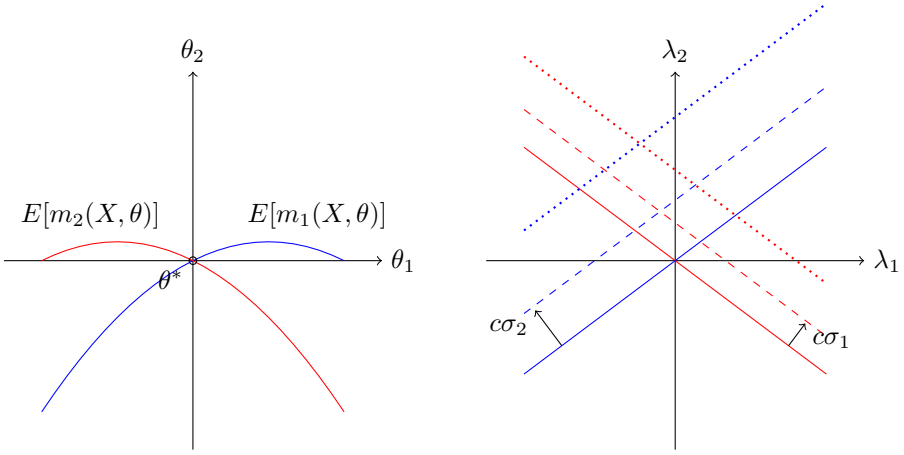

Figure 4.1: Moment inequalities (left) and linearized constraints (right)

where

$$\mathbb{G}_n(\theta^*) = \begin{pmatrix} \sqrt{n}(\bar{X}^{(1)} + \bar{X}_n^{(2)})/\sqrt{2}\sigma_x \\ \sqrt{n}(\bar{X}^{(3)} + \bar{X}_n^{(4)})/\sqrt{2}\sigma_x \end{pmatrix}, \tag{4.4}$$

$$D_P(\theta^*) = \begin{pmatrix} -2\mu_x/\sqrt{2}\sigma_x & 1/\sqrt{2}\sigma_x \\ 2\mu_x/\sqrt{2}\sigma_x & 1/\sqrt{2}\sigma_x \end{pmatrix}, \tag{4.5}$$

$$\gamma_{1,P,n}(\theta^*) = \begin{pmatrix} 0 & 0 \end{pmatrix}'. \tag{4.6}$$

This program is infeasible in the sense that it uses unknown population objects, in particular, the knowledge that both population moment inequalities bind at θ^*, hence $\gamma_{1,P,n}(\theta^*) = (0,0)'$. Though infeasible, it gives useful insights. Figure 4.1 shows the original nonlinear constraints and linearized constraints around θ^* perturbed by $\mathbb{G}_n$. The key idea of our procedure is to find $\hat{c}_n(\theta^*)$ such that the optimal value of the perturbed linear program in (4.3) is greater than or equal to 0 with probability $1-\alpha$, and use it in the original nonlinear problem upon projecting $\mathcal{C}_n(\cdot)$.

In Example 4.1, the value of the linear program in (4.3) has a closed form, namely $p'D_P^{-1}([c\ c]' - \mathbb{G}_n) = \sqrt{2}\sigma_x(c - W_n)$, where $W_n = (\mathbb{G}_{n,1} + \mathbb{G}_{n,2})/2$ has a limiting distribution $N(0, 1/2)$ (under a fixed (θ^*, P)). Therefore, by setting $\hat{c}_n(\theta^*)$ to 1.16, the 95%-quantile of $N(0,1/2)$, one can ensure the optimal value in (4.3) is nonnegative with probability 95% asymptotically.[22] This infeasible critical value is the baseline of our method. In practice, the researcher does not know whether a given θ is on the boundary of the identification region nor the population objects: the distribution of $\mathbb{G}_n(\theta)$ and $(D_P(\theta), \gamma_{1,P,n}(\theta))$. Our bootstrap

[22] This argument is based on a pointwise asymptotics, which fixes (θ^*, P) and sends n to ∞. This is done only for illustration purposes to obtain a specific value for $\hat{c}_n(\theta^*)$. Our proof does not use this argument. Note that the critical value calculated under this pointwise asymptotics depends on the covariance matrix of $\mathbb{G}_n$. For example, if $corr(\mathbb{G}_{n,1}, \mathbb{G}_{n,2})$=-0.9, it suffices to set $\hat{c}_n(\theta^*)$ to 0.37.

procedure therefore replaces them with suitable estimators.

In this extremely well-behaved example and when ignoring the ρ-box constraint, one can show that leading alternative approaches (BCS, PPHI) asymptotically agree with each other and with ours. Indeed, the support function here equals $p'\theta^*$ and is estimated by $p'\hat{\theta}$, where θ^* is the intersection of two constraints and $\hat{\theta}$ is its sample analog. Under assumptions maintained throughout the literature, $\hat{\theta}$ is asymptotically normal. Thus, if one knew a priori that this situation obtains, one could use a one-sided t-statistic based (bootstrap or plug-in asymptotic) confidence interval in this special case. Indeed, all of the aforementioned approaches asymptotically recover this interval. They can be thought of as generalizing it in different directions.

A caveat to this is that adding the ρ-box constraints conservatively distorts our confidence interval. Our proposal, explained later, for selecting ρ is designed to make the distortion small in well-behaved cases, but it is a distortion nonetheless.[23]

The similarity breaks down if the example is changed to an "overidentified" corner, e.g. a corner in $\mathbb{R}^2$ at which 3 constraints intersect. Note that GMS with hard thresholding makes such scenarios generic in the sense that their realization in sample is not knife-edge. Simulation of the support function, as advocated in PPHI, now leads to (potentially much) longer confidence intervals than our method. For a drastic but simple example, consider the minimum of two means in $\mathbb{R}$: We want to estimate $\min\{\mu_1, \mu_2\}$ and observe two signals $[\hat{\mu}_1, \hat{\mu}_1]'$ that have a bivariate Normal distribution with mean $[\mu_1, \mu_2]'$, covariance equal to zero, and variances, respectively, 1 and σ^2. Assume that $\mu_1 = \mu_2 = 0$; this setting could, of course, reflect recentering by a hard thresholding GMS procedure. Then the bootstrap sample support function is $\min\{\hat{\mu}_1, \hat{\mu}_2\}$. If $\sigma \gg 1$, the left tail of the distribution of $\min\{\hat{\mu}_1, \hat{\mu}_2\}$ is essentially determined by the distribution of $\hat{\mu}_2$, and the PPHI confidence interval is approximately $(-\infty, \min\{\hat{\mu}_1, \hat{\mu}_2\} + 1.645\sigma]$. (The approximation is favorable since tail probabilities of the more precise signal were ignored.) In contrast, AS and our method agree (because there is no projection in this example in $\mathbb{R}$) and approximately recover Bonferroni, thus our interval is similar to $(-\infty, \min\{\hat{\mu}_1 + 1.96, \hat{\mu}_2 + 1.96\sigma\}]$. (The approximation is unfavorable because our method actually exploits independence of $\hat{\mu}_1$ and $\hat{\mu}_2$.) For $\sigma = 10$, numerical evaluation without these approximations reveals that the upper bounds of the intervals have expected values 12.4 and 1.8, respectively, to be compared with a true value $\min\{\mu_1, \mu_2\} = 0$ and an expected value of the estimator $E(\min\{\hat{\mu}_1, \hat{\mu}_2\}) = -4$. The difference between the confidence intervals can be made arbitrarily large by increasing σ.

4.2 Flat Faces and Drifting-to-Flat Faces

Next, we consider a setting where the projection is maximized at multiple points. For this, we add, to the constraints in Example 4.1, one more inequality restriction whose moment

[23] In the present example, we would recommend $\rho \approx 2.8$, with negligible effect on c and on true coverage.

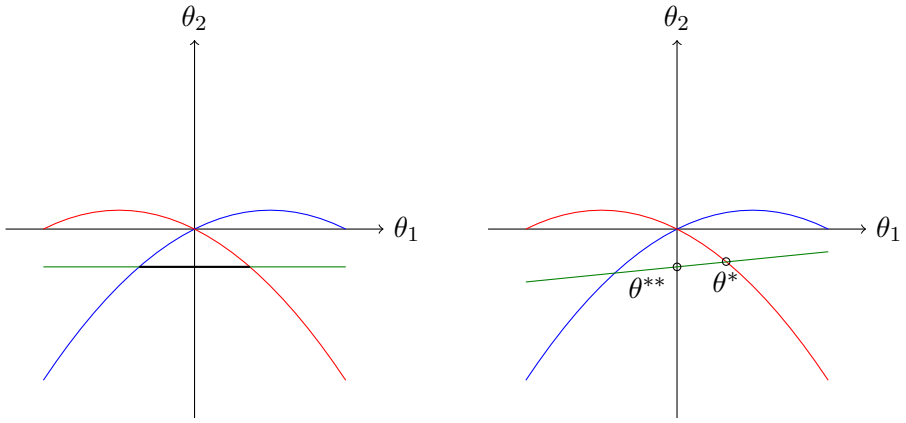

Figure 4.2: Flat face (left) and a near flat face (right)

function is given by

$$m_3(x,\theta) = x^{(5)}\theta_1 + \theta_2 + x^{(6)}, \tag{4.7}$$

where $X^{(5)}$ and $X^{(6)}$ are independent random variables independent from $X^{(1)}, \cdots, X^{(4)}$ with mean $E_P[X^{(5)}] = 0$, $E_P[X^{(6)}] = \mu_x$ and variance $Var_P(X^{(5)}) = Var_P(X^{(6)}) = \sigma_x^2$. (See Figure 4.2, left panel.)

The projection of $\theta \in \Theta_I$ is then maximized over the following set:

$$H(p,\Theta_I) = \{\theta \in \Theta : \theta_1 \in [1-\sqrt{2}, -1+\sqrt{2}], \theta_2 = -\mu_x\}. \tag{4.8}$$

In other words, the identification region has a flat face toward direction p. At each $\theta \in H(p,\Theta_I)$, one can study the infeasible linear program. For example, at $\theta^* = (1-\sqrt{2}, -\mu_x)$, the first and third moment inequalities bind, but not the second one. Then, the approximating linear program in (4.3) holds with $\sqrt{n}\gamma_{1,P,n}(\theta^*) = (0, -\sqrt{n}(4-2\sqrt{2})\mu_x, 0)'$. If the magnitude of the second component of $\sqrt{n}\gamma_{1,P,n}(\theta^*)$ is large, or along any sequence (θ_n, P_n) such that $\sqrt{n}\gamma_{1,P_n,2,n}(\theta_n) \to -\infty$, the second moment inequality becomes negligible. Solving for the optimal value using the two remaining constraints then yields $\sqrt{2}\sigma_x(c - W_n)$, where $W_n = \mathbb{G}_{n,3}(\theta^*)$ approximately follows the standard normal distribution, which suggests that $\hat{c}_n(\theta^*) = 1.645$, the usual one-sided critical value, can be used. However, if $\sqrt{n}\gamma_{1,P,2,n}(\theta^*)$ is close to 0, the second constraint is also relevant. In such cases, GMS will asymptotically replace $\sqrt{n}\gamma_{1,P,2,n}(\theta^*)$ with 0 and thus add the second inequality as an additional constraint. Like any tightening of constraints, this will increases $\hat{c}_n(\theta^*)$. The same argument applies to every θ in the support set. For example at $\theta = (0, -\mu_x)$, the third moment inequality is the only one that binds, which again defines another approximating linear program with a different local slackness parameter. Hence, the amount of relaxation needed to ensure the

one-sided coverage differs across points in $H(p, \Theta_I)$ due to different values of the slackness parameter.[24] Furthermore, the analysis also extends to settings where the identification region has a face whose normal vector is nearly aligned with p as shown in Figure 4.2, right panel. We come back to this case later in this section.

The presence of a flat face or more generally a non-singleton support set does not complicate our inference procedure because we calibrate the level at each θ. On the other hand, these features raise a nontrivial challenge for methods that use test statistics whose limiting distributions depend on $H(p, \Theta_I)$. For example, consider again the method that constructs a confidence interval from the support function of the estimated identified set. If the support set is not a singleton, the distribution of the normalized support function $S_n = \sqrt{n}[s(p, \mathcal{C}_n(0)) - s(p, \Theta_I(P))]$ can be shown to be approximated by the supremum of the optimal value in (4.3) over $H(p, \Theta_I)$; see, e.g., Kaido (2012). Hence, the support set becomes a nuisance parameter that affects the distribution of the statistic. Uniform size control then becomes challenging. In particular, for a sequence of DGPs P_n along which the support sets are singletons (i.e. $H(p, \Theta_I(P_n)) = \{\theta_n\}$ for all n) but non-singleton in the limit, the limiting distribution of the statistic changes discontinuously. We call such a setting "drifting-to-flat face". In the present example, one can construct such a sequence P_n by letting $E_{P_n}[X^{(5)}] > 0$ for all n and letting it drift to 0 (see Figure 4.2, right panel). To handle this issue, one must either assume away flat faces (toward direction p) or introduce a conservative distortion. Beresteanu and Molinari (2008, Assumption 4.5), PPHI, and Kaido and Santos (2014, Assumption 4.1) take the first approach, rendering them inapplicable to some commonly studied examples.[25]

Drifting-to-flat faces are also assumed away in the recent work of BCS. They consider testing the hypothesis $\mathcal{H}_0 : p'\theta = \beta_0$ and constructing a confidence interval through a test inversion. Their method is based on bootstrapping a profiled test statistic $\inf_{\{\theta : p'\theta = \beta_0\}} a_n Q_n(\theta)$, where Q_n is a sample criterion function which includes the use of GMS. A key role in profiling is played by the subset of elements of $\Theta_I(P)$ that satisfy the null hypothesis $\mathcal{H}_0$. When $\beta_0 = s(p, \Theta_I(P))$, this set coincides with the support set $H(p, \Theta_I(P))$. Although BCS's inference is valid over a class of distributions under which $H(p, \Theta_I(P))$ is not necessarily singleton-valued, they require that the population criterion function increases as a polynomial function of the distance from θ to $H(p, \Theta_I(P))$ when θ deviates from this set along the hyperplane $\{\theta : p'\theta = s(p, \Theta_I(P))\}$.[26] This requirement, however, excludes DGPs that

[24] In this specific example, $\hat{c}_n$ converges to 1.645 at all points in the support set because the constraint whose gradient is orthogonal to p reduces the problem to a one-sided testing problem. Finite sample critical levels will differ across $H(p, \Theta_I)$, though.

[25] For example, Beresteanu and Molinari (2008) show that the identification region for the best linear predictor of an interval-valued outcome variable with discrete covariates has flat faces. See also Freyberger and Horowitz (2015) for a nonparametric IV example with discrete variables.

[26] Without this requirement, their estimator of $H(p, \Theta_I(P))$ may include points at which population moment (in)equalities are violated but by not much. At such points, the sample moment inequalities may even realize as slack constraints, and hence replacing the (violated) population local slackness parameter with the GMS

exhibit drifting-to-flat faces. For example, in the right panel of Figure 4.2, consider deviating from θ^* toward direction $(-1, 0)$. Because of the third constraint drifting to a flat face, one can make the population criterion function increase arbitrarily slowly along such a deviation.

4.3 Role of the ρ-Box Constraint

We next discuss why we impose the additional constraint $\lambda \in \rho B^d$. To do so, we return to Example 4.1 (without the additional constraint (4.7)). Recall that

$$D_P(\theta^*) = \begin{pmatrix} -2\mu_x/\sqrt{2}\sigma_x & 1/\sqrt{2}\sigma_x \\ 2\mu_x/\sqrt{2}\sigma_x & 1/\sqrt{2}\sigma_x \end{pmatrix} \tag{4.9}$$

and consider a sequence of DGPs such that $\mu_x \to 0$. As we saw before, under each DGP with $\mu_x > 0$, the infeasible linear program calibrates $\hat{c}_n(\theta^*) = 1.16$. In the limit, however, the moment inequalities reduce to

$$\theta_2 - E_P[X^{(2)}] \leq 0 \tag{4.10}$$

$$\theta_2 - E_P[X^{(4)}] \leq 0. \tag{4.11}$$

In other words, θ_2's upper bound is given by the minimum of the two means: $E_P[X^{(2)}]$ and $E_P[X^{(4)}]$. This structure is also known as "intersection bounds" (Hirano and Porter, 2012). The value of the linear program in (4.3) is then $\min\{c - \mathbb{G}_{n,1}, c - \mathbb{G}_{n,2}\}$. To ensure coverage, one needs a critical level of $\hat{c}_n(\theta^*) = 1.95$ instead of 1.16 (the slight difference to 1.96 is because we exploit independence of error terms). This discontinuity presents another challenge for uniform validity of inference. For any setting where the constraints are close to the minimum of the two means, an inference method that does not take into account this feature would have poor size control.

This type of example is the main reason why we restrict the localization parameter λ into the ρ-box. To see the benefit, consider Figure 4.3. The figure shows the DGP on the left panel and a realization of a constraint in the bootstrap problem on the right panel. Due to sampling variation, the estimated gradients $\hat{D}_{n,1}$ and $\hat{D}_{n,2}$ differ slightly from the population gradients. Without the ρ-box constraint, the maximum is attained at λ^*. Since the estimated gradients are fixed across bootstrap replications, $p'\lambda^*$ behaves as approximately normal, and by the previous argument we would end up with $\hat{c}_n(\theta^*) = 1.16$. With the ρ-box, however, the optimum is attained at λ^{**} whose projection is the minimum of the projections of two points at which the two constraints intersect with the right boundary of ρB^d. Therefore, our bootstrap procedure mimics the minimum of the two-means problem. This scenario is very likely to occur in bootstrap samples whenever the population gradients are close to this

function does not necessarily provide conservative approximations. For details, we refer to discussions provided in Bugni, Canay, and Shi (2015, page 265).

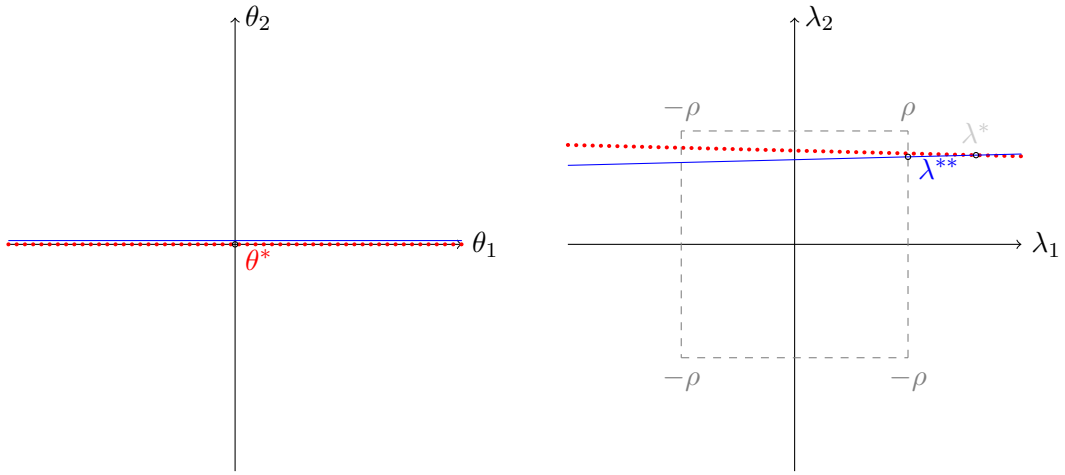

Figure 4.3: Minimum of two means and a ρ-box

situation, and hence restricting λ to the ρ-box is key to uniform validity of our procedure.[27]

The drifting-to-flat face example in Figure 4.2, right panel, can be handled analogously. For example, for some points such as θ^{**}, the relevant constraint is the third constraint, which is drifting-to-flat. A linearized problem around θ^{**} then looks akin to the right panel of Figure 4.3 without the dotted line. Calculating a bootstrap critical value then yields a one-sided critical value $\hat{c}_n(\theta) = 1.645$ as before.

In practice, the choice of ρ requires trading off how much conservative bias one is willing to bear in well-behaved cases (e.g., Example 4.1) against how much finite-sample size distortion one is willing to bear in ill-behaved cases such as the minimum of two means example just described. We propose a heuristic approach to calibrate ρ focusing on conservative bias in well behaved cases. In these cases, the optimal value is distributed asymptotically normal as a linear combination of d binding inequalities. When in fact $J_1 + J_2 = d$, constraining $\lambda \in \rho B^d$ increases the coverage probability by at most $\beta = 1 - [1 - 2\Phi(-\rho)]^d$. The parameter ρ can therefore be calibrated to achieve a conservative bias of at most β. When $J_1 + J_2 > d$, we propose to calibrate ρ using the benchmark

$$\beta = 1 - [1 - 2\Phi(-\rho)]^{d\binom{J_1+J_2}{d}},$$

again inferring ρ so as to achieve a target conservative bias (in well-behaved cases) of β. A few numerical examples with $\beta = 0.01$ yield, with $J_1 + J_2 = 10$ and $d = 3$ a value of $\rho = 4.2$; with $J_1 + J_2 = 100$ and $d = 10$, $\rho = 8.4$.

[27] This reasoning does not go through if the two constraints are perfectly correlated; their bootstrap resamples might then always intersect inside the ρ-box despite the very acute angle formed. This is precisely why we restrict the correlation matrix of moments, but also why we only need this restriction for moment conditions whose boundaries may intersect.

Table 5.1: DGPs used in the Monte Carlo experiments 1-4

DGP	Moment Conditions	Projections of Θ_I	Description
DGP-1	$\theta_1 + \theta_2 \leq E_P[X_1]$ $-\theta_1 + \theta_2 \leq E_P[X_2]$ $\theta_1 - \theta_2 \leq E_P[X_3] + 2$ $-\theta_1 - \theta_2 \leq E_P[X_4] + 2$	$\theta_1 \in [-1, 1]$ $\theta_2 \in [-2, 0]$	Θ_I is a square.
DGP-2	$\theta_1\sqrt{n} + \theta_2 \leq E_P[X_1] - 1 + 1/\sqrt{n}$ $-\theta_1\sqrt{n} + \theta_2 \leq E_P[X_2] - 1 + 1/\sqrt{n}$ $\theta_1\sqrt{n} - \theta_2 \leq E_P[X_3] + 1 + 1/\sqrt{n}$ $-\theta_1\sqrt{n} - \theta_2 \leq E_P[X_4] + 1 + 1/\sqrt{n}$	$\theta_1 \in [-1, 1]$ $\theta_2 \in [-1 - \frac{1}{\sqrt{n}}, -1 + \frac{1}{\sqrt{n}}]$	Θ_I is local to a thin face.
DGP-3	$\theta_1 + \theta_2 \leq E_P[X_1] + 1/\sqrt{n}$ $-\theta_1 + \theta_2 \leq E_P[X_2] + 1/\sqrt{n}$ $\theta_1 - \theta_2 \leq E_P[X_3] + 1/\sqrt{n}$ $-\theta_1 - \theta_2 \leq E_P[X_4] + 1/\sqrt{n}$	$\theta_1 \in [-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$ $\theta_2 \in [-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$	Θ_I is local to point identification.
DGP-4	$\theta_1 + \theta_2 \leq E_P[X_5]$ $-\theta_1 + \theta_2 \leq E_P[X_6]$ $\theta_1 - \theta_2 \leq E_P[X_7] + 2$ $-\theta_1 - \theta_2 \leq E_P[X_8] + 2$ and the inequalities in DGP-1.	$\theta_1 \in [-1, 1]$ $\theta_2 \in [-2, 0]$	The corners of Θ_I are overidentified.

Table notes: (1) For each DGP, the projection of interest is defined by $p'\theta : \theta \in \Theta_I$ with $p = (0, 1)'$ and $\theta = (\theta_1, \theta_2)'$; (2) $X_1, \cdots, X_8$ are i.i.d. Normal random variables, with $Var(X_k) = 1,\ k = 1, \ldots, 4$, $Var(X_5) = Var(X_7) = 4$ and $Var(X_6) = Var(X_8) = 9$.

5 Monte Carlo Simulations

We evaluate the performance of our confidence intervals in two sets of Monte Carlo experiments. The first set examines linear restrictions in a two-dimensional parameter space. This illustrates the performance of our procedure under DGPs that make inference for projections nontrivial, but where our method is still easy to visualize. The second set of experiments is about a two-player entry game commonly studied in the literature. With $d = J_1 = J_2 = 8$, the DGPs considered there have interesting complexity.

Another important class of models are moment inequalities that arise from revealed preference considerations in games, as laid out in Pakes, Porter, Ho, and Ishii (2015). We refer the reader to Mohapatra and Chatterjee (2015) for an empirical application of our method in such a model with $d = 5$, $J_1 = 44$, and $J_2 = 0$.

5.1 Linear Restrictions in $\mathbb{R}^2$

All DGPs in this subsection are parameterized by $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$. We take the second component to be the projection of interest, thus $p = (0, 1)'$. Further details of the specifications are listed in Table 5.1.

DGP-1 has a square-shaped identified set defined by four linear inequalities, with length

Table 5.2: Simulation result for DGPs 1-4 with $n = 3000$, $MCs = 1000$.

	$1-\alpha$	Average CI		KMS Coverage		AS Coverage		Average $\hat{c}$		Excess Length	
		KMS	AS	Upper	Lower	Upper	Lower	KMS	AS	KMS	AS
DGP-1	95%	[-2.021,0.021]	[-2.041,0.040]	94.5%	94.9%	100%	99.9%	1.161	1.955	0.042	0.071
	90%	[-2.017,0.016]	[-2.036,0.035]	89.6%	90.8%	99.8%	99.6%	0.906	1.634	0.033	0.059
	85%	[-2.013,0.013]	[-2.032,0.032]	84.4%	85.3%	99.5%	98.9%	0.732	1.419	0.026	0.052
	80%	[-2.011,0.010]	[-2.029,0.029]	78.2%	79.7%	99.2%	98.0%	0.595	1.251	0.021	0.045
	75%	[-2.009,0.008]	[-2.027,0.027]	74.0%	75.5%	98.6%	97.4%	0.476	1.108	0.017	0.040
DGP-2	95%	[-1.058,-0.942]	[-1.059,-0.941]	100%	99.9%	100%	99.9%	2.175	2.236	0.079	0.081
	90%	[-1.053,-0.948]	[-1.054,-0.947]	99.7%	99.3%	99.8%	99.6%	1.888	1.945	0.069	0.071
	85%	[-1.049,-0.951]	[-1.050,-0.950]	99.4%	98.6%	99.5%	98.9%	1.699	1.755	0.062	0.064
	80%	[-1.047,-0.954]	[-1.048,-0.953]	99.0%	97.8%	99.2%	98.0%	1.552	1.607	0.056	0.058
	75%	[-1.044,-0.956]	[-1.045,-0.955]	98.2%	96.8%	98.6%	97.4%	1.429	1.482	0.052	0.054
DGP-3	95%	[-0.042,0.041]	[-0.055,0.055]	98.4%	96.9%	100%	99.9%	1.305	2.074	0.047	0.073
	90%	[-0.038,0.037]	[-0.050,0.049]	95.6%	94.7%	99.7%	99.5%	1.087	1.785	0.039	0.063
	85%	[-0.035,0.035]	[-0.046,0.046]	92.2%	93.0%	99.4%	98.4%	0.945	1.598	0.034	0.056
	80%	[-0.033,0.033]	[-0.044,0.043]	89.4%	90.0%	98.6%	97.7%	0.831	1.452	0.030	0.050
	75%	[-0.031,0.031]	[-0.041,0.041]	85.6%	87.6%	98.1%	96.9%	0.736	1.330	0.026	0.046
DGP-4	95%	[-2.024,0.024]	[-2.038,0.038]	95.0%	94.2%	99.8%	99.5%	1.610	2.234	0.048	0.076
	90%	[-2.019,0.018]	[-2.032,0.031]	89.0%	89.2%	98.7%	98.5%	1.373	1.940	0.037	0.063
	85%	[-2.014,0.014]	[-2.027,0.027]	83.5%	84.1%	97.3%	96.3%	1.211	1.746	0.029	0.055
	80%	[-2.011,0.011]	[-2.024,0.024]	77.4%	79.7%	95.4%	93.9%	1.082	1.595	0.022	0.048
	75%	[-2.008,0.008]	[-2.021,0.021]	72.3%	74.6%	92.0%	91.8%	0.974	1.468	0.016	0.042

Table notes: (1) The projection of interest is θ_2 for $(\theta_1, \theta_2) \in \Theta_I$. (2) "Upper" coverage refers to coverage of $\max\{p'\theta : \theta \in \Theta_I\}$, and similarly for "Lower". (3) The excess length of a confidence interval (CI) is computed as length of CI - length of population projection. (4) $B = 2001$ bootstrap draws.

of the projection of interest equal to 2. DGP-2 is similar, but the slope of each constraint equals $n^{-1/2}$ in absolute value so the length of projection is $2/\sqrt{n}$. Therefore, as n grows, the identified set converges to the line segment spanned by $\{(-1,-1),(1,-1)\}$. This specification is used to examine the performance of the confidence intervals when the identified set is local to a thin face in the direction of projection. In particular, note that DGP-2 converges to point identification of $p'\theta$ but not of θ. DGP-3 is again similar to DGP-1, but shrinks toward the singleton $\{(0,-1)\}$ as n grows, so that point identification of θ is approached; length of projection is again $2/\sqrt{n}$. Finally, DGP-4 adds four additional inequalities to DGP-1. These have exactly the same form as the ones in DGP-1, hence the length of projection stays at 2, but differ in their variances. This specification is used to evaluate the performance of the confidence interval when some of the boundary points of Θ_I are overidentified.

Table 5.2 reports the results of the Monte Carlo experiments under alternative nominal coverage levels (95%, 90%, 85%, 80%, 75%). The DGPs are simple enough so that this table can be computed using Matlab's `fmincon` command as well as our E-A-M algorithm. We did both, with identical results. We report separate coverage probabilities for the upper and lower bound on $p'\theta$, average critical levels $\hat{c}_n$ at the upper and lower support point of $\mathcal{C}_n$, and average excess lengths of confidence intervals. By "excess length," we mean the difference between the length of the confidence interval and that of the projection of the identified set.

For simplicity, we refer to "AS confidence intervals" below when we mean the projections of AS confidence regions.

DGP-1 is the benchmark specification. In this setting, the coverage probabilities of our confidence intervals are very close to nominal levels, while those of the AS confidence intervals are much higher and close to 100% in most cases. This is also reflected in the higher critical levels and larger excess lengths displayed in the table. The comparison of $\hat{c}_n$ and $\hat{c}_n^{AS}$ also provides an example of the theoretical result presented in Section 3.2, where we showed that our critical level is strictly lower than AS's unless all constraints are orthogonal to the direction of projection. In this well-behaved example, the difference is large and would be even larger if the example were extended to higher dimensions.[28] Results are similar under DGP-4, suggesting that our confidence interval performs well in the presence of overidentifying moment restrictions.

A notable specification in which our CI and AS's CI perform similarly is DGP-2. Recall that, in this setting, the identified set is local to a thin face in the direction of projection; at our sample size of $n = 3000$, the numerical value of the slope is $\pm.018$. In the presence of the ρ-box, this makes our linearization of the inference problem similar to an overidentified two-sided test. Therefore, $\hat{c}_n$ and $\hat{c}_n^{AS}$ are close to each other and also to Bonferroni correction (not displayed). They also have similar coverage properties.[29]

Under DGP-3, where the identified set is local to point identification, the coverage probabilities of our CIs are again strictly above the nominal levels. This conservatism reflects GMS and is shared by both reported, and other uniformly valid, approaches. Projection of AS incurs considerable additional conservatism.

In sum, these experiments confirm that our confidence interval controls size well under various DGPs and is substantially less conservative than AS, except for the special case where the DGP is statistically indistinguishable from one that does not involve projection conservatism.

5.2 An Entry Game in $\mathbb{R}^8$

Consider the following variation on Example 3.1:

	$Y_2 = 0$	$Y_2 = 1$
$Y_1 = 0$	$0, 0$	$0, Z_2'\beta_1 + u_2$
$Y_1 = 1$	$Z_1'\beta_1 + u_1, 0$	$Z_1'(\beta_1 + \Delta_1) + u_1, Z_2'(\beta_2 + \Delta_2) + u_2$

[28] To get an idea, compare the average values of critical levels for $1 - \alpha = 95\%$ to the corresponding entries for $n = 2$ in Table 3.1. This comparison also corroborates numerical accuracy of our simulations, as well as minimal influence of the ρ-box constraints in well-behaved examples.

[29] They are conservative because, as long as inequalities are not exactly parallel, our $\hat{c}_n$ for DGP-1 would actually do, but this fact is not knowable to the researcher. Compare the discussion of Figure 4.3.

where $Y_k \in \{0,1\}$, Z_k, and u_k denote, respectively, player k's binary action, observed characteristics, and unobserved characteristics. The strategic interaction effects $Z_k'\Delta_k < 0, k = 1,2$ measure the impacts of the opponent's entry into the market. In what follows, we let $X \equiv (Y_1, Y_2, Z_1', Z_2')'$ and $\theta \equiv (\beta_1', \beta_2', \Delta_1', \Delta_2')'$. We generate $Z = (Z_1, Z_2)$ as an i.i.d. random vector taking values in a finite set whose distribution $p_z = P(Z = z)$ is assumed known. We then generate $u = (u_1, u_2)$ as standard bivariate Normal random variables independent of Z. The outcome $Y = (Y_1, Y_2)$ is generated as a pure strategy Nash equilibrium of the game. For some value of Z and u, the model predicts monopoly outcomes $Y = (0,1)$ and $(1,0)$ as multiple equilibria. When this is the case, we select Y by independent Bernoulli trials with fixed parameter $\tau \in [0,1]$.

The model gives rise to the following moment equality and inequality restrictions (Tamer, 2003; Ciliberto and Tamer, 2009):

$$P((0,0)|Z) = P(u_1 \leq -Z_1'\beta_1,\ u_2 \leq -Z_2'\beta_2) \tag{5.1}$$

$$P((1,1)|Z) = P(u_1 > -Z_1'(\beta_1 + \Delta_1),\ u_2 > -Z_2'(\beta_2 + \Delta_2)) \tag{5.2}$$

$$P((0,1)|Z) \leq P(u_1 \leq -Z_1'(\beta_1 + \Delta_1),\ u_2 > -Z_2'\beta_2) \tag{5.3}$$

$$\begin{aligned} P((0,1)|Z) \geq\ & P(u_1 \leq -Z_1'(\beta_1 + \Delta_1),\ u_2 > -Z_2'(\beta_2 + \Delta_2)) \\ & + P(u_1 \leq -Z_1'\beta_1,\ -Z_2'\beta_2 \leq u_2 \leq -Z_2'(\beta_2 + \Delta_2)). \end{aligned} \tag{5.4}$$

The inequality restrictions (5.3)-(5.4) bound the probability of an outcome that can be selected from multiple equilibria. Using our specification, it is straightforward to rewrite the restrictions as follows:

$$E[1\{Y = (0,0)\}1\{Z = z\}] - \Phi(-z_1'\beta_1)\Phi(-z_2'\beta_2)p_z = 0 \tag{5.5}$$

$$E[1\{Y = (1,1)\}1\{Z = z\}] - (1 - \Phi(-z_1'(\beta_1 + \Delta_1)))(1 - \Phi(-z_2'(\beta_2 + \Delta_2)))p_z = 0 \tag{5.6}$$

$$E[1\{Y = (0,1)\}1\{Z = z\}] - \Phi(-z_1'(\beta_1 + \Delta_1))(1 - \Phi(-z_2'\beta_2))p_z \leq 0 \tag{5.7}$$

$$\begin{aligned} & - E[1\{Y = (0,1)\}1\{Z = z\}] \\ & + \Big[\Phi(-z_1'(\beta_1 + \Delta_1))(1 - \Phi(-z_2'(\beta_2 + \Delta_2))) + \Phi(-z_1'\beta_1)(\Phi(-z_2'(\beta_2 + \Delta_2)) - \Phi(-z_2'\beta_2))\Big]p_z \leq 0, \end{aligned} \tag{5.8}$$

where Φ is the CDF of the standard normal distribution.

The complexity of this model depends on the support of Z. We work with a constant and a player specific, binary covariate, so $Z_1 \in \{(1,-1),(1,1)\}$ and $Z_2 \in \{(1,-1),(1,1)\}$. Z therefore takes four different values, giving rise to 8 moment equalities and 8 moment inequalities, i.e. $J = 24$ restrictions. The standard deviation of each moment takes the form $\big(E[1\{Y = y\}1\{Z = z\}](1 - E[1\{Y = y\}1\{Z = z\}])\big)^{1/2}$, which we estimate by its sample analog. The gradients of each moment can be computed analytically using (5.5)-(5.8). The

Table 5.3: Simulation result for DGP-5 with $n = 4000$, $MCs = 1000$.

	$1-\alpha$	Median CI		Coverage		Average CI Length	
		KMS	AS	KMS	AS	KMS	AS
$\beta_1^{[1]} = 0.50$	0.95	[0.344,0.763]	[0.125,0.941]	95.7%	100.0%	0.425	0.815
	0.90	[0.368,0.723]	[0.169,0.903]	92.2%	100.0%	0.356	0.735
	0.85	[0.381,0.698]	[0.194,0.880]	88.3%	99.7%	0.326	0.685
$\beta_1^{[2]} = 0.25$	0.95	[0.098,0.367]	[-0.003,0.490]	96.6%	100.0%	0.275	0.508
	0.90	[0.117,0.349]	[0.021,0.465]	93.1%	99.8%	0.236	0.455
	0.85	[0.128,0.340]	[0.035,0.449]	90.5%	99.6%	0.217	0.423
$\Delta_1^{[1]} = -1$	0.95	[-1.386,-0.701]	[-1.717,-0.292]	96.3%	100.0%	0.692	1.432
	0.90	[-1.327,-0.744]	[-1.654,-0.367]	92.3%	100.0%	0.588	1.291
	0.85	[-1.291,-0.775]	[-1.614,-0.412]	88.4%	99.9%	0.522	1.207
$\Delta_1^{[2]} = -1$	0.95	[-1.183,-0.753]	[-1.445,-0.494]	96.6%	100.0%	0.438	0.955
	0.90	[-1.154,-0.787]	[-1.400,-0.541]	93.1%	99.9%	0.375	0.862
	0.85	[-1.134,-0.811]	[-1.371,-0.570]	88.7%	99.9%	0.337	0.805

Table notes: (1) Population projection length is zero in this DGP. (2) $B = 2001$ bootstrap draws.

estimator of the normalized gradients can then be computed by dividing each gradient by the corresponding estimated standard deviation.

In our DGP-5, we set $\beta_1 = (.5, .25)'$ and $\Delta_1 = (-1, -1)'$. DGP-6 differs by setting $\Delta_1 = (-1, -.75)'$. In both cases, $(\beta_2, \Delta_2) = (\beta_1, \Delta_1)$ and the equilibrium selection probability is $\tau = 0.5$; we only report results for (β_1, Δ_1). Although parameter values are similar, there is a qualitative difference: In DGP-5, parameters turn out to be point identified. In DGP-6, they are not but the identified set is still not large compared to sampling uncertainty, specifically: for $\beta_1^{[1]}$, the projection of the identified set is $[0.405, 0.589]$; for $\beta_1^{[2]}$, it is $[0.236, 0.26]$; for $\Delta_1^{[1]}$, it is $[-1.158, -0.832]$; for $\Delta_1^{[2]}$, it is $[-0.790, -0.716]$. We therefore expect all methods that use GMS to be conservative in DGP-6. Finally, the marginal distribution of $(Z_1^{[2]}, Z_2^{[2]})$ on its support $\{(-1,-1), (-1,1), (1,-1), (1,1)\}$ is specified as $(0.1, 0.2, 0.3, 0.4)$.

An interesting feature of this model is that despite being (in general, and in one of our specifications) partially identified, it is also testable because moment conditions are overidentifying in some dimensions. More specifically, it can be verified that one of the four constraints corresponding to (5.5), and similarly one of the four constraints in (5.6), can be expressed as nonlinear function of the others. Indeed, this is one reason why DGP-6 is partially identified despite the presence of 8 equalities in $\mathbb{R}^8$. The additional constraints do, however, restrict the distribution of observables and therefore make the models testable.

"Supernumerary" or "partially overidentifying" moment conditions raise interesting questions. For one thing, their presence means that the sample analog of Θ_I is generically empty, a frequent feature of empirical applications but one that makes consistent estimation of identified sets difficult. Also, in our framework, they increase $\hat{c}_n$ because they act like implicit specification tests: Their rejection will cause the confidence interval to be empty. Ceteris

Table 5.4: Simulation result for DGP-6 with $n = 4000$, $MCs = 1000$.

	$1-\alpha$	Median CI		KMS Coverage		AS Coverage		Excess Length	
		KMS	AS	Lower	Upper	Lower	Upper	KS	AS
$\beta_1^{[1]} = 0.50$	0.95	[0.218,0.819]	[-0.009,1.002]	98.3%	98.2%	100.0%	100.0%	0.424	0.828
	0.90	[0.253,0.787]	[0.039,0.968]	95.8%	95.8%	100.0%	100.0%	0.351	0.748
	0.85	[0.272,0.762]	[0.069,0.947]	92.3%	92.9%	100.0%	100.0%	0.308	0.697
$\beta_1^{[2]} = 0.25$	0.95	[0.100,0.389]	[-0.003,0.524]	98.1%	98.2%	100.0%	100.0%	0.268	0.515
	0.90	[0.119,0.368]	[0.021,0.498]	95.1%	95.2%	99.8%	100.0%	0.229	0.460
	0.85	[0.131,0.355]	[0.035,0.481]	92.3%	91.2%	99.8%	99.7%	0.204	0.428
$\Delta_1^{[1]} = -1$	0.95	[-1.525,-0.470]	[-1.867,-0.015]	98.3%	98.3%	100.0%	100.0%	0.730	1.501
	0.90	[-1.472,-0.529]	[-1.803,-0.101]	95.9%	95.9%	100.0%	100.0%	0.616	1.374
	0.85	[-1.436,-0.565]	[-1.764,-0.156]	92.9%	92.1%	100.0%	100.0%	0.546	1.281
$\Delta_1^{[2]} = -0.75$	0.95	[-0.986,-0.490]	[-1.277,-0.230]	98.0%	98.0%	100.0%	100.0%	0.430	0.981
	0.90	[-0.956,-0.522]	[-1.226,-0.275]	95.9%	95.6%	100.0%	100.0%	0.368	0.884
	0.85	[-0.937,-0.543]	[-1.194,-0.304]	92.6%	92.4%	100.0%	100.0%	0.326	0.824

Table notes: (1) Population projections are as follows: for $\beta_1^{[1]}$, $[0.405, 0.589]$; for $\beta_1^{[2]}$, $[0.236, 0.26]$; for $\Delta_1^{[1]}$, $[-1.158, -0.832]$; for $\Delta_1^{[2]}$, $[-0.790, -0.716]$. (2) "Upper" coverage refers to coverage of $\max\{p'\theta : \theta \in \Theta_I\}$, and similarly for "Lower". (3) The excess length of a confidence interval (CI) is computed as length of CI - length of population projection. (4) $B = 2001$ bootstrap draws.

paribus, this makes confidence intervals larger, but it has to be traded off against potentially more efficient estimation. Unlike with the point identified case, the trade-off is not obvious, and we leave its analysis for future research.

Tables 5.3-5.4 summarize our findings. DGP-5 is characterized by moderate, and DGP-6 by considerable, conservatism due to GMS contracting several constraints in most samples.[30] In both examples, we decisively outperform AS both in terms of finite sample size as well as length of confidence interval. Last but not least, Tables 5.3-5.4 serve as proof of feasibility: With 3 different coverage probabilities and 1000 MC replications, we computed the confidence interval for each component and for each method (our own and AS) 3000 times. Our ability to do so in a speedy manner critically relies on the E-A-M algorithm.

6 Conclusions

This paper introduces a bootstrap-based confidence interval for linear functions of parameter vectors that are partially identified through finitely many moment (in)equalities. The extreme points of our confidence interval are obtained by minimizing and maximizing $p'\theta$ subject to the sample analog of the moment (in)equality conditions properly relaxed. This relaxation amount, or critical level, is computed to insure that $p'\theta$, rather than θ itself, is uniformly asymptotically covered with a prespecified probability. Calibration of the critical levels is computationally attractive because it is based on repeatedly checking feasibility of (bootstrap)

[30] This diagnosis is corroborated by: (i) closed-form analysis of simple high-dimensional models, which indicates that GMS can have a strong effect; (ii) simulations with $\kappa_n \approx 0$, in which we encountered slight undercoverage. Details are available from the authors.

linear programming problems. Computation of the extreme points of the confidence intervals is also computationally attractive thanks to an application, novel to this paper, of the response surface method for global optimization that can be of independent interest in the partial identification literature.

The class of DGPs over which we can establish validity of our procedure is non-nested with the class over which the main alternative to our method (Romano and Shaikh, 2008; Bugni, Canay, and Shi, 2014) is asymptotically valid. For example, our method yields asymptotically uniformly valid confidence intervals for linear functions of best linear predictor parameters in models with interval valued outcomes and discrete covariates, while profiling based methods do not. The price to pay is the use of one additional (non-drifting) tuning parameter.

The confidence region that we propose is by construction an interval. As such, it does not pick up gaps in the projection. To do so, one could replace our proposed confidence interval with the mathematical projection of $\mathcal{C}_n(\hat{c}_n)$, that is, with

$$\left\{ p'\theta : n^{-1} \sum_{i=1}^{n} m_j(X_i, \theta)/\hat{\sigma}_{n,j}(\theta) \leq \hat{c}_n(\theta),\ j = 1, \ldots, J \right\}.$$

Theorems 3.1 and 3.2 apply to this object, including if the projection of the AS confidence region is defined analogously to the above (and therefore also captures gaps). However, computation of this object is much harder, and so we recommend it only if possible gaps in the identified set for $p'\theta$ are genuinely interesting.

Also, and similarly to confidence regions proposed in AS, Stoye (2009), and elsewhere, our confidence interval can be empty, namely if the sample analog of the identified set is empty and if violations of moment inequalities exceed $\hat{c}_n(\theta)$ at each θ. Emptiness of CI_n can be interpreted as rejection of maintained assumptions. See Stoye (2009) and especially AS for further discussion and Bugni, Canay, and Shi (2015) for a paper that focuses on this interpretation and improves on $\hat{c}_n^{AS}$ for the purpose of specification testing. We leave a detailed analysis of our implied specification test to future research.

In applications, a researcher might wish to obtain a confidence interval for a nonlinear function $f : \Theta \mapsto \mathbb{R}$. Examples might include policy analysis and counterfactual estimation in the presence of partial identification or demand extrapolation subject to rationality constraints. While our results are formally derived for the case that f is linear in θ, the extension to uniformly continuously differentiable functions f is immediate. In particular, we propose to calibrate $\hat{c}_n$ as

$$\hat{c}_n(\theta) \equiv \inf\{c \geq 0 : P(\Lambda_n^b(\theta, \rho, c) \cap \{\nabla_\theta f(\theta)\lambda = 0\} \neq \emptyset) \geq 1 - \alpha\}, \tag{6.1}$$

where $\nabla_\theta f(\theta)$ is the gradient of $f(\theta)$. The lower and upper points of the confidence interval

are then obtained solving

$$\min_{\theta\in\Theta} / \max_{\theta\in\Theta} f(\theta) \quad \text{s.t.} \quad \sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta) \leq \hat{c}_n(\theta), \ j = 1, \cdots, J.$$

A related extension is inference on $E_P(f(X_i, \theta))$ for some known function f. Note that, while f is known, the expectation needs to be estimated even if θ is known. To handle this situation, we propose to apply our method to the augmented parameter vector $\tilde{\theta} = (E(f(X_i, \theta)), \theta')'$ and direction of optimization $p = (1, 0, ..., 0)$.[31]

Another extension that is of interest in applications is one where the moment conditions depend on a point identified parameter vector Π for which a consistent and asymptotically normal estimator $\hat{\Pi}_n(\theta_0)$ exists when θ_0 is the true value of θ. The sample moment functions are then of the form $\bar{m}_{n,j}(\theta) = \bar{m}_{n,j}(\theta, \hat{\Pi}_n(\theta))$. As explained in AS, the estimator of the asymptotic variance of $\sqrt{n}\bar{m}_{n,j}(\theta)$ needs to be adjusted to reflect that Π is replaced by an estimator. With this modification, and in line with AS, our results remain valid under conditions provided by AS to guarantee that $n^{-1}\sum_{i=1}^{n} m_j(X_i, \theta, \hat{\Pi}_n(\theta))$ is asymptotically Normal.

Yet another extension considers projection in a direction p that is unknown but is $\sqrt{n}$-consistently estimated by $\hat{p}$.[32] Our method applies without modification, treating the estimator $\hat{p}$ as if it were the true direction p, by retilting the gradients of the constraints. Combinations of each of these extensions are of course possible.

While our analysis is carried out with the criterion function in equation (3.7), it is also easy to show that our method (including the bootstrap procedure described in Section 2.2) applies similarly to a criterion function of the form

$$\tilde{T}_n(\theta) = \sum_{j=1,\cdots,J_1} \sqrt{n}[\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)]_+ + \sum_{j=J_1+1,\cdots,J_1+J_2} \sqrt{n}|\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)|, \tag{6.2}$$

Criterion function T_n corresponds to criterion function S_3 in AS; criterion function $\tilde{T}_n$ is akin to criterion function S_1 in AS. In addition, AS consider a QLR based test statistic previously proposed in Rosen (2008). This test statistic does not lend itself easily to linearization, and as such we do not consider it in this paper.

Finally, our method employs generalized moment selection in order to conservatively determine which inequalities bind at a given θ. Implementation of GMS requires the use of a tuning parameter $\kappa_n = o(n^{1/2})$, which can be difficult to choose in practice. An interesting avenue for future research would combine the method proposed in this paper with the method proposed by Romano, Shaikh, and Wolf (2014) for the choice of κ_n.

[31] We thank Kei Hirano for suggesting this adaptation of our method.

[32] This case occurs in Gafarov and Montiel-Olea (2015), who study inference for maximum and minimum responses to impulses in structural vector autoregression models. Bounds on treatment effects frequently have this form as well: Demuynck (2015) rewrites numerous such bounds as values of a linear program with estimated direction p and varying, estimated constraints.

Appendix A Proof of Main Results

This Appendix is organized as follows. Section A.1 provides in Table A.0 a summary of the notation used throughout, and in Figure A.1 and Table A.1 a flow diagram and heuristic explanation of how each lemma contributes to the proof of Theorem 3.1. Section A.2 contains proofs of our two main results, Theorem 3.1 and Theorem 3.2. Section B contains the statements and proofs of the lemmas used to establish Theorem 3.1. Section C contains the statements and proofs of two auxiliary lemmas used in the main text (Sections 2.2 and 3.2). Throughout the Appendix we use the convention $\infty 0 = 0$.

A.1 Notation and Structure of Proofs

Table A.0: Important notation. Here $\rho > 0$ is fixed as described in Section 4, $(P_n, \theta_n) \in \{(P,\theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ is a subsequence as defined in (A.3)-(A.4) below, $\theta_n' \in \theta_n + \rho/\sqrt{n}B^d$, and $\lambda \in \mathbb{R}^d$.

$\mathbb{G}_{n,j}(\cdot)$	=	$\frac{\sqrt{n}(\bar{m}_{n,j}(\cdot) - E_P(m_j(X_i,\cdot)))}{\sigma_{P,j}(\cdot)},\ j = 1,\ldots,J$	Sample empirical process.
$\mathbb{G}^b_{n,j}(\cdot)$	=	$\frac{\sqrt{n}(\bar{m}^b_{n,j}(\cdot) - \bar{m}_{n,j}(\cdot))}{\hat{\sigma}_{n,j}(\cdot)},\ j = 1,\ldots,J$	Bootstrap empirical process.
$\eta_{n,j}(\cdot)$	=	$\frac{\sigma_{P,j}(\cdot)}{\hat{\sigma}_{n,j}(\cdot)} - 1,\ j = 1,\ldots,J$	Estimation error in sample moments' asymptotic standard deviation.
$D_{P,j}(\cdot)$	=	$\nabla_\theta \left(\frac{E_P(m_j(X_i,\cdot))}{\sigma_{P,j}(\cdot)}\right),\ j = 1,\ldots,J$	Gradient of population moments w.r.t. θ, with estimator $\hat{D}_{n,j}(\cdot)$.
$\gamma_{1,P_n,j}(\cdot)$	=	$\frac{E_{P_n}(m_j(X_i,\cdot))}{\sigma_{P_n,j}(\cdot)},\ j = 1,\ldots,J$	Studentized population moments.
$\pi_{1,j}$	=	$\lim_{n\to\infty} \kappa_n^{-1}\sqrt{n}\gamma_{1,P_n,j}(\theta_n')$	Limit of rescaled population moments, constant $\forall \theta_n' \in \theta_n + \frac{\rho}{\sqrt{n}}B^d$ by Lemma B.5.
$\pi^*_{1,j}$	=	$\begin{cases} 0, & \text{if } \pi_{1,j} = 0, \\ -\infty, & \text{if } \pi_{1,j} < 0. \end{cases}$	"Oracle" GMS.
$\hat{\xi}_{n,j}(\cdot)$	=	$\begin{cases} \kappa_n^{-1}\sqrt{n}\bar{m}_{n,j}(\cdot)/\hat{\sigma}_{n,j}(\cdot), & j = 1,\ldots,J_1 \\ 0, & j = J_1+1,\ldots,J \end{cases}$	Rescaled studentized sample moments, set to 0 for equalities.
$u_{n,j,\theta_n}(\lambda)$	=	$\{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\}(1 + \eta_{n,j}(\theta_n))$	Mean value expansion of nonlinear constraints with sample empirical process and "oracle" GMS, with $\bar{\theta}_n$ componentwise between θ_n and $\theta_n + \lambda/\sqrt{n}$.
$U_n(\theta_n, c)$	=	$\{\lambda \in \rho B^d : p'\lambda = 0 \cap u_{n,j,\theta_n}(\lambda) \le c,\ \forall j = 1,\ldots,J\}$	Feasible set for nonlinear sample problem intersected with $p'\lambda = 0$.
$v^b_{n,j,\theta_n'}(\lambda)$	=	$\mathbb{G}^b_{n,j}(\theta_n') + \hat{D}_{n,j}(\theta_n')\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta_n'))$	Linearized constraints with bootstrap empirical process and sample GMS.
$V^b_n(\theta_n', c)$	=	$\{\lambda \in \rho B^d : p'\lambda = 0 \cap v^b_{n,j,\theta_n'}(\lambda) \le c,\ \forall j = 1,\ldots,J\}$	Feasible set for linearized bootstrap problem with sample GMS and $p'\lambda = 0$.
$v_{n,j,\theta_n'}(\lambda)$	=	$\mathbb{G}^b_{n,j}(\theta_n') + \hat{D}_{n,j}(\theta_n')\lambda + \pi^*_{1,j}$	Linearized constraints with bootstrap empirical process and "oracle" GMS.
$V_n(\theta_n', c)$	=	$\{\lambda \in \rho B^d : p'\lambda = 0 \cap v_{n,j,\theta_n'}(\lambda) \le c,\ \forall j = 1,\ldots,J\}$	Feasible set for linearized bootstrap problem with "oracle" GMS and $p'\lambda = 0$.
$w_{j,\theta_n'}(\lambda)$	=	$\mathbb{G}_{P,j}(\theta_n') + D_{P,j}(\theta_n')\lambda + \pi^*_{1,j}$	Linearized constraints with limit Gaussian process and "oracle" GMS.
$W(\theta_n, c)$	=	$\{\lambda \in \rho B^d : p'\lambda = 0 \cap w_{j,\theta_n}(\lambda) \le c,\ \forall j = 1,\ldots,J\}$	Feasible set for linearized Gaussian problem with $p'\lambda = 0$.
$\hat{c}_n(\theta)$	=	$\inf\{c \in \mathbb{R}_+ : P^*(V^b_n(\theta, c) \neq \emptyset) \ge 1 - \alpha\}$	Bootstrap critical level.
$\hat{c}^*_n(\theta)$	=	$\inf_{\lambda \in \rho B^d} \hat{c}_n(\theta + \frac{\lambda}{\sqrt{n}})$	Smallest value of the bootstrap critical level in a $\frac{\rho}{\sqrt{n}}B^d$ neighborhood of θ.

Figure A.1: Structure of Lemmas used in the proof of Theorem 3.1.

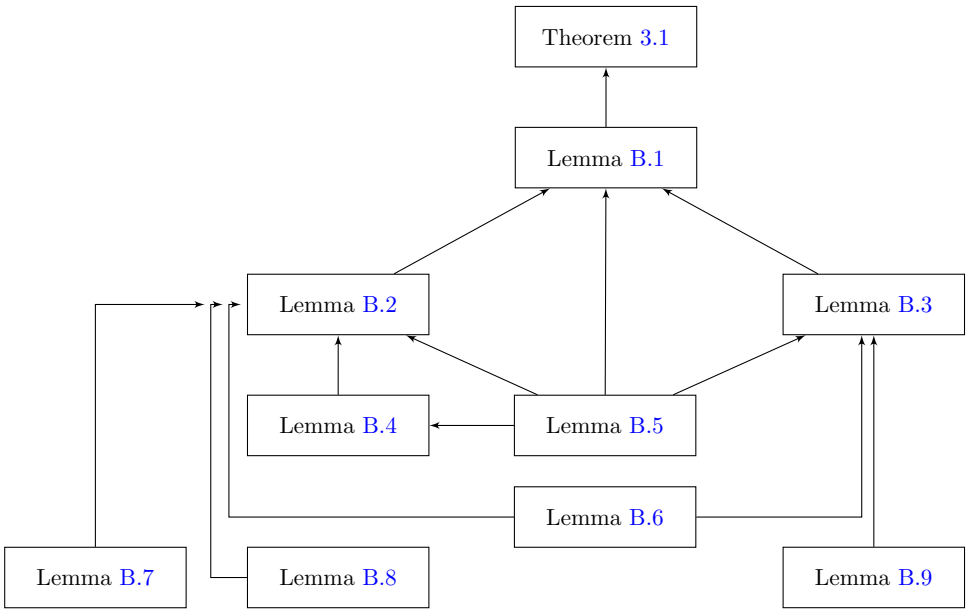

Table A.1: Heuristics for the role of each Lemma in the proof of Theorem 3.1. Notes: (i) Uniformity in Theorem 3.1 is enforced arguing along subsequences; (ii) When needed, random variables are realized on the same probability space as shown in Lemma B.1; (iii) Here $(P_n, \theta_n) \in \{(P, \theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ is a subsequence as defined in (A.3)-(A.4) below; (iv) All results hold for any $\theta_n' \in \theta_n + \rho/\sqrt{n}B^d$.

Theorem 3.1	$P_n(p'\theta_n \in CI) \geq P_n\left(U_n(\theta_n, \hat{c}_n^*(\theta_n)) \neq \emptyset\right) + o_{\mathcal{P}}(1)$. Coverage is conservatively estimated by the probability that U_n is nonempty.
Lemma B.1	$P_n\left(U_n(\theta_n, \hat{c}_n^*(\theta_n)) \neq \emptyset\right) \geq 1 - \alpha + o_{\mathcal{P}}(1)$.
Lemma B.2	$P_n(U(\theta_n, \hat{c}_n^*(\theta_n)) \neq \emptyset, V_n(\theta_n', \hat{c}_n(\theta_n')) = \emptyset) + P_n(U(\theta_n, \hat{c}_n^*(\theta_n)) = \emptyset, V(\theta_n', \hat{c}_n(\theta_n')) \neq \emptyset) = o_{\mathcal{P}}(1)$. Argued by comparing both U and V to their common limit W (after coupling).
Lemma B.3	$P_n(V_n(\theta_n', \hat{c}_n(\theta_n')) \neq \emptyset) \geq 1 - \alpha + o_{\mathcal{P}}(1)$. V_n differs from V_n^b by substituting "oracle" GMS (π_1^*) for sample GMS; any resulting distortion is conservative.
Lemma B.4	$\max\left\{\sup_{\lambda \in \rho B^d} \left\| \max_j (u_{n,j,\theta_n}(\lambda) - \hat{c}_n^*) - \max_j (w_{j,\theta_n'}(\lambda) - \hat{c}_n), \sup_{\lambda \in \rho B^d} \left\| \max_j w_{j,\theta_n'}(\lambda) - \max_j v_{n,j,\theta_n'}(\lambda) \right| \right\} = o_{\mathcal{P}}(1)$. The criterion functions entering U, V, and W, converge to each other.
Lemma B.5	Local-to-binding constraints are selected by GMS uniformly over the ρ-box (intuition: $\rho n^{-1/2} = o_{\mathcal{P}}(\kappa_n^{-1})$), and $\|\hat{\xi}_n(\theta_n')) - \kappa_n^{-1}\sqrt{n}\sigma_{P_n,j}^{-1}(\theta_n')E_{P_n}[m_j(X_i, \theta_n')]\| = o_{\mathcal{P}}(1)$.
Lemma B.6	$\forall \eta > 0\ \exists \delta > 0, N \in \mathbb{N} : P_n(\{W(\theta_n', c) \neq \emptyset\} \cap \{W^{-\delta}(\theta_n', c) = \emptyset\}) < \eta,\ \forall n \geq N$, and similarly for V_n. It is unlikely that these sets are nonempty but become empty upon slightly tightening stochastic constraints.
Lemma B.7	Intersections of constraints whose gradients are almost linearly dependent are unlikely to realize inside W. Hence, we can ignore irregularities that occur as linear dependence is approached.
Lemma B.8	If there are weakly more equality constraints than parameters, then c is uniformly bounded away from zero. This simplifies some arguments.
Lemma B.9	If two paired inequalities are local to binding, then they are also asymptotically identical up to sign. This justifies "merging" them.

A.2 Proof of Theorems 3.1 and 3.2

Proof of Theorem 3.1

Following Andrews and Guggenberger (2009), we index distributions by a vector of nuisance parameters relevant for the asymptotic size. For this, let $\gamma_P \equiv (\gamma_{1,P}, \gamma_{2,P}, \gamma_{3,P})$, where $\gamma_{1,P} = (\gamma_{1,P,1}, \cdots, \gamma_{1,P,J})$ with

$$\gamma_{1,P,j}(\theta) = \sigma_{P,j}^{-1}(\theta) E_P[m_j(X_i, \theta)],\ j = 1, \cdots, J, \tag{A.1}$$

$\gamma_{2,P} = (s(p, \Theta_I(P)), vech(\Omega_P(\theta)), vec(D_P(\theta)))$, and $\gamma_{3,P} = P$. We proceed in steps.

Step 1. Let $\{P_n, \theta_n\} \in \{(P, \theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ be a sequence such that

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(p'\theta \in CI_n) = \liminf_{n\to\infty} P_n(p'\theta_n \in CI_n), \tag{A.2}$$

with $CI_n = [-s(-p, \mathcal{C}_n(\hat{c}_n)), s(p, \mathcal{C}_n(\hat{c}_n))]$. We then let $\{l_n\}$ be a subsequence of $\{n\}$ such that

$$\liminf_{n\to\infty} P_n(p'\theta_n \in CI_n) = \lim_{n\to\infty} P_{l_n}(p'\theta_{l_n} \in CI_{l_n}). \tag{A.3}$$

Then there is a further subsequence $\{a_n\}$ of $\{l_n\}$ such that

$$\lim_{a_n\to\infty} \kappa_{a_n}^{-1} \sqrt{a_n} \sigma_{P_{a_n},j}^{-1}(\theta_{a_n}) E_{P_{a_n}}[m_j(X_i, \theta_{a_n})] = \pi_{1,j} \in \mathbb{R}_{[-\infty]},\ j = 1, \ldots, J. \tag{A.4}$$

To avoid multiple subscripts, with some abuse of notation we write (P_n, θ_n) to refer to (P_{a_n}, θ_{a_n}) throughout this Appendix. We let

$$\pi_{1,j}^* = \begin{cases} 0 & \text{if } \pi_{1,j} = 0, \\ -\infty & \text{if } \pi_{1,j} < 0. \end{cases} \tag{A.5}$$

The projection of θ_n is covered when

$$\begin{aligned}
&-s(-p, \mathcal{C}_n(\hat{c}_n)) \le p'\theta_n \le s(p, \mathcal{C}_n(\hat{c}_n)) \\
&\Leftrightarrow \left\{ \begin{array}{c} \inf p'\vartheta \\ s.t.\ \vartheta \in \Theta,\quad \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \le \hat{c}_n(\vartheta), \forall j \end{array} \right\} \le p'\theta_n \le \left\{ \begin{array}{c} \sup p'\vartheta \\ s.t.\ \vartheta \in \Theta,\quad \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \le \hat{c}_n(\vartheta), \forall j \end{array} \right\} \\
&\Leftrightarrow \left\{ \begin{array}{c} \inf_\lambda p'\lambda \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n),\quad \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta_n + \lambda/\sqrt{n})} \le \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\} \le 0 \\
&\qquad \le \left\{ \begin{array}{c} \sup_\lambda p'\lambda \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n),\quad \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta_n + \lambda/\sqrt{n})} \le \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\} \\
&\Leftrightarrow \left\{ \begin{array}{c} \inf_\lambda p'\lambda \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n), \\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1 + \eta_{n,j}(\theta_n)) \le \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\} \le 0 \\
&\qquad \le \left\{ \begin{array}{c} \sup_\lambda p'\lambda \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n), \\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1 + \eta_{n,j}(\theta_n)) \le \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\},
\end{aligned} \tag{A.6}$$

with $\eta_{n,j}(\cdot) \equiv \sigma_{P,j}(\cdot)/\hat{\sigma}_{n,j}(\cdot) - 1$ and where we took a mean value expansion yielding $\forall j$

$$\frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta_n + \lambda/\sqrt{n})} = \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1 + \eta_{n,j}(\theta_n)). \tag{A.7}$$

The event in (A.6) is implied by

$$\left\{ \begin{array}{c} \inf_\lambda p'\lambda \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n) \cap \rho B^d, \\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1 + \eta_{n,j}(\theta_n)) \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\} \leq 0$$

$$\leq \left\{ \begin{array}{c} \sup_\lambda p'\lambda \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n) \cap \rho B^d, \\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1 + \eta_{n,j}(\theta_n)) \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\}, \tag{A.8}$$

with $B^d = \{x \in \mathbb{R}^d : |x_i| \leq 1, i = 1, \ldots, d\}$.

Step 2. This step is used only when Assumption 3.3′ is invoked. For each $j = 1, \ldots, J_{11}$ such that

$$\pi^*_{1,j} = \pi^*_{1,j+J_{11}} = 0, \tag{A.9}$$

where π^*_1 is defined in (A.5), let

$$\tilde{\mu}_j = \left\{ \begin{array}{cc} 1 & \text{if } \gamma_{1,P_n,j}(\theta_n) = 0 = \gamma_{1,P_n,j+J_{11}}(\theta_n), \\ \frac{\gamma_{1,P_n,j+J_{11}}(\theta_n)(1+\eta_{n,j+J_{11}}(\theta_n))}{\gamma_{1,P_n,j+J_{11}}(\theta_n)(1+\eta_{n,j+J_{11}}(\theta_n))+\gamma_{1,P_n,j}(\theta_n)(1+\eta_{n,j}(\theta_n))} & \text{otherwise,} \end{array} \right. \tag{A.10}$$

$$\tilde{\mu}_{j+J_{11}} = \left\{ \begin{array}{cc} 0 & \text{if } \gamma_{1,P_n,j}(\theta_n) = 0 = \gamma_{1,P_n,j+J_{11}}(\theta_n), \\ \frac{\gamma_{1,P_n,j}(\theta_n)(1+\eta_{n,j}(\theta_n))}{\gamma_{1,P_n,j+J_{11}}(\theta_n)(1+\eta_{n,j+J_{11}}(\theta_n))+\gamma_{1,P_n,j}(\theta_n)(1+\eta_{n,j}(\theta_n))} & \text{otherwise,} \end{array} \right. \tag{A.11}$$

For each $j = 1, \ldots, J_{11}$, replace the constraint indexed by j, that is

$$\frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta_n + \lambda/\sqrt{n})} \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \tag{A.12}$$

with the following weighted sum of the paired inequalities

$$\tilde{\mu}_j \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta_n + \lambda/\sqrt{n})} - \tilde{\mu}_{j+J_{11}} \frac{\sqrt{n}\bar{m}_{j+J_{11},n}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j+J_{11}}(\theta_n + \lambda/\sqrt{n})} \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \tag{A.13}$$

and for each $j = 1, \ldots, J_{11}$, replace the constraint indexed by $j + J_{11}$, that is

$$\frac{\sqrt{n}\bar{m}_{j+J_{11},n}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j+J_{11}}(\theta_n + \lambda/\sqrt{n})} \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \tag{A.14}$$

with

$$-\tilde{\mu}_j \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta_n + \lambda/\sqrt{n})} + \tilde{\mu}_{j+J_{11}} \frac{\sqrt{n}\bar{m}_{j+J_{11},n}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j+J_{11}}(\theta_n + \lambda/\sqrt{n})} \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \tag{A.15}$$

It then follows from Assumption 3.3′ (iii-4) that these replacements are conservative with probability approaching one because

$$P_n\left(\frac{\bar{m}_{j+J_{11},n}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j+J_{11}}(\theta_n + \lambda/\sqrt{n})} \leq -\frac{\bar{m}_{n,j}(\theta_n + \lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta_n + \lambda/\sqrt{n})} \right) \to 1,$$

and therefore with probability approaching one, (A.13) implies (A.12) and (A.15) implies (A.14).

Step 3. Next, we make the following comparisons:

$$\pi^*_{1,j} = 0 \Rightarrow \pi^*_{1,j} \geq \sqrt{n}\gamma_{1,P_n,j}(\theta_n), \tag{A.16}$$

$$\pi^*_{1,j} = -\infty \Rightarrow \sqrt{n}\gamma_{1,P_n,j}(\theta_n) \to -\infty. \tag{A.17}$$

For any constraint j for which $\pi^*_{1,j} = 0$, (A.16) yields that replacing $\sqrt{n}\gamma_{1,P_n,j}(\theta_n)$ in (A.8) with $\pi^*_{1,j}$ introduces a conservative distortion. Under Assumption 3.3′, for any j such that (A.9) holds, the substitutions in (A.13) and (A.15) yield $\tilde{\mu}_j\sqrt{n}\gamma_{1,P_n,j}(\theta_n)(1+\eta_{n,j}(\theta_n)) - \tilde{\mu}_{j+J_{11}}\sqrt{n}\gamma_{1,P_n,j+J_{11}}(\theta_n)(1+\eta_{n,j+J_{11}}(\theta_n)) = 0$, and therefore replacing this term with $\pi^*_{1,j} = 0 = \pi^*_{1,j+J_{11}}$ is inconsequential.

For any j for which $\pi^*_{1,j} = -\infty$, (A.17) yields that for n large enough, $\sqrt{n}\gamma_{1,P_n,j}(\theta_n)$ can be replaced with $\pi^*_{1,j}$. To see this, note that by the Cauchy-Schwarz inequality, Assumption 3.4 (i)-(ii), and $\lambda \in \rho B^d$, it follows that

$$D_{P_n,j}(\bar{\theta}_n)\lambda \leq \rho\sqrt{d}(\|D_{P_n,j}(\bar{\theta}_n) - D_{P_n,j}(\theta_n)\| + \|D_{P_n,j}(\theta_n)\|) \leq \rho\sqrt{d}(\bar{M} + \rho M/\sqrt{n}), \tag{A.18}$$

where $\bar{M}$ and M are as defined in Assumption 3.4-(i) and (ii) respectively, and we used that $\bar{\theta}_n$ lies component-wise between θ_n and $\theta_n + \lambda/\sqrt{n}$. Using that $\mathbb{G}_{n,j}$ is asymptotically tight by Assumption 3.5, we have that for any $\tau > 0$, there exists a $T > 0$ and $N_1 \in \mathbb{N}$ such that

$$\begin{aligned}
&P_n\left(\max_{j:\pi^*_{1,j}=-\infty}\{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n)) \leq 0,\ \forall\lambda \in \rho B^d\right)\\
&\geq P_n\left(\max_{j:\pi^*_{1,j}=-\infty}\{T + \rho(\bar{M} + \rho M/\sqrt{n}) + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n)) \leq 0 \cap \max_{j:\pi^*_{1,j}=-\infty}\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) \leq T\right)\\
&= P_n\left(\max_{j:\pi^*_{1,j}=-\infty}\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) \leq T\right) > 1 - \tau/2,\ \forall n \geq N_1.
\end{aligned} \tag{A.19}$$

We therefore have that for $n \geq N_1$,

$$\begin{aligned}
&P_n\left(\left\{\begin{array}{c}\inf_\lambda p'\lambda\\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n) \cap \rho B^d,\\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n)) \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j\end{array}\right\} \leq 0\right.\\
&\qquad\left.\leq \left\{\begin{array}{c}\sup_\lambda p'\lambda\\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n) \cap \rho B^d,\\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n)) \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j\end{array}\right\}\right)
\end{aligned} \tag{A.20}$$

$$\begin{aligned}
&\geq P_n\left(\left\{\begin{array}{c}\inf_\lambda p'\lambda\\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n) \cap \rho B^d,\\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\}(1+\eta_{n,j}(\theta_n)) \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j\end{array}\right\} \leq 0\right.\\
&\qquad\left.\leq \left\{\begin{array}{c}\sup_\lambda p'\lambda\\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n) \cap \rho B^d,\\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\}(1+\eta_{n,j}(\theta_n)) \leq \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j\end{array}\right\}\right) - \tau/2.
\end{aligned} \tag{A.21}$$

Since the choice of τ is arbitrary, the limit of the term in (A.20) is not smaller than the limit of the first term in (A.21). Hence, we continue arguing for the event whose probability is evaluated in (A.21).

Finally, by definition $\hat{c}_n(\theta) \geq 0$ and therefore $\inf_{\lambda\in\rho B^d}\hat{c}_n(\theta_n + \lambda/\sqrt{n})$ exists. Therefore, the event

whose probability is evaluated in (A.21) is implied by the event

$$\left\{ \begin{array}{c} \inf_\lambda p'\lambda \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n) \cap \rho B^d, \\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\}(1 + \eta_{n,j}(\theta_n)) \leq \inf_{\lambda \in \rho B^d} \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\} \leq 0$$

$$\leq \left\{ \begin{array}{c} \sup_\lambda p'\lambda \\ s.t. \lambda \in \sqrt{n}(\Theta - \theta_n) \cap \rho B^d, \\ \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\}(1 + \eta_{n,j}(\theta_n)) \leq \inf_{\lambda \in \rho B^d} \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\} \tag{A.22}$$

For each $\lambda \in \mathbb{R}^d$, define

$$u_{n,j,\theta_n}(\lambda) \equiv \{\mathbb{G}_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\}(1 + \eta_{n,j}(\theta_n)), \tag{A.23}$$

where under Assumption 3.3′ when $\pi^*_{1,j} = 0$ and $\pi^*_{1,j+J_{11}} = 0$ the substitutions of equation (A.12) with equation (A.13) and of equation (A.14) with equation (A.15) have been performed. Let

$$U_n(\theta_n, c) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap u_{n,j,\theta_n}(\lambda) \leq c, \ \forall j = 1, \ldots, J\right\}, \tag{A.24}$$

and define

$$\hat{c}^*_n(\theta) \equiv \inf_{\lambda \in \rho B^d} \hat{c}_n(\theta + \lambda/\sqrt{n}). \tag{A.25}$$

Then by (A.22) and the definition of U_n, we obtain

$$P_n(p'\theta_n \in CI_n) \geq P_n\left(U_n(\theta_n, \hat{c}^*_n(\theta_n)) \neq \emptyset\right), \tag{A.26}$$

because whenever $U_n(\theta_n, \hat{c}^*_n(\theta_n)) \neq \emptyset$, the event in (A.22) attains. By Lemma B.1,

$$\lim_{n\to\infty} P_n\left(U_n(\theta_n, \hat{c}^*_n(\theta_n)) \neq \emptyset\right) \geq 1 - \alpha. \tag{A.27}$$

The conclusion of the theorem then follows from (A.2), (A.3), (A.26), and (A.27). □

Proof of Theorem 3.2

For given θ, the event

$$\max_{j=1,\ldots,J} \left\{\mathbb{G}^b_{n,j}(\theta) + \varphi_j(\hat{\xi}_{n,j}(\theta))\right\} \leq c \tag{A.28}$$

implies the event

$$\max_{\lambda \in \Lambda^b_n(\theta,\rho,c)} p'\lambda \geq 0 \geq \min_{\lambda \in \Lambda^b_n(\theta,\rho,c)} p'\lambda, \tag{A.29}$$

with Λ^b_n defined in (2.7). This is so because if $\max_{j=1,\ldots,J}\left\{\mathbb{G}^b_{n,j}(\theta) + \varphi_j(\hat{\xi}_{n,j}(\theta))\right\} \leq c$, $\lambda = 0$ is feasible in both optimization problems in (A.29), hence the event in (A.29) is implied. In turn this yields that for each $n \in \mathbb{N}$ and $\theta \in \Theta$,

$$c^{AS}_n(\theta) \geq \hat{c}_n(\theta), \tag{A.30}$$

and therefore the result follows. □

Appendix B Main Lemmas

Throughout this Appendix, we maintain Assumptions 3.1, 3.2, 3.3 or 3.3′, 3.4, and 3.5. We let $(P_n, \theta_n) \in \{(P, \theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ be a subsequence as defined in (A.3)-(A.4).

Fix $\rho > 0$ as discussed in Section 4 and $c \geq 0$. For each $\lambda \in \mathbb{R}^d$ and $\theta \in \theta_n + \frac{\rho}{\sqrt{n}} B^d$, let

$$v_{n,j,\theta}(\lambda) \equiv \mathbb{G}^b_{n,j}(\theta) + \hat{D}_{n,j}(\theta)\lambda + \pi^*_{1,j}, \tag{B.1}$$

$$w_{j,\theta}(\lambda) \equiv \mathbb{G}_{P,j}(\theta) + D_{P,j}(\theta)\lambda + \pi^*_{1,j}, \tag{B.2}$$

where $\pi^*_{1,j}$ is defined in (A.5) and we used Lemma B.5. Under Assumption 3.3′ if

$$\pi^*_{1,j} = 0 = \pi^*_{1,j+J_{11}}, \tag{B.3}$$

we replace the constraints

$$\mathbb{G}^b_{n,j}(\theta) + \hat{D}_{n,j}(\theta)\lambda \leq c, \tag{B.4}$$

$$\mathbb{G}_{P,j}(\theta) + D_{P,j}(\theta)\lambda \leq c, \tag{B.5}$$

$$\mathbb{G}^b_{n,j+J_{11}}(\theta) + \hat{D}_{n,j+J_{11}}(\theta)\lambda \leq c, \tag{B.6}$$

$$\mathbb{G}_{P,j+J_{11}}(\theta) + D_{P,j+J_{11}}(\theta)\lambda \leq c, \tag{B.7}$$

respectively with

$$\mu_j\{\mathbb{G}^b_{n,j}(\theta) + \hat{D}_{n,j}(\theta)\lambda\} - \mu_{j+J_{11}}\{\mathbb{G}^b_{n,j+J_{11}}(\theta) + \hat{D}_{n,j+J_{11}}(\theta)\lambda\} \leq c, \tag{B.8}$$

$$\mu_j\{\mathbb{G}_{P,j}(\theta) + D_{P,j}(\theta)\lambda\} - \mu_{j+J_{11}}\{\mathbb{G}_{P,j+J_{11}}(\theta) + D_{P,j+J_{11}}(\theta)\lambda\} \leq c, \tag{B.9}$$

$$-\mu_j\{\mathbb{G}^b_{n,j}(\theta) + \hat{D}_{n,j}(\theta)\lambda\} + \mu_{j+J_{11}}\{\mathbb{G}^b_{n,j+J_{11}}(\theta) + \hat{D}_{n,j+J_{11}}(\theta)\lambda\} \leq c, \tag{B.10}$$

$$-\mu_j\{\mathbb{G}_{P,j}(\theta) + D_{P,j}(\theta)\lambda\} + \mu_{j+J_{11}}\{\mathbb{G}_{P,j+J_{11}}(\theta) + D_{P,j+J_{11}}(\theta)\lambda\} \leq c. \tag{B.11}$$

where

$$\mu_j = \begin{cases} 1 & \text{if } \gamma_{1,P_n,j}(\theta) = 0 = \gamma_{1,P_n,j+J_{11}}(\theta), \\ \frac{\gamma_{1,P_n,j+J_{11}}(\theta)}{\gamma_{1,P_n,j+J_{11}}(\theta)+\gamma_{1,P_n,j}(\theta)} & \text{otherwise,} \end{cases} \tag{B.12}$$

$$\mu_{j+J_{11}} = \begin{cases} 0 & \text{if } \gamma_{1,P_n,j}(\theta) = 0 = \gamma_{1,P_n,j+J_{11}}(\theta), \\ \frac{\gamma_{1,P_n,j}(\theta)}{\gamma_{1,P_n,j+J_{11}}(\theta)+\gamma_{1,P_n,j}(\theta)} & \text{otherwise,} \end{cases} \tag{B.13}$$

Let the level sets associated with the so defined functions $v_{n,j,\theta}(\lambda)$ and $w_{j,\theta}(\lambda)$ be

$$V_n(\theta, c) \equiv \big\{\lambda \in \rho B^d : p'\lambda = 0 \cap v_{n,j,\theta}(\lambda) \leq c,\ \forall j = 1, \ldots, J\big\}, \tag{B.14}$$

$$W(\theta, c) \equiv \big\{\lambda \in \rho B^d : p'\lambda = 0 \cap w_{j,\theta}(\lambda) \leq c,\ \forall j = 1, \ldots, J\big\}. \tag{B.15}$$

Due to the substitutions in equations (B.8)-(B.11), the paired inequalities (i.e., inequalities for which (B.3) holds under Assumption 3.3′) are now genuine equalities relaxed by c. With some abuse of notation, we index them among the $j = J_1 + 1, \ldots, J$. With that convention, for given $\delta \in \mathbb{R}$, define

$$V^\delta_n(\theta, c) \equiv \big\{\lambda \in \rho B^d : p'\lambda = 0 \cap v_{n,j,\theta}(\lambda) \leq c + \delta,\ \forall j = 1, \ldots, J_1, \\ \cap\, v_{n,j,\theta}(\lambda) \leq c,\ \forall j = J_1 + 1, \ldots, J\big\}. \tag{B.16}$$

and

$$W^{\delta}(\theta, c) \equiv \big\{\lambda \in \rho B^d : p'\lambda = 0 \cap w_{j,\theta}(\lambda) \leq c + \delta,\ \forall j = 1, \ldots, J_1, \\ \cap\, w_{j,\theta}(\lambda) \leq c,\ \forall j = J_1 + 1, \ldots, J\big\}. \tag{B.17}$$

Define the $(J + 2d + 2) \times d$ matrix

$$K_P(\theta) \equiv \begin{bmatrix} [D_{P,j}(\theta)]_{j=1}^{J_1+J_2} \\ [-D_{P,j-J_1}(\theta)]_{j=J_1+J_2+1}^{J} \\ I_d \\ -I_d \\ p' \\ -p' \end{bmatrix}. \tag{B.18}$$

Given a square matrix A, we let $\text{eig}(A)$ denote its smallest eigenvalue. In all Lemmas below, α is assumed less than $1/2$.

LEMMA B.1: *For each $\theta \in \Theta_I(P_n)$, let $\hat{c}_n^*(\theta) \equiv \inf_{\lambda \in \rho B^d} \hat{c}_n(\theta + \lambda/\sqrt{n})$. Then*

$$\lim_{n\to\infty} P_n\left(U_n(\theta_n, \hat{c}_n^*(\theta_n)) \neq \emptyset\right) \geq 1 - \alpha. \tag{B.19}$$

Proof. For any $\epsilon > 0$, there exists $\lambda^\epsilon \in \rho B^d$ such that

$$\hat{c}_n(\theta_n + \lambda^\epsilon/\sqrt{n}) \leq \inf_{\lambda \in \rho B^d} \hat{c}_n(\theta_n + \lambda/\sqrt{n}) + \epsilon. \tag{B.20}$$

In what follows, let

$$\theta_n^\epsilon \equiv \theta_n + \lambda^\epsilon/\sqrt{n} \tag{B.21}$$

denote the value at which $\hat{c}_n$ is evaluated in equation (B.20).

By simple addition and subtraction,

$$\begin{aligned} P_n\Big(U_n(\theta_n, \hat{c}_n^*(\theta_n)) \neq \emptyset\Big) &= P_n^*\Big(V_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\Big) \\ &+ \left[P_n\Big(U_n(\theta_n, \hat{c}_n^*(\theta_n)) \neq \emptyset\Big) - P_n\Big(W_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\Big)\right] \\ &+ \left[P_n\Big(W_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\Big) - P_n^*\Big(V_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\Big)\right]. \end{aligned} \tag{B.22}$$

By passing to a further subsequence $\{n\}$, we may assume that

$$D_{P_n}(\theta_n) \to D \tag{B.23}$$

for some $J \times d$ matrix D such that $\|D\| \leq M$.

By Lemma D.1 in Bugni, Canay, and Shi (2015), $\mathbb{G}_n \xrightarrow{d} \mathbb{G}_P$ in $l^\infty(\Theta)$ uniformly in $\mathcal{P}$. Using the same argument as in the proof of Theorem 3.2 with all moments binding, one can show that for any sequence $\{\theta_n\} \subset \Theta$, $\hat{c}_n(\theta_n)$ and $\hat{c}_n^*(\theta_n)$ are asymptotically bounded by the $(1 - \alpha/J)$ quantile of the standard Normal distribution, and hence are asymptotically tight. Therefore, the sequence $\{(\mathbb{G}_n, \hat{c}_n(\theta_n^\epsilon), \hat{c}_n^*(\theta_n))\}$ is asymptotically tight. By Prohorov's theorem and passing to a further sub-

sequence, we may then assume that

$$(\mathbb{G}_n, \hat{c}_n(\theta_n^\epsilon), \hat{c}_n^*(\theta_n)) \xrightarrow{d} (\mathbb{G}, \hat{c}, \hat{c}^*). \tag{B.24}$$

for some tight Borel random element $(\mathbb{G}, \hat{c}, \hat{c}^*) \in \ell^\infty(\Theta) \times \mathbb{R}_+ \times \mathbb{R}_+$. Moreover, by Assumption 3.3-(ii) (or Assumption 3.3′-(ii)) $\sup_{\theta\in\Theta} \|\eta_n(\theta)\| \xrightarrow{p} 0$ uniformly in $\mathcal{P}$, so that

$$(\mathbb{G}_n, \eta_n(\theta_n), \hat{c}_n(\theta_n^\epsilon), \hat{c}_n^*(\theta_n)) \xrightarrow{d} (\mathbb{G}, 0, \hat{c}, \hat{c}^*). \tag{B.25}$$

In what follows, using Lemma 1.10.4 in van der Vaart and Wellner (2000) we take $(\mathbb{G}_n^*, \eta_n^*, c_n, c_n^*)$ to be the almost sure representation of $(\mathbb{G}_n, \eta_n(\theta_n), \hat{c}_n(\theta_n^\epsilon), \hat{c}_n^*(\theta_n))$ defined on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that $(\mathbb{G}_n^*, \eta_n^*, c_n, c_n^*) \xrightarrow{a.s.} (\mathbb{G}^*, 0, c, c^*)$, where $(\mathbb{G}^*, c, c^*) \overset{d}{=} (\mathbb{G}, \hat{c}, \hat{c}^*)$.

Similarly, again by Lemma D.2.8 in Bugni, Canay, and Shi (2015), $\mathbb{G}_n^b \xrightarrow{d} \mathbb{G}_P$ in $l^\infty(\Theta)$ uniformly in $\mathcal{P}$ conditional on $\{X_1, \cdots, X_n\}$, and by Assumption 3.4 $|\hat{D}_n(\theta_n^\epsilon) - D_{P_n}(\theta_n)\| \xrightarrow{p} 0$. Hence, by (B.23) and (B.24),

$$(\mathbb{G}_n^b, \hat{D}_n(\theta_n^\epsilon), \hat{c}_n(\theta_n^\epsilon), \hat{c}_n^*(\theta_n)) \xrightarrow{d} (\mathbb{G}, D, \hat{c}, \hat{c}^*). \tag{B.26}$$

We take $(\tilde{\mathbb{G}}_n^{b,*}, \tilde{D}_n^*, \tilde{c}_n, \tilde{c}_n^*)$ to be the almost sure representation of $(\mathbb{G}_n^b,\ \hat{D}_n(\theta_n^\epsilon),\ \hat{c}_n(\theta_n^\epsilon),\ \hat{c}_n^*(\theta_n))$ defined on another probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbf{P}})$ such that $(\tilde{\mathbb{G}}_n^{b,*},\ \tilde{D}_n^*,\ \tilde{c}_n,\ \tilde{c}_n^*) \xrightarrow{a.s.} (\tilde{\mathbb{G}}^*,\ D,\ \tilde{c},\ \tilde{c}^*)$, where $(\tilde{\mathbb{G}}^*, \tilde{c}, \tilde{c}^*) \overset{d}{=} (\mathbb{G}, \hat{c}, \hat{c}^*)$.

For each $\lambda \in \mathbb{R}^d$, we define analogs to the quantities in (A.23), (B.1) and (B.2) as

$$u_{n,j,\theta_n}^*(\lambda) \equiv \{\mathbb{G}_{n,j}^*(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \pi_{1,j}^*\}(1 + \eta_{n,j}^*(\theta_n)), \tag{B.27}$$

$$v_{n,j,\theta_n^\epsilon}^*(\lambda) \equiv \tilde{\mathbb{G}}_{n,j}^{b,*}(\theta_n^\epsilon) + \tilde{D}_{n,j}^*\lambda + \pi_{1,j}^*, \tag{B.28}$$

$$w_{j,\theta_n^\epsilon}^*(\lambda) \equiv \mathbb{G}_j^*(\theta_n^\epsilon) + D_{P_n,j}(\theta_n^\epsilon)\lambda + \pi_{1,j}^*, \tag{B.29}$$

$$\tilde{w}_{j,\theta_n^\epsilon}^*(\lambda) \equiv \tilde{\mathbb{G}}_j^*(\theta_n^\epsilon) + D_{P_n,j}(\theta_n^\epsilon)\lambda + \pi_{1,j}^*, \tag{B.30}$$

where we used that by Lemma B.5 $\kappa_n^{-1}\sqrt{n}\gamma_{1,P,j}(\theta_n) - \kappa_n^{-1}\sqrt{n}\gamma_{1,P,j}(\theta_n') = o(1)$ uniformly over $\theta_n' \in \theta_n + \rho/\sqrt{n}B^d$ and therefore $\pi_{1,j}^*$ is constant over this neighborhood, and we applied a similar replacement as described in equations (B.4)-(B.11) for the case that $\pi_{1,j}^* = 0 = \pi_{1,j+J_{11}}^*$. Similarly, we define analogs to the sets in (A.24), (B.14) and (B.15) as

$$U_n^*(\theta_n, c_n^*) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap u_{n,j,\theta_n}^*(\lambda) \le c_n^*,\ \forall j = 1, \ldots, J\right\}, \tag{B.31}$$

$$V_n^*(\theta_n^\epsilon, \tilde{c}_n) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap v_{n,j,\theta_n^\epsilon}^*(\lambda) \le \tilde{c}_n,\ \forall j = 1, \ldots, J\right\}, \tag{B.32}$$

$$W^*(\theta_n^\epsilon, c_n) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap w_{j,\theta_n^\epsilon}^*(\lambda) \le c_n,\ \forall j = 1, \ldots, J\right\}, \tag{B.33}$$

$$\tilde{W}^*(\theta_n^\epsilon, \tilde{c}_n) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap \tilde{w}_{j,\theta_n^\epsilon}^*(\lambda) \le \tilde{c}_n,\ \forall j = 1, \ldots, J\right\}. \tag{B.34}$$

It then follows that equation (B.22) can be rewritten as

$$\begin{aligned} P_n\Big(U_n(\theta_n, \hat{c}_n^*(\theta_n)) \neq \emptyset\Big) = \tilde{\mathbf{P}}\Big(V_n^*(\theta_n^\epsilon, \tilde{c}_n) \neq \emptyset\Big) + \left[\mathbf{P}\Big(U_n^*(\theta_n, c_n^*) \neq \emptyset\Big) - \mathbf{P}\Big(W^*(\theta_n^\epsilon, c_n) \neq \emptyset\Big)\right] \\ + \left[\tilde{\mathbf{P}}\Big(\tilde{W}^*(\theta_n^\epsilon, \tilde{c}_n) \neq \emptyset\}\Big) - \tilde{\mathbf{P}}\Big(V_n^*(\theta_n^\epsilon, \tilde{c}_n) \neq \emptyset\Big)\right]. \end{aligned} \tag{B.35}$$

By the Skorokhod representation and by Lemma B.3,

$$\lim_{n\to\infty} \tilde{\mathbf{P}}\left(V_n^*(\theta_n^\epsilon, \tilde{c}_n) \neq \emptyset\right) = \lim_{n\to\infty} P_n^*\left(\{V_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\right) \geq 1-\alpha. \tag{B.36}$$

We are left to show that the two terms in square brackets in (B.35) converge to zero as $n \to \infty$. Define

$$\mathcal{J}^* \equiv \{j = 1, \cdots, J : \pi_{1,j}^* = 0\}. \tag{B.37}$$

Case 1. Suppose first that $\mathcal{J}^* = \emptyset$, which implies $J_2 = 0$ and $\pi_{1,j}^* = -\infty$ for all j. Then we have

$$U_n^*(\theta_n, c_n^*) = W^*(\theta_n^\epsilon, c_n) = \tilde{W}^*(\theta_n, \tilde{c}_n^*) = V_n^*(\theta_n^\epsilon, \tilde{c}_n) = \{\lambda \in \rho B^d : p'\lambda = 0\}, \tag{B.38}$$

with probability 1, and hence

$$\mathbf{P}\Big(\{U_n^*(\theta_n, c_n^*) \neq \emptyset\} \cap \{W^*(\theta_n^\epsilon, c_n) \neq \emptyset\}\Big) = 1. \tag{B.39}$$

This in turn implies that

$$\left|\mathbf{P}\Big(U_n^*(\theta_n, c_n^*) \neq \emptyset\}\Big) - \mathbf{P}\Big(W^*(\theta_n^\epsilon, c_n) \neq \emptyset\}\Big)\right| = 0, \tag{B.40}$$

where we used $|\mathbf{P}(A) - \mathbf{P}(B)| \leq \mathbf{P}(A \Delta B) \leq 1 - \mathbf{P}(A \cap B)$ for any pair of events A and B. Hence, the first term in square brackets in (B.35) is 0.

We now turn to the second term in square brackets in (B.35). By (B.38), the same argument yielding to (B.39) applies, now for the sets $\tilde{W}^*(\theta_n, c_n^*)$ and $V_n^*(\theta_n^\epsilon, \tilde{c}_n)$, yielding

$$\left|\tilde{\mathbf{P}}\Big(\tilde{W}^*(\theta_n, c_n^*) \neq \emptyset\Big) - \tilde{\mathbf{P}}\Big(V_n^*(\theta_n^\epsilon, \tilde{c}_n) \neq \emptyset\Big)\right| = 0. \tag{B.41}$$

Hence, the second term in square brackets in (B.35) is also 0. The claim of the Lemma then follows by (B.36).

Case 2. Now consider the case that $\mathcal{J}^* \neq \emptyset$. We show that the terms in square brackets in (B.35) converge to 0. To that end, note that for any events A, B,

$$\Big|\mathbf{P}(A \neq \emptyset) - \mathbf{P}(B \neq \emptyset)\Big| \leq \Big|\mathbf{P}(\{A = \emptyset\} \cap \{B \neq \emptyset\}) + \mathbf{P}(\{A \neq \emptyset\} \cap \{B = \emptyset\})\Big| \tag{B.42}$$

Hence, we aim to establish that for $A = U_n^*(\theta_n, c_n^*)$, $B = W^*(\theta_n^\epsilon, c_n)$, and for $A = \tilde{W}^*(\theta_n^\epsilon, \tilde{c}_n)$, $B = V_n^*(\theta_n^\epsilon, \tilde{c}_n)$, the right hand side of equation (B.42) converges to zero. But this is guaranteed by Lemma B.2. □

LEMMA B.2: *Let (P_n, θ_n) have the almost sure representations given in Lemma B.1, and let $\mathcal{J}^*$ be defined as in (B.37). Assume that $\mathcal{J}^* \neq \emptyset$. Let θ_n^ϵ be as defined in (B.21). Then for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that*

$$\tilde{\mathbf{P}}\Big(\{\tilde{W}^*(\theta_n^\epsilon, \tilde{c}_n) \neq \emptyset\} \cap \{V_n^*(\theta_n^\epsilon, \tilde{c}_n) = \emptyset\}\Big) \leq \eta/2, \tag{B.43}$$

$$\tilde{\mathbf{P}}\Big(\{\tilde{W}^*(\theta_n^\epsilon, \tilde{c}_n) = \emptyset\} \cap \{V_n^*(\theta_n^\epsilon, \tilde{c}_n) \neq \emptyset\}\Big) \leq \eta/2, \tag{B.44}$$

$$\mathbf{P}\Big(\{U_n^*(\theta_n, c_n^*) \neq \emptyset\} \cap \{W^*(\theta_n^\epsilon, c_n) = \emptyset\}\Big) \leq \eta/2, \tag{B.45}$$

$$\mathbf{P}\Big(\{U_n^*(\theta_n, c_n^*) = \emptyset\} \cap \{W^*(\theta_n^\epsilon, c_n) \neq \emptyset\}\Big) \leq \eta/2, \tag{B.46}$$

for all $n \geq N$, where the sets in the above expressions are defined in equations (B.31), (B.32), (B.33),

and (B.34)

Proof. We begin by observing that for $j \notin \mathcal{J}^*$, $\pi^*_{1,j} = -\infty$, and therefore the corresponding inequalities

$$\begin{aligned}(\mathbb{G}^*_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j})(1 + \eta^*_{n,j}(\theta_n)) &\leq c^*_n,\\ \tilde{\mathbb{G}}^{b,*}_{n,j}(\theta^\epsilon_n) + \tilde{D}^*_{n,j}\lambda + \pi^*_{1,j} &\leq \tilde{c}_n,\\ \mathbb{G}^*_j(\theta^\epsilon_n) + D_{P_n,j}(\theta^\epsilon_n)\lambda + \pi^*_{1,j} &\leq c_n,\\ \tilde{\mathbb{G}}^*_j(\theta^\epsilon_n) + D_{P_n,j}(\theta^\epsilon_n)\lambda + \pi^*_{1,j} &\leq \tilde{c}_n,\end{aligned}$$

are satisfied with probability approaching one by similar arguments as in (A.19). Hence, we can redefine the sets of interest as

$$U^*_n(\theta_n, c^*_n) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap u^*_{n,j,\theta_n}(\lambda) \leq c^*_n,\ \forall j \in \mathcal{J}^*\right\}, \tag{B.47}$$

$$V^*_n(\theta^\epsilon_n, \tilde{c}_n) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap v^*_{n,j,\theta^\epsilon_n}(\lambda) \leq \tilde{c}_n,\ \forall j \in \mathcal{J}^*\right\}, \tag{B.48}$$

$$W^*(\theta^\epsilon_n, c_n) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap w^*_{j,\theta^\epsilon_n}(\lambda) \leq c_n,\ \forall j \in \mathcal{J}^*\right\}, \tag{B.49}$$

$$\tilde{W}^*(\theta^\epsilon_n, \tilde{c}_n) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap \tilde{w}^*_{j,\theta^\epsilon_n}(\lambda) \leq \tilde{c}_n,\ \forall j \in \mathcal{J}^*\right\}. \tag{B.50}$$

We first show (B.43). For this, we start by defining the events

$$A_n \equiv \left\{\sup_{\lambda \in \rho B^d} \max_{j \in \mathcal{J}^*} \left|(v^*_{n,j,\theta^\epsilon_n}(\lambda) - \tilde{w}^*_{j,\theta^\epsilon_n}(\lambda)\right| \geq \delta\right\}, \tag{B.51}$$

with $v^*_{n,j,\theta^\epsilon_n}(\lambda)$ and $\tilde{w}^*_{j,\theta^\epsilon_n}(\lambda)$ as defined in equations (B.28) and (B.30), respectively. Then by Lemma B.4, using the assumption that $\mathcal{J}^* \neq \emptyset$, for any $\eta > 0$ there exists $N' \in \mathbb{N}$ such that

$$\tilde{\mathbf{P}}(A_n) < \eta/2,\ \forall n \geq N'. \tag{B.52}$$

Define the sets of λs, $\tilde{W}^{*,+\delta}$ and $V^{*,+\delta}_n$ by relaxing the constraints shaping $\tilde{W}^*$ and V^*_n by δ:

$$V^{*,+\delta}_n(\theta^\epsilon_n, c) \equiv \{\lambda \in \rho B^d : p'\lambda = 0 \cap v^*_{n,j,\theta^\epsilon_n}(\lambda) \leq c + \delta,\ j \in \mathcal{J}^*\}, \tag{B.53}$$

$$\tilde{W}^{*,+\delta}(\theta^\epsilon_n, c) \equiv \{\lambda \in \rho B^d : p'\lambda = 0 \cap \tilde{w}^*_{j,\theta^\epsilon_n}(\lambda) \leq c + \delta,\ j \in \mathcal{J}^*\}. \tag{B.54}$$

Compared to the sets in equations (B.16) and (B.17), here we replace $v^*_{n,j,\theta^\epsilon_n}(\lambda)$ for $v_{n,j,\theta^\epsilon_n}(\lambda)$ and $\tilde{w}^*_{j,\theta^\epsilon_n}(\lambda)$ for $w_{j,\theta^\epsilon_n}(\lambda)$, we retain only constraints in $\mathcal{J}^*$, and we relax all such constraints by $\delta > 0$ instead of relaxing only those in $\{1, \ldots, J_1\}$. Next, define the event $L_n \equiv \{\tilde{W}^*(\theta^\epsilon_n, \tilde{c}_n) \subset V^{*,+\delta}_n(\theta^\epsilon_n, \tilde{c}_n)\}$ and note that $A^c_n \subseteq L_n$.

We may then bound the left hand side of (B.43) as

$$\begin{aligned}\tilde{\mathbf{P}}\Big(\{\tilde{W}^*(\theta^\epsilon_n, \tilde{c}_n) \neq \emptyset\} \cap \{V^*_n(\theta^\epsilon_n, \tilde{c}_n) = \emptyset\}\Big) &\leq \tilde{\mathbf{P}}\Big(\{\tilde{W}^*(\theta^\epsilon_n, \tilde{c}_n) \neq \emptyset\} \cap \{V^{*,+\delta}_n(\theta^\epsilon_n, \tilde{c}_n) = \emptyset\}\Big)\\ &\quad + \tilde{\mathbf{P}}\Big(\{V^{*,+\delta}_n(\theta^\epsilon_n, \tilde{c}_n) \neq \emptyset\} \cap \{V^*_n(\theta^\epsilon_n, \tilde{c}_n) = \emptyset\}\Big),\end{aligned} \tag{B.55}$$

where we used $P(A \cap B) \leq P(A \cap C) + P(B \cap C^c)$ for any events A, B, and C. The first term on the right hand side of (B.55) can further be bounded as

$$\begin{aligned}\tilde{\mathbf{P}}\Big(\{\tilde{W}^*(\theta^\epsilon_n, \tilde{c}_n) \neq \emptyset\} \cap \{V^{*,+\delta}_n(\theta^\epsilon_n, \tilde{c}_n) = \emptyset\}\Big) &\leq \tilde{\mathbf{P}}\Big(\{\tilde{W}^*(\theta^\epsilon_n, \tilde{c}_n) \not\subseteq V^{*,+\delta}_n(\theta^\epsilon_n, \tilde{c}_n)\}\Big)\\ &= \tilde{\mathbf{P}}(L^c_n) \leq \tilde{\mathbf{P}}(A_n) < \eta/2,\ \forall n \geq N',\end{aligned} \tag{B.56}$$

where the penultimate inequality follows from $A^c_n \subseteq L_n$ as argued above, and the last inequality

follows from (B.52). For the second term on the left hand side of (B.55), by Lemma B.6, there exists $N'' \in \mathbb{N}$ such that

$$\tilde{\mathbf{P}}\Big(\{V_n^{*,+\delta}(\theta_n^\epsilon, \tilde{c}_n) \neq \emptyset\} \cap \{V_n^*(\theta_n^\epsilon, \tilde{c}_n) = \emptyset\}\Big) \leq \eta/2, \ \forall n \geq N''. \tag{B.57}$$

Hence, (B.43) follows from (B.55), (B.56), and (B.57). The result in (B.44) follows similarly.

To establish (B.45), define

$$B_n \equiv \left\{ \sup_{\lambda \in \rho B^d} \max_{j \in \mathcal{J}^*} \left| (w^*_{j,\theta_n^\epsilon}(\lambda) - c_n) - (u^*_{n,j,\theta_n}(\lambda) - c_n^*) \right| \geq \delta \right\}. \tag{B.58}$$

Then by Lemma B.4, for any $\eta > 0$ there exists $N' \in \mathbb{N}$ such that

$$\mathbf{P}(B_n) < \eta/2, \ \forall n \geq N'. \tag{B.59}$$

Define

$$W^{*,+\delta}(\theta_n^\epsilon, c) \equiv \{\lambda \in \rho B^d : p'\lambda = 0 \cap w^*_{j,\theta_n^\epsilon}(\lambda) \leq c + \delta, \ j \in \mathcal{J}^*\}. \tag{B.60}$$

Further define the event $R_{1n} \equiv \{U_n^*(\theta_n, c_n^*) \subset W^{*,+\delta}(\theta_n^\epsilon, c_n)\}$, and note that $B_n^c \subseteq R_{1n}$. The result in equation (B.45) then follows using similar steps to (B.55)-(B.57).

To establish (B.46), we distinguish three cases.

Case 1. Suppose first that $J_2 = 0$ (recalling that under Assumption 3.3′ this means that there is no $j = 1, \ldots, J_{11}$ such that $\pi^*_{1,j} = 0 = \pi^*_{1,j+J_{11}}$), and hence one has only moment inequalities. In this case, by (B.47) and (B.49), one may write

$$U_n^*(\theta_n, c) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap u^*_{n,j,\theta_n}(\lambda) \leq c, \ j \in \mathcal{J}^*\right\}, \tag{B.61}$$

$$W^{*,-\delta}(\theta_n^\epsilon, c) \equiv \left\{\lambda \in \rho B^d : p'\lambda = 0 \cap w^*_{j,\theta_n^\epsilon}(\lambda) \leq c - \delta, \ j \in \mathcal{J}^*\right\}, \tag{B.62}$$

where $W^{*,-\delta}$, $\delta > 0$, is obtained by tightening the inequality constraints shaping W^*. Define the event

$$R_{2n} \equiv \{W^{*,-\delta}(\theta_n^\epsilon, c_n) \subset U_n^*(\theta_n, c_n^*)\}, \tag{B.63}$$

and note that $B_n^c \subseteq R_{2n}$. The result in equation (B.46) then follows by Lemma B.6 using again similar steps to (B.55)-(B.57).

Case 2. Next suppose that $J_2 \geq d$. In this case, we define $W^{*,-\delta}$ to be the set obtained by tightening by δ the inequality constraints as well as each of the two opposing inequalities obtained from the equality constraints. That is,

$$W^{*,-\delta}(\theta_n^\epsilon, c) \equiv \{\lambda \in \rho B^d : p'\lambda = 0 \cap w^*_{j,\theta_n^\epsilon}(\lambda) \leq c - \delta, \ j \in \mathcal{J}^*\}, \tag{B.64}$$

that is, the same set as in (B.112) with $w^*_{j,\theta_n^\epsilon}(\lambda)$ replacing $w_{j,\theta_n^\epsilon}(\lambda)$ and defining the set using only inequalities in $\mathcal{J}^*$. Note that, by Lemma B.8, there exists $N \in \mathbb{N}$ such that for all $n \geq N$ $\hat{c}_n(\theta)$ is bounded from below by some $\underline{c} > 0$ with probability approaching one uniformly in $P \in \mathcal{P}$ and $\theta \in \Theta_I(P)$. This ensures the limit c of c_n is bounded from below by $\underline{c} > 0$. This in turn allows us to construct a non-empty tightened constraint set with probability approaching 1. Namely, for $\delta < \underline{c}$, $W_n^{*,-\delta}(\theta, c_n)$ is nonempty with probability approaching 1 by Lemma B.6, and hence its superset $W_n^*(\theta, c_n)$ is also non-empty with probability approaching 1. However, note that $B_n^c \subseteq R_{2n}$, where R_{2n} is in (B.63) now defined using the tightened constraint set $W_n^{*,-\delta}(\theta, c_n)$ being defined as in (B.64),

and therefore the same argument as in the previous case applies.

Case 3. Finally, suppose that $1 \leq J_2 < d$. Recall that

$$c = \lim_{n\to\infty} c_n, \tag{B.65}$$

and note that by construction $c \geq 0$. Consider first the case that $c > 0$. Then, by taking $\delta < c$, the argument in Case 2 applies.

Next consider the case that $c = 0$. Observe that

$$\begin{aligned}\mathbf{P}\Big(\{U_n^*(\theta_n, c_n^*) = \emptyset\} \cap \{W^*(\theta_n^\epsilon, c_n) \neq \emptyset\}\Big) \leq\ & \mathbf{P}\Big(\{U_n^*(\theta_n, c_n^*) = \emptyset\} \cap \{W^{*,-\delta}(\theta_n^\epsilon, 0) \neq \emptyset\}\Big) \\ &+ \mathbf{P}\Big(\{W^{*,-\delta}(\theta_n^\epsilon, 0) = \emptyset\} \cap \{W^*(\theta_n^\epsilon, 0) \neq \emptyset\}\Big) \\ &+ \mathbf{P}\Big(\{W^*(\theta_n^\epsilon, 0) = \emptyset\} \cap \{W^*(\theta_n^\epsilon, c_n) \neq \emptyset\}\Big),\end{aligned} \tag{B.66}$$

with $W^{*,-\delta}(\theta_n^\epsilon, 0)$ defined as in (B.17) with $c = 0$ and with $w^*_{j,\theta_n^\epsilon}(\lambda)$ replacing $w_{j,\theta_n^\epsilon}(\lambda)$.

By Lemma B.6, for any $\eta > 0$ there exists $\delta > 0$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\Big(\{W^{*,-\delta}(\theta_n^\epsilon, 0) = \emptyset\} \cap \{W^*(\theta_n^\epsilon, 0) \neq \emptyset\}\Big) < \eta/3 \ \forall n \geq N. \tag{B.67}$$

Moreover, because $c_n \xrightarrow{a.s.} 0$, an easy adaptation of the proof of Lemma B.6 yields that, for any $\eta > 0$, there exists $\delta > 0$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\Big(\{W^*(\theta_n^\epsilon, 0) = \emptyset\} \cap \{W^*(\theta_n^\epsilon, c_n) \neq \emptyset\}\Big) < \eta/3 \ \forall n \geq N. \tag{B.68}$$

In particular, $W^*(\theta_n^\epsilon, 0)$ relates to $W^*(\theta_n^\epsilon, c_n)$ by tightening each constraint $j \in \mathcal{J}^*$ and not only constraints $j \in \mathcal{J}^* \cap \{1, \ldots, J_1\}$. Consequently, τ in the proof of Lemma B.6 must be defined to have entries of 1 corresponding to all elements of $\mathcal{J}^*$, followed by $2d + 2$ entries of 0. Then most steps go through immediately. Case 2-(b) needs to be slightly modified: In that case, one now has $\sum_{j\in\mathcal{J}^*} \nu_j^t g_{P,j}(\theta) = c_n \sum_{j\in\{J_1+1,\cdots,J\}\cap\mathcal{J}^*} \nu_j^t$ and $\sum_{j\in\mathcal{J}^*} \nu_j^t \tau_j = c_n \sum_{j\in\{J_1+1,\cdots,J\}\cap\mathcal{J}^*} \nu_j^t$, so the argument for case 1 applies. In sum, the last two terms on the right hand side of (B.66) are arbitrarily small.

We now consider the first term on the right hand side of (B.66). Let $g_{P_n}(\theta_n^\epsilon)$ be a $J + 2d + 2$ vector with

$$g_{P_n,j}(\theta_n^\epsilon) \equiv \begin{cases} -\mathbb{G}_j^*(\theta_n^\epsilon), & \text{if } j \in \mathcal{J}^*, \\ \rho, & \text{if } j = J+1, \ldots, J+2d, \\ 0, & \text{if } j = J+2d+1, J+2d+2 \\ 0, & \text{otherwise,} \end{cases} \tag{B.69}$$

where we used that $\pi^*_{1,j} = 0$ for $j \in \mathcal{J}^*$ and where the last assignment is without loss of generality because of the considerations leading to the sets in (B.47)-(B.50). For a given set $C \subset \{1, \ldots, J + 2d + 2\}$, let the vector $g^C_{P_n}(\theta_n^\epsilon)$ collect the entries of $g_{P_n}(\theta_n^\epsilon)$ corresponding to indexes in C, and let the matrix $K^C_{P_n}(\theta_n^\epsilon)$ collect the rows of $K_{P_n}(\theta_n^\epsilon)$ corresponding to indexes in C and K_{P_n} as defined in (B.18) with P_n replacing P.

Let $\widetilde{\mathcal{C}}$ collect all size d subsets C of $\{1, ..., J + 2d + 2\}$ ordered lexicographically by their smallest, then second smallest, etc. elements. Let the random variable $\mathcal{C}(\theta)$ (dependence on many other quantities is suppressed) equal the first element of $\widetilde{\mathcal{C}}$ s.t. $\det K^C_P(\theta) \neq 0$ and $\lambda^C = \left(K^C_P(\theta)\right)^{-1} g^C_P(\theta) \in W^{*,-\delta}(\theta, 0)$ if such an element exists; else, let $\mathcal{C}(\theta) = \{J + 1, ..., J + d\}$ and $\lambda^C = \rho \mathbf{1}_d$, where

$\mathbf{1}_d$ denotes a d vector with each entry equal to 1. Recall that $W^{*,-\delta}(\theta,0)$ is a (possibly empty) measurable random polyhedron in a compact subset of $\mathbb{R}^d$, see, e.g., Molchanov (2005, Definition 1.1.1). Thus, if $W^{*,-\delta}(\theta,0)\neq\emptyset$, then $W^{*,-\delta}(\theta,0)$ has extreme points, each of which is characterized as the intersection of d (not necessarily unique) linearly independent constraints interpreted as equalities. Therefore, $W^{*,-\delta}(\theta,0)\neq\emptyset$ implies that $\lambda^{\mathcal{C}(\theta)}\in W^{*,-\delta}(\theta,0)$ and therefore also that $\mathcal{C}(\theta)\subset\mathcal{J}^*\cup\{J+1,\ldots,J+2d+2\}$. Note that the associated random vector $\lambda^{\mathcal{C}}(\theta)$ is a measurable selection of a random closet set that equals $W^{*,-\delta}(\theta,0)$ if $W^{*,-\delta}(\theta,0)\neq\emptyset$ and equals ρB^d otherwise, see, e.g., Molchanov (2005, Definition 1.2.2).

Lemma B.7 establishes that for any $\eta>0$, there exist $\alpha_\eta>0$ and N s.t. $n\geq N$ implies

$$\mathbf{P}\left(W^{*,-\delta}(\theta_n^\epsilon,0)\neq\emptyset,\left|\det K_P^{\mathcal{C}(\theta_n^\epsilon)}(\theta_n^\epsilon)\right|\leq\alpha_\eta\right)\leq\eta, \tag{B.70}$$

which in turn, given our definition of $\mathcal{C}(\theta_n^\epsilon)$, yields that there is $M>0$ and N such that

$$\mathbf{P}\Big(\left|\det\left(K_{P_n}^{\mathcal{C}(\theta_n^\epsilon)}(\theta_n^\epsilon)\right)^{-1}\right|\leq M\Big)\geq 1-\eta,\ \forall n\geq N. \tag{B.71}$$

For each n and $\lambda\in\rho B^d$, define the mapping $\phi_n:\rho B^d\to\mathbb{R}^d_{[\pm\infty]}$ by

$$\phi_n(\lambda)\equiv\left(K_{P_n}^{\mathcal{C}(\theta_n^\epsilon)}(\bar\theta(\theta_n,\lambda))\right)^{-1}\tilde g_n^{\mathcal{C}(\theta_n^\epsilon)}(\theta_n+\lambda/\sqrt{n}), \tag{B.72}$$

where the notation $\bar\theta(\theta,\lambda)$ emphasizes that $\bar\theta$ depends on θ and λ because it lies component-wise between θ and $\theta+\lambda/\sqrt{n}$, and where the vector $\tilde g_n^{\mathcal{C}(\theta_n^\epsilon)}(\theta_n+\lambda/\sqrt{n})$ collects the entries corresponding to indexes in $\mathcal{C}(\theta_n^\epsilon)$ of a $J+2d+2$ vector $\tilde g_n(\theta_n+\lambda/\sqrt{n})$ with

$$\tilde g_{n,j}(\theta+\lambda/\sqrt{n})\equiv\begin{cases}c_n^*/(1+\eta_{n,j}^*)-\mathbb{G}_{n,j}^*(\theta+\lambda/\sqrt{n}) & \text{if } j\in\mathcal{J}^*,\\ \rho, & \text{if } j=J+1,\ldots,J+2d,\\ 0, & \text{if } j=J+2d+1,J+2d+2,\\ 0, & \text{otherwise,}\end{cases} \tag{B.73}$$

using again that $\pi_{1,j}^*=0$ for $j\in\mathcal{J}^*$ and that the last assignment is without loss of generality.

We show that ϕ_n is a contraction mapping and hence has a fixed point. To simplify notation, in what follows we omit the dependence of $\mathcal{C}$ on θ_n^ϵ.

For any $\lambda,\lambda'\in\rho B^d$ write

$$\begin{aligned}&\|\phi_n(\lambda)-\phi_n(\lambda')\|\\ &\quad=\left\|\left(K_{P_n}^{\mathcal{C}}(\bar\theta(\theta_n,\lambda))\right)^{-1}\tilde g_n^{\mathcal{C}}(\theta_n+\lambda/\sqrt{n})-\left(K_{P_n}^{\mathcal{C}}(\bar\theta(\theta_n,\lambda'))\right)^{-1}\tilde g_n^{\mathcal{C}}(\theta_n+\lambda'/\sqrt{n})\right\|\\ &\quad\leq\left\|\left(K_{P_n}^{\mathcal{C}}(\bar\theta(\theta_n,\lambda))\right)^{-1}\right\|_{op}\left\|\tilde g_n^{\mathcal{C}}(\theta_n+\lambda/\sqrt{n})-\tilde g_n^{\mathcal{C}}(\theta_n+\lambda'/\sqrt{n})\right\|\\ &\qquad+\left\|\left(K_{P_n}^{\mathcal{C}}(\bar\theta(\theta_n,\lambda))\right)^{-1}-\left(K_{P_n}^{\mathcal{C}}(\bar\theta(\theta_n,\lambda'))\right)^{-1}\right\|_{op}\left\|\tilde g_n^{\mathcal{C}}(\theta_n+\lambda'/\sqrt{n})\right\|,\end{aligned} \tag{B.74}$$

where $\|\cdot\|_{op}$ denotes the operator norm.

By Assumption 3.5 (i), for any $\eta>0$, $k>0$, there is $N\in\mathbb{N}$ such that

$$\begin{aligned}&\mathbf{P}\left(\left\|\tilde g_n^{\mathcal{C}}(\theta_n+\lambda/\sqrt{n})-\tilde g_n^{\mathcal{C}}(\theta_n+\lambda'/\sqrt{n})\right\|\leq k\|\lambda-\lambda'\|\right)\\ &\quad=\mathbf{P}\big(\|\mathbb{G}_n^{*,\mathcal{C}}(\theta_n+\lambda/\sqrt{n})-\mathbb{G}_n^{*,\mathcal{C}}(\theta_n+\lambda'/\sqrt{n})\|\leq k\|\lambda-\lambda'\|\big)\geq 1-\eta,\ \forall n\geq N.\end{aligned} \tag{B.75}$$

Moreover, by arguing as in equation (A.19), for any η there exist $0 < L < \infty$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\big(\sup_{\lambda' \in \rho B^d} \big\| \tilde{g}_n^{\mathcal{C}}(\theta_n + \lambda'/\sqrt{n}) \big\| \leq L \big) \geq 1 - \eta, \ \forall n \geq N. \tag{B.76}$$

For any invertible matrix K, $\|K^{-1}\|_{op} \leq |\det(K)^{-1}| \|adj(K)\|_{op}$. Hence, by Assumption 3.4-(i) and equation (B.71), for any $\eta > 0$, there exist $0 < L < \infty$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\big(\big\| \big(K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon}) \big)^{-1} \big\|_{op} \leq L \big) \geq \mathbf{P}\big(\big| \det\big(K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon}) \big)^{-1} \big| b(\bar{M} + \rho M/\sqrt{n}) \leq L \big) \geq 1 - \eta, \ \forall n \geq N, \tag{B.77}$$

where $b > 0$ is a constant that depends only on d, $\bar{M}$ is defined in Assumption 3.4-(i) and M is defined in Assumption 3.4-(ii). By Horn and Johnson (1985, ch. 5.8), for any invertible matrices $K, \tilde{K}$ such that $\|\tilde{K}^{-1}(K - \tilde{K})\|_{op} < 1$,

$$\|K^{-1} - \tilde{K}^{-1}\|_{op} \leq \frac{\|\tilde{K}^{-1}(K - \tilde{K})\|_{op}}{1 - \|\tilde{K}^{-1}(K - \tilde{K})\|_{op}} \|\tilde{K}^{-1}\|_{op}. \tag{B.78}$$

By (B.78), $\|K^{-1}\|_{op} \leq |\det(K)^{-1}| \|adj(K)\|_{op}$, and the triangle inequality, for any $\eta > 0$, there exist $0 < L < \infty$ and $N \in \mathbb{N}$ such that

$$\begin{aligned}
&\mathbf{P}\big(\sup_{\lambda \in \rho B^d} \big\| \big(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n, \lambda)) \big)^{-1} \big\|_{op} \leq 2L \big) \\
&\geq \mathbf{P}\big(\big\| \big(K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon}) \big)^{-1} \big\|_{op} + \sup_{\lambda \in \rho B^d} \|K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n, \lambda))^{-1} - K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon})^{-1}\|_{op} \leq 2L \big) \\
&\geq \mathbf{P}\big(\big\| \big(K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon}) \big)^{-1} \big\|_{op} \leq L, \big| \det\big(K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon}) \big)^{-1} \big| b(\bar{M} + \rho M/\sqrt{n}) \sup_{\lambda \in \rho B^d} \|K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n, \lambda)) - K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon})\|_{op} \leq L \big) \\
&\geq \mathbf{P}\big(\big\| \big(K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon}) \big)^{-1} \big\|_{op} \leq L, \big| \det\big(K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon}) \big)^{-1} \big| b(\bar{M} + \rho M/\sqrt{n}) \rho M/\sqrt{n} \leq L \big) \geq 1 - 2\eta, \ \forall n \geq N,
\end{aligned} \tag{B.79}$$

where the last inequality follows from $\|K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n, \lambda)) - K_{P_n}^{\mathcal{C}}(\theta_n^{\epsilon})\|_{op} \leq \|D(\bar{\theta}(\theta_n, \lambda)) - D(\theta_n^{\epsilon})\|_{op} \leq M\rho/\sqrt{n}$ by Assumption 3.4 (ii), (B.71) and (B.77). Again by applying (B.78), for any $k > 0$, there exists $N \in \mathbb{N}$ such that

$$\begin{aligned}
&\mathbf{P}\big(\big\| \big(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n, \lambda)) \big)^{-1} - \big(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n, \lambda')) \big)^{-1} \big\|_{op} \leq k\|\lambda - \lambda'\| \big) \\
&\geq \mathbf{P}\big(\sup_{\lambda \in \rho B^d} \big\| \big(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n, \lambda)) \big)^{-1} \big\|_{op}^2 M \|\bar{\theta}(\theta_n, \lambda) - \bar{\theta}(\theta_n, \lambda')\| \leq k\|\lambda - \lambda'\| \big) \geq 1 - \eta, \ \forall n \geq N,
\end{aligned} \tag{B.80}$$

where the first inequality follows from $\|K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n, \lambda)) - K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n, \lambda'))\|_{op} \leq M\|\bar{\theta}(\theta_n, \lambda) - \bar{\theta}(\theta_n, \lambda')\| \leq M/\sqrt{n}\|\lambda - \lambda'\|$ by Assumption 3.4 (ii), and the last inequality follows from (B.79).

By (B.74)-(B.76) and (B.79)-(B.80), it then follows that there exists $\beta \in [0, 1)$ such that for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that

$$\mathbf{P}\left(|\phi_n(\lambda) - \phi_n(\lambda')| \leq \beta\|\lambda - \lambda'\|, \ \ \forall \lambda, \lambda' \in \rho B^d \right) \geq 1 - \eta, \ \forall n \geq N. \tag{B.81}$$

This implies that with probability approaching 1, each $\phi_n(\cdot)$ is a contraction, and therefore by the Contraction Mapping Theorem it has a fixed point (e.g., Pata (2014, Theorem 1.3)). This in turn implies that for any $\eta > 0$ there exists a $N \in \mathbb{N}$ such that

$$\mathbf{P}\left(\exists \lambda_n^f : \lambda_n^f = \phi_n(\lambda_n^f) \right) \geq 1 - \eta, \ \forall n \geq N. \tag{B.82}$$

Next, define the mapping

$$\psi_n(\lambda) \equiv \left(K^{\mathcal{C}}_{P_n}(\theta^{\epsilon}_n)\right)^{-1} g^{\mathcal{C}}_{P_n}(\theta^{\epsilon}_n). \tag{B.83}$$

This map is constant in λ and hence is uniformly continuous and a contraction with Lipschitz constant equal to zero. It therefore has $\lambda^{\mathcal{C}}_n$ as its fixed point. Moreover, by (B.72) and (B.83) arguing as in (B.74), it follows that for any $\lambda \in \rho B^d$,

$$\begin{aligned}\|\psi_n(\lambda) - \phi_n(\lambda)\| \leq & \left\|\left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n,\lambda))\right)^{-1}\right\|_{op}\left\|g^{\mathcal{C}}_{P_n}(\theta^{\epsilon}_n) - \tilde{g}^{\mathcal{C}}_n(\theta_n + \lambda/\sqrt{n})\right\| \\ & + \left\|\left(K^{\mathcal{C}}_{P_n}(\theta^{\epsilon}_n)\right)^{-1} - \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n,\lambda))\right)^{-1}\right\|_{op}\left\|g^{\mathcal{C}}_{P_n}(\theta^{\epsilon}_n)\right\|. \end{aligned} \tag{B.84}$$

By (B.69) and (B.73)

$$\begin{aligned}\left\|g^{\mathcal{C}}_{P_n}(\theta^{\epsilon}_n) - \tilde{g}^{\mathcal{C}}_n(\theta_n + \lambda/\sqrt{n})\right\| &\leq \max_{j\in\mathcal{J}^*} |-\mathbb{G}^*_j(\theta^{\epsilon}_n) - c^*_n/(1+\eta^*_{n,j}) + \mathbb{G}^*_{n,j}(\theta_n + \lambda/\sqrt{n})| \\ &\leq \max_{j\in\mathcal{J}^*} |\mathbb{G}^*_j(\theta^{\epsilon}_n) - \mathbb{G}^*_{n,j}(\theta_n + \lambda/\sqrt{n})| + \max_{j\in\mathcal{J}^*} |c^*_n/(1+\eta^*_{n,j})|. \end{aligned} \tag{B.85}$$

We note that when Assumption 3.3′ is used, for each $j = 1,\ldots,J_{11}$ such that $\pi^*_{1,j} = 0 = \pi^*_{1,j+J_{11}}$ we have that $|\tilde{\mu}_j - \mu_j| = o_{\mathcal{P}}(1)$ because $\sup_{\theta\in\Theta}|\eta_j(\theta)| = o_{\mathcal{P}}(1)$, where $\tilde{\mu}_j$ and μ_j were defined in (A.10)-(A.11) and (B.12)-(B.13) respectively. Moreover, Assumption 3.5 (ii) implies $\mathbb{G}^*_n \overset{a.s.}{\to} \mathbb{G}^*$ in $l^\infty(\Theta)$ and (B.65) implies $c^*_n \to 0$, so that we have

$$\sup_{\lambda\in\rho B^d} \|g^{\mathcal{C}}_{P_n}(\theta^{\epsilon}_n) - \tilde{g}^{\mathcal{C}}_n(\theta_n + \lambda/\sqrt{n})\| \overset{a.s.}{\to} 0. \tag{B.86}$$

Further, by (B.78) and Assumption 3.4-(ii),

$$\sup_{\lambda\in\rho B^d}\left\|\left(K^{\mathcal{C}}_{P_n}(\theta^{\epsilon}_n)\right)^{-1} - \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n,\lambda))\right)^{-1}\right\|_{op} \leq M \sup_{\lambda\in\rho B^d}\|\bar{\theta}(\theta_n,\lambda) - \theta^{\epsilon}_n\| \leq M\rho/\sqrt{n}. \tag{B.87}$$

In sum, by (B.76), (B.79), and (B.85)-(B.87), for any $\eta,\nu > 0$, there exists $N \geq \mathbb{N}$ such that

$$\mathbf{P}\left(\sup_{\lambda\in\rho B^d}\|\psi_n(\lambda) - \phi_n(\lambda)\| < \nu\right) \geq 1-\eta,\ \forall n \geq \mathbb{N}. \tag{B.88}$$

Hence, for a specific choice of $\nu = \kappa(1-\beta)$, where β is defined in equation (B.81), we have that $\sup_{\lambda\in\rho B^d}\|\psi_n(\lambda) - \phi_n(\lambda)\| < \kappa(1-\beta)$ implies

$$\begin{aligned}\|\lambda^{\mathcal{C}}_n - \lambda^f_n\| &= \|\psi_n(\lambda^{\mathcal{C}}_n) - \phi_n(\lambda^f_n)\| \\ &\leq \|\psi_n(\lambda^{\mathcal{C}}_n) - \phi_n(\lambda^{\mathcal{C}}_n)\| + \|\phi_n(\lambda^{\mathcal{C}}_n) - \phi_n(\lambda^f_n)\| \\ &\leq \kappa(1-\beta) + \beta\|\lambda^{\mathcal{C}}_n - \lambda^f_n\| \end{aligned} \tag{B.89}$$

Rearranging terms, we obtain $\|\lambda^{\mathcal{C}}_n - \lambda^f_n\| \leq \kappa$. Note that by Assumptions 3.4 (i) and 3.5 (i), for any $\delta > 0$, there exists $\kappa_\delta > 0$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\big(\sup_{\|\lambda-\lambda'\|\leq\kappa_\delta}|u^*_{n,j,\theta_n}(\lambda) - u^*_{n,j,\theta_n}(\lambda')| < \delta\big) \geq 1-\eta,\ \forall n \geq \mathbb{N}, \tag{B.90}$$

For $\lambda^{\mathcal{C}}_n \in W^{*,-\delta}(\theta^{\epsilon}_n, 0)$, one has

$$w^*_{j,\theta^{\epsilon}_n}(\lambda^{\mathcal{C}}_n) + \delta \leq 0,\ j \in \{1,\cdots,J_1\} \cap \mathcal{J}^*. \tag{B.91}$$

Hence, by (B.59), (B.65), and (B.90)-(B.91), $\|\lambda_n^{\mathcal{C}} - \lambda_n^f\| \leq \kappa_{\delta/4}$, for each $j \in \{1, \cdots, J_1\} \cap \mathcal{J}^*$ we have

$$u^*_{n,j,\theta_n}(\lambda_n^f) - c_n^*(\theta_n) \leq u^*_{n,j,\theta_n}(\lambda_n^{\mathcal{C}}) - c_n^*(\theta_n) + \delta/4 \leq w^*_{j,\theta_n}(\lambda_n^{\mathcal{C}}) + \delta/2 \leq 0. \tag{B.92}$$

For $j \in \{J_1 + 1, \cdots, 2J_2\} \cap \mathcal{J}^*$, the inequalities hold by construction given the definition of $\mathcal{C}$.

In sum, for any $\eta > 0$ there exists $\delta > 0$ and $N \in \mathbb{N}$ such that for all $n \geq N$ we have

$$\begin{aligned}\mathbf{P}\Big(\{U_n^*(\theta_n, c_n^*) = \emptyset\} \cap \{W^{*,-\delta}(\theta_n^\epsilon, 0) \neq \emptyset\}\Big) &\leq \mathbf{P}\Big(\nexists \lambda_n^f \in U_n^*(\theta_n, c_n^*), \exists \lambda_n^{\mathcal{C}} \in W^{*,-\delta}(\theta_n^\epsilon, 0)\Big) \\ &\leq \mathbf{P}\left(\left\{\sup_{\lambda \in \rho B^d} \|\psi_n(\lambda) - \phi_n(\lambda)\| < \kappa_\delta(1-\beta) \cap B_n\right\}^c\right) \leq \eta/3, \end{aligned} \tag{B.93}$$

where A^c denotes the complement of the set A, and the last inequality follows from (B.59) and (B.88). Combining equations (B.67), (B.68), and (B.93) yields the desired result for Case 3. □

LEMMA B.3: *Let θ_n^ϵ be as defined in (B.21). Then*

$$\lim_{n\to\infty} P_n^*\left(V_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\right) \geq 1 - \alpha. \tag{B.94}$$

Proof. Let

$$\begin{aligned} V_n^b(\theta, c) &\equiv \{\lambda \in \rho B^d : \mathbb{G}^b_{n,j}(\theta) + \hat{D}_{n,j}(\theta)\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta)) \leq c,\ j = 1, \ldots, J\} \cap \{p'\lambda = 0\} \\ &= \Lambda_n^b(\theta, \rho, c) \cap \{p'\lambda = 0\}, \end{aligned} \tag{B.95}$$

with $\Lambda_n^b(\theta, \rho, c)$ defined in (2.7). By construction, see (2.11), for all $\theta \in \Theta$,

$$P_n^*\left(\{V_n^b(\theta, \hat{c}_n(\theta)) \neq \emptyset\right) \geq 1 - \alpha. \tag{B.96}$$

Inspection of equations (2.7) and (B.14) shows that $V_n^b(\theta, c)$ and $V_n(\theta, c)$ differ exclusively in that the first set features sample GMS, $\varphi_j(\hat{\xi}_{n,j}(\theta_n^\epsilon))$, in the stochastic inequalities, whereas the second set features $\pi^*_{1,j}$. Observe that

$$P_n^*(V_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset) \geq P_n^*\Big(V_n^b(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\Big) - P_n^*\Big(\{V_n^b(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\} \cap \{V_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) = \emptyset\}\Big), \tag{B.97}$$

where we used that given any two events A, B,

$$P(A \neq \emptyset) \geq P(B \neq \emptyset) - P(\{B \neq \emptyset\} \cap \{A = \emptyset\})$$

We now establish that the last term in (B.97) is $o_{\mathcal{P}}(1)$. We have

$$\begin{aligned} &P_n^*\Big(\{V_n^b(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\} \cap \{V_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) = \emptyset\}\Big) \\ &\quad \leq P_n^*\Big(\{V_n^b(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\} \cap \{V_n^{b,-\delta}(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) = \emptyset\}\Big) \\ &\qquad + P_n^*\Big(\{V_n^{b,-\delta}(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\} \cap \{V_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) = \emptyset\}\Big). \end{aligned} \tag{B.98}$$

By Lemma B.6, for any $\eta > 0$ there exists $\delta > 0$ and $N \in \mathbb{N}$ such that

$$P_n^*\Big(\{V_n^b(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \neq \emptyset\} \cap \{V_n^{b,-\delta}(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) = \emptyset\}\Big) < \eta/2,\ \forall n \geq N. \tag{B.99}$$

Consider first the case that Assumption 3.3 holds. The result then follows from (B.99) and the fact that by Lemma B.5, if $\pi_j^* = 0$ then $\hat{\xi}_{n,j}(\theta_n^\epsilon) = o_{\mathcal{P}}(1)$, so that for n large enough, with probability

at least $1-\eta/2$, by Assumption 3.2 $|\varphi_j(\hat{\xi}_{n,j}(\theta_n^\epsilon))| < \delta$. Observing that $\pi_j^* \in \{0, -\infty\}$, we have that for all $\eta > 0$ there is $\delta > 0$ and $N \in \mathbb{N}$ such that

$$P_n^*\Big(V_n^{b,-\delta}(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon)) \subseteq V_n(\theta_n^\epsilon, \hat{c}_n(\theta_n^\epsilon))\Big) \geq 1-\eta/2, \ \forall n \geq N, \tag{B.100}$$

yielding the result.

Consider next the case that Assumption 3.3′ holds. In this case, the set $V_n^b(\theta_n^\epsilon, c)$ is defined with hard threshold GMS as in equation (2.9). The same argument of proof as just provided applies. The only case that might create concern is one in which

$$\pi_{1,j}^* = -\infty, \text{ and } \pi_{j+J_{11}}^* = 0,$$
$$\varphi_j(\hat{\xi}_{n,j}(\theta_n^\epsilon) = 0, \text{ and } \varphi_{j+J_{11}}(\hat{\xi}_{n,j+J_{11}}(\theta_n^\epsilon) = 0,$$

so that in the set $V_n^b(\theta_n^\epsilon, c)$ inequality $j + J_{11}$, which is

$$\mathbb{G}_{n,j+J_{11}}^b(\theta_n^\epsilon) + \hat{D}_{n,j+J_{11}}(\theta_n^\epsilon)\lambda \leq c,$$

is replaced with inequality

$$-\mathbb{G}_{n,j}^b(\theta_n^\epsilon) - \hat{D}_{n,j}(\theta_n^\epsilon)\lambda \leq c,$$

as explained in Section 3.1. For this case, Lemma B.9 establishes that

$$\|\mathbb{G}_{n,j+J_{11}}^b(\theta_n^\epsilon) + \hat{D}_{n,j+J_{11}}(\theta_n^\epsilon) + \mathbb{G}_{n,j}^b(\theta_n^\epsilon) + \hat{D}_{n,j}(\theta_n^\epsilon)\lambda\| = o_{\mathcal{P}}(1). \tag{B.101}$$

Note that (B.96) continues to hold if inequality

$$\mathbb{G}_{n,j}^b(\theta_n^\epsilon) + \hat{D}_{n,j}(\theta_n^\epsilon)\lambda \leq c,$$

is dropped from the set $V_n^b(\theta_n^\epsilon, c)$, because the program is thereby relaxed. We can then define set $V_n^{b,-\delta}(\theta_n^\epsilon, c)$ with inequality j dropped, and including a delta contraction of the inequality that replaces inequality $j + J_{11}$, namely

$$-\mathbb{G}_{n,j}^b(\theta_n^\epsilon) - \hat{D}_{n,j}(\theta_n^\epsilon)\lambda \leq c - \delta.$$

Therefore, using (B.101) the same argument of proof applies as for the case that Assumption 3.3 holds. □

LEMMA B.4: *Let (P_n, θ_n) have the almost sure representations given in Lemma B.1, let $\mathcal{J}^*$ be defined as in (B.37), and assume that $\mathcal{J}^* \neq \emptyset$. Let θ_n^ϵ be as defined in (B.21). Then, for any $\varepsilon, \eta > 0$, there exists $N' \in \mathbb{N}$ and $N'' \in \mathbb{N}$ such that*

$$\mathbf{P}\left(\sup_{\lambda \in \rho B^d \cap \sqrt{n}(\Theta - \theta_n)} \left| \max_{j=1,\cdots,J}(u_{n,j,\theta_n}^*(\lambda) - c_n^*) - \max_{j=1,\cdots,J}(w_{j,\theta_n^\epsilon}^*(\lambda) - c_n) \right| \geq \varepsilon \right) < \eta, \tag{B.102}$$

for all $n \geq N'$ and

$$\tilde{\mathbf{P}}\left(\sup_{\lambda \in \rho B^d \cap \sqrt{n}(\Theta - \theta_n)} \left| \max_{j=1,\cdots,J} \tilde{w}_{j,\theta_n^\epsilon}^*(\lambda) - \max_{j=1,\cdots,J} v_{n,j,\theta_n^\epsilon}^*(\lambda) \right| \geq \varepsilon \right) < \eta, \tag{B.103}$$

for all $n \geq N''$, where the functions $u^*, v^*, w^*, \tilde{w}^*$ are defined in equations (B.27),(B.28),(B.29) and (B.30).

Proof. We first establish (B.102). By definition, $\pi^*_{1,j} = -\infty$ for all $j \notin \mathcal{J}^*$, and therefore

$$\mathbf{P}(\sup_{\lambda\in\rho B^d\cap\sqrt{n}(\Theta-\theta_n)} |\max_{j=1,\cdots,J}(u^*_{n,j,\theta_n}(\lambda)-c^*_n) - \max_{j=1,\cdots,J}(w^*_{j,\theta^\epsilon_n}(\lambda)-c_n)| \geq \varepsilon)$$
$$= \mathbf{P}(\sup_{\lambda\in\rho B^d\cap\sqrt{n}(\Theta-\theta_n)} |\max_{j\in\mathcal{J}^*}(u^*_{n,j,\theta_n}(\lambda)-c^*_n) - \max_{j\in\mathcal{J}^*}(w^*_{j,\theta^\epsilon_n}(\lambda)-c_n)| \geq \varepsilon). \quad \text{(B.104)}$$

Hence, for the conclusion of the lemma, it suffices to show

$$\lim_{n\to\infty}\mathbf{P}(\sup_{\lambda\in\rho B^d\cap\sqrt{n}(\Theta-\theta_n)} |\max_{j\in\mathcal{J}^*}(u^*_{n,j,\theta_n}(\lambda)-c^*_n) - \max_{j\in\mathcal{J}^*}(w^*_{j,\theta^\epsilon_n}(\lambda)-c_n)| \geq \varepsilon) = 0.$$

For each $\lambda\in\mathbb{R}^d$, define $r_{n,j,\theta_n}(\lambda) \equiv (u^*_{n,j,\theta_n}(\lambda)-c^*_n)-(w^*_{j,\theta^\epsilon_n}(\lambda)-c_n)$. Using the fact that $\pi^*_{1,j}=0$ for $j\in\mathcal{J}^*$, and the triangle and Cauchy-Schwarz inequalities, for any $\lambda\in\rho B^d\cap\sqrt{n}(\Theta-\theta_n)$ and $j\in\mathcal{J}^*$, we have

$$\begin{aligned}|r_{n,j,\theta_n}(\lambda)| &\leq |\mathbb{G}^*_{n,j}(\theta_n+\lambda/\sqrt{n}) - \mathbb{G}^*_j(\theta^\epsilon_n)| + \|D_{P_n,j}(\bar\theta_n) - D_{P_n,j}(\theta^\epsilon_n)\|\|\lambda\| \\ &\quad + |\mathbb{G}^*_{n,j}(\theta_n+\lambda/\sqrt{n}) + D_{P_n,j}(\bar\theta_n)\lambda|\eta_{n,j}(\theta_n+\lambda/\sqrt{n}) + |c^*_n - c_n| \\ &\leq |\mathbb{G}^*_{n,j}(\theta_n+\lambda/\sqrt{n}) - \mathbb{G}^*_j(\theta^\epsilon_n)| + o_{\mathcal{P}}(1) + \{O_{\mathcal{P}}(1)+O(1)\}o_{\mathcal{P}}(1) + o_{\mathcal{P}}(1), \quad \text{(B.105)}\end{aligned}$$

where the last inequality follows using the fact that $\|\theta_n-\theta^\epsilon_n\| = O(1/\sqrt{n})$ together with the Lipschitz continuity of $D_{P,j}$ (Assumption 3.4-(ii)) and $\bar\theta_n$ being a mean value between θ_n and $\theta_n+\lambda/\sqrt{n}$, $\|\lambda\|\leq\rho$, $\|\mathbb{G}_{n,j}(\theta+\lambda/\sqrt{n})\| = O_{\mathcal{P}}(1)$, $\|D_{P,j}(\theta)\|$ being uniformly bounded (Assumption 3.4-(i)), $\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)| = o_{\mathcal{P}}(\kappa_n/\sqrt{n})$ by Assumption 3.3-(ii) (or Assumption 3.3′-(ii)), and equation (B.20). We note that when Assumption 3.3′ is used, for each $j=1,\ldots,J_{11}$ such that $\pi^*_{1,j}=0=\pi^*_{1,j+J_{11}}$ we have that $|\tilde\mu_j-\mu_j| = o_{\mathcal{P}}(1)$ because $\sup_{\theta\in\Theta}|\eta_j(\theta)| = o_{\mathcal{P}}(1)$, where $\tilde\mu_j$ and μ_j were defined in (A.10)-(A.11) and (B.12)-(B.13) respectively.

By (B.105) and the uniform stochastic equicontinuity of $\{\mathbb{G}_{n,j}\}$ (Assumption 3.5) inherited by its almost sure representation, and the fact that $j\in\mathcal{J}^*$, we have

$$\sup_{\lambda\in\rho B^d\cap\sqrt{n}(\Theta-\theta_n)} |\max_{j\in\mathcal{J}^*}(u^*_{n,j,\theta_n}(\lambda)-c^*_n) - \max_{j\in\mathcal{J}^*}(w^*_{j,\theta^\epsilon_n}(\lambda)-c_n)|$$
$$\leq \sup_{\lambda\in\rho B^d\cap\sqrt{n}(\Theta-\theta_n)} \max_{j\in\mathcal{J}^*}|r_{n,j,\theta_n}(\lambda)| = o_{\mathcal{P}}(1). \quad \text{(B.106)}$$

The conclusion of the lemma then follows from (B.104) and (B.106).

The result in (B.103) follows from similar arguments. □

LEMMA B.5: *Given a sequence $\{Q_n,\vartheta_n\}\in\{(P,\theta): P\in\mathcal{P},\theta\in\Theta_I(P)\}$ such that $\lim_{n\to\infty}\kappa_n^{-1}\sqrt{n}\gamma_{1,Q_n,j}(\vartheta_n)$ exists for each $j=1,\ldots,J$, let $\chi_j(\{Q_n,\vartheta_n\})$ be a function of the sequence $\{Q_n,\vartheta_n\}$ defined as*

$$\chi_j(\{Q_n,\vartheta_n\}) \equiv \begin{cases} 0, & \text{if } \lim_{n\to\infty}\kappa_n^{-1}\sqrt{n}\gamma_{1,Q_n,j}(\vartheta_n)=0, \\ -\infty, & \text{if } \lim_{n\to\infty}\kappa_n^{-1}\sqrt{n}\gamma_{1,Q_n,j}(\vartheta_n)<0. \end{cases} \quad \text{(B.107)}$$

Then for any $\theta'_n\in\theta_n+\frac{\rho}{\sqrt{n}}B^d$ for all n, one has: (i) $\kappa_n^{-1}\sqrt{n}\gamma_{1,P_n,j}(\theta_n)-\kappa_n^{-1}\sqrt{n}\gamma_{1,P_n,j}(\theta'_n)=o(1)$; (ii) $\chi(\{P_n,\theta_n\})=\chi(\{P_n,\theta'_n\})=\pi^*_{1,j}$; and (iii) $\kappa_n^{-1}\frac{\sqrt{n}\bar m_{n,j}(\theta'_n)}{\hat\sigma_{n,j}(\theta'_n)}-\kappa_n^{-1}\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta'_n)]}{\sigma_{P_n,j}(\theta'_n)}=o_{\mathcal{P}}(1)$.

Proof. For (i), the mean value theorem yields

$$\sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P),\theta'\in\theta+\rho/\sqrt{n}B^d}\left|\frac{\sqrt{n}E_P(m_j(X,\theta))}{\kappa_n\sigma_{P,j}(\theta)}-\frac{\sqrt{n}E_P(m_j(X,\theta'))}{\kappa_n\sigma_{P,j}(\theta')}\right|$$
$$\leq \sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P),\theta'\in\theta+\rho/\sqrt{n}B^d}\frac{\sqrt{n}\|D_{P,j}(\tilde{\theta})\|\|\theta'-\theta\|}{\kappa_n}=o(1), \quad \text{(B.108)}$$

where $\tilde{\theta}$ represents a mean value that lies componentwise between θ and θ' and where we used the fact that $D_{P,j}(\theta)$ is Lipschitz continuous and $\sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P)}\|D_{P,j}(\theta)\|\leq\bar{M}$.

Result (ii) then follows immediately from (B.107).

For (iii), note that

$$\sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d}\left|\kappa_n^{-1}\frac{\sqrt{n}\bar{m}_{n,j}(\theta_n')}{\hat{\sigma}_{n,j}(\theta_n')}-\kappa_n^{-1}\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')}\right|$$
$$\leq \sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d}\left|\kappa_n^{-1}\frac{\sqrt{n}(\bar{m}_{n,j}(\theta_n')-E_{P_n}[m_j(X_i,\theta_n')])}{\sigma_{n,j}(\theta_n')}(1+\eta_{n,j}(\theta_n'))+\kappa_n^{-1}\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')}\eta_{n,j}(\theta_n')\right|$$
$$\leq \sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d}|\kappa_n^{-1}\mathbb{G}_n(\theta_n')(1+\eta_{n,j}(\theta_n'))|+\left|\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\kappa_n\sigma_{P_n,j}(\theta_n')}\eta_{n,j}(\theta_n')\right|=o_{\mathcal{P}}(1), \quad \text{(B.109)}$$

where the last equality follows from $\sup_{\theta\in\Theta}|\mathbb{G}_n(\theta)|=O_{\mathcal{P}}(1)$ due to asymptotic tightness of $\{\mathbb{G}_n\}$ (uniformly in P) by Lemma D.1 in Bugni, Canay, and Shi (2015), Theorem 3.6.1 and Lemma 1.3.8 in van der Vaart and Wellner (2000), and $\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)|=o_{\mathcal{P}}(\kappa_n/\sqrt{n})$ by Assumption 3.3 (ii) respectively 3.3′ (ii). □

LEMMA B.6: *For any $\theta_n'\in\theta_n+\frac{\rho}{\sqrt{n}}B^d$,*

(i) For any $\eta>0$, there exist $\delta>0$ and $N\in\mathbb{N}$ such that

$$\sup_{c\geq 0}P_n(\{W(\theta_n',c)\neq\emptyset\}\cap\{W^{-\delta}(\theta_n',c)=\emptyset\})<\eta,\ \forall n\geq N. \quad \text{(B.110)}$$

Moreover, for any $\eta>0$, there exist $\delta>0$ and $N\in\mathbb{N}$ such that

$$\sup_{c\geq 0}P_n^*(\{V_n(\theta_n',c)\neq\emptyset\}\cap\{V_n^{-\delta}(\theta_n',c)=\emptyset\})<\eta,\ \forall n\geq N. \quad \text{(B.111)}$$

(ii) Fix $\underline{c}>0$ and redefine

$$W^{-\delta}(\theta_n',c)\equiv\left\{\lambda\in\rho B^d: p'\lambda=0\cap w_{j,\theta_n'}(\lambda)\leq c-\delta,\ \forall j=1,\ldots,J\right\}, \quad \text{(B.112)}$$

and

$$V_n^{-\delta}(\theta_n',c)\equiv\left\{\lambda\in\rho B^d: p'\lambda=0\cap v_{n,j,\theta_n'}(\lambda)\leq c-\delta,\ \forall j=1,\ldots,J\right\}. \quad \text{(B.113)}$$

Then for any $\eta>0$, there exist $\delta>0$ and $N\in\mathbb{N}$ such that

$$\sup_{c\geq\underline{c}}P_n(\{W(\theta_n',c)\neq\emptyset\}\cap\{W^{-\delta}(\theta_n',c)=\emptyset\})<\eta,\ \forall n\geq N, \quad \text{(B.114)}$$

with $W^{-\delta}(\theta_n',c)$ defined in (B.112). Moreover, for any $\eta>0$, there exist $\delta>0$ and $N\in\mathbb{N}$ such

that

$$\sup_{c \geq \underline{c}} P_n^*(\{V_n(\theta_n', c) \neq \emptyset\} \cap \{V_n^{-\delta}(\theta_n', c) = \emptyset\}) < \eta, \ \forall n \geq N, \tag{B.115}$$

with $V_n^{-\delta}(\theta_n', c)$ *defined in* (B.113).

Proof. We first show (B.110). If $\mathcal{J}^* = \emptyset$, with $\mathcal{J}^*$ as defined in (B.37), then the result is immediate. Assume then that $\mathcal{J}^* \neq \emptyset$. Any inequality indexed by $j \notin \mathcal{J}^*$ is satisfied with probability approaching one by similar arguments as in (A.19) (both with c and with $c - \delta$). Hence, one could argue for sets $W(\theta_n', c), W^{-\delta}(\theta_n', c)$ defined as in equations (B.15) and (B.17) but with $j \in \mathcal{J}^*$. To keep the notation simple, below we argue as if all $j = 1, \ldots, J$ belong to $\mathcal{J}^*$.

Let $c \geq 0$ be given. Let $g_{P_n}(\theta_n')$ be a $J + 2d + 2$ vector with entries

$$g_{P_n,j}(\theta_n') = \begin{cases} c - \mathbb{G}_{P_n,j}(\theta_n') - \pi_{1,j}^*, & j = 1, \ldots, J, \\ \rho, & j = J+1, \ldots, J+2d, \\ 0, & j = J+2d+1, J+2d+2, \end{cases} \tag{B.116}$$

recalling that $\pi_{1,j}^* = 0$ for $j = J_1 + 1, \cdots, J$. Let τ be a $(J + 2d + 2)$ vector with entries

$$\tau_j = \begin{cases} 1, & j = 1, \ldots, J_1, \\ 0, & j = J_1 + 1, \ldots, J + 2d + 2. \end{cases} \tag{B.117}$$

Then we can express the sets of interest as

$$W(\theta_n', c) = \{\lambda : K_{P_n}(\theta_n')\lambda \leq g_{P_n}(\theta_n')\}, \tag{B.118}$$

$$W^{-\delta}(\theta_n', c) = \{\lambda : K_{P_n}(\theta_n')\lambda \leq g_{P_n}(\theta_n') - \delta\tau\}. \tag{B.119}$$

By Farkas' Lemma, e.g. Rockafellar (1970, Theorem 22.1), a solution to the system of linear inequalities in (B.118) exists if and only if for all $\mu \in \mathbb{R}_+^{J+2d+2}$ such that $\mu' K_{P_n}(\theta_n') = 0$, one has $\mu' g_{P_n}(\theta_n') \geq 0$. Similarly, a solution to the system of linear inequalities in (B.119) exists if and only if for all $\mu \in \mathbb{R}^{J+2d+2}$ such that $\mu' K_{P_n}(\theta_n') = 0$, one has $\mu'(g_{P_n}(\theta_n') - \delta\tau) \geq 0$. Define

$$\mathcal{M}(\theta_n') \equiv \{\mu \in \mathbb{R}_+^{J+2d+2} : \mu' K_{P_n}(\theta_n') = 0\}. \tag{B.120}$$

Then, one may write

$$\begin{aligned} & P_n(\{W(\theta_n', c) \neq \emptyset\} \cap \{W^{-\delta}(\theta_n', c) = \emptyset\}) \\ =& P_n(\{\mu' g_{P_n}(\theta_n') \geq 0, \forall \mu \in \mathcal{M}(\theta_n')\} \cap \{\mu'(g_{P_n}(\theta_n') - \delta\tau) < 0, \exists \mu \in \mathcal{M}(\theta_n')\}) \\ =& P_n(\{\mu' g_{P_n}(\theta_n') \geq 0, \forall \mu \in \mathcal{M}(\theta_n')\} \cap \{\mu' g_{P_n}(\theta_n') < \delta\mu'\tau, \exists \mu \in \mathcal{M}(\theta_n')\}). \end{aligned} \tag{B.121}$$

Note that the set $\mathcal{M}(\theta_n')$ is a non-stochastic polyhedral cone. Hence, by Minkowski-Weyl's theorem (see, e.g. Rockafellar and Wets (2005, Theorem 3.52)), there exist $\{\nu^t \in \mathcal{M}(\theta_n'), t = 1, \cdots, T\}$, with $T < \infty$ a constant that depends only on J and d, such that any $\mu \in \mathcal{M}(\theta_n')$ can be represented as

$$\mu = b \sum_{t=1}^{T} a_t \nu^t, \tag{B.122}$$

where $b > 0$ and $a_t \geq 0, \ t = 1, \ldots, T, \ \sum_{t=1}^{T} a_t = 1$. Hence, if $\mu \in \mathcal{M}(\theta_n')$ satisfies $\mu' g_{P_n}(\theta_n') < \delta\mu'\tau$,

denoting $\nu^{t\prime}$ the transpose of vector ν^t, we have

$$\sum_{t=1}^{T} a_t \nu^{t\prime} g_{P_n}(\theta_n') < \delta \sum_{t=1}^{T} a_t \nu^{t\prime} \tau. \tag{B.123}$$

However, due to $a_t \geq 0, \forall t$ and $\nu^t \in \mathcal{M}(\theta_n')$, this means $\nu^{t\prime} g_{P_n}(\theta_n') < \delta \nu^{t\prime}\tau$ for some $t \in \{1, \ldots, T\}$. Furthermore, since $\nu^t \in \mathcal{M}(\theta_n')$, we have $0 \leq \nu^{t\prime} g_{P_n}(\theta_n')$. Therefore,

$$\begin{aligned} &P_n(\{\mu' g_{P_n}(\theta_n') \geq 0, \forall \mu \in \mathcal{M}(\theta_n')\} \cap \{\mu' g_{P_n}(\theta_n') < \delta \mu'\tau, \exists \mu \in \mathcal{M}(\theta_n')\}) \\ &\quad \leq P_n(0 \leq \nu^{t\prime} g_{P_n}(\theta_n') < \delta \nu^{t\prime}\tau, \exists t \in \{1, \cdots, T\}) \leq \sum_{t=1}^{T} P_n\left(0 \leq \nu^{t\prime} g_{P_n}(\theta_n') < \delta \nu^{t\prime}\tau\right). \end{aligned} \tag{B.124}$$

Case 1. Consider first any $t = 1, \ldots, T$ such that ν^t assigns positive weight only to constraints in $\{J+1, \ldots, J+2d+2\}$. Then

$$\begin{aligned} \nu^{t\prime} g_{P_n}(\theta_n') &= \rho \sum_{j=J+1}^{J+2d} \nu_j^t, \\ \delta \nu^{t\prime}\tau &= \delta \sum_{j=J+1}^{J+2d+2} \nu_j^t \tau_j = 0, \end{aligned}$$

where the last equality follows by (B.117). Therefore $P_n\left(0 \leq \nu^{t\prime} g_{P_n}(\theta_n') < \delta \nu^{t\prime}\tau\right) = 0$.

Case 2. Consider now any $t = 1, \ldots, T$ such that ν^t assigns positive weight also to constraints in $\{1, \ldots, J\}$. Recall that indexes $j = J_1 + 1, \ldots, J_1 + 2J_2$ correspond to moment equalities, each of which is written as two moment inequalities, therefore yielding a total of $2J_2$ inequalities with $D_{P_n, j+J_2}(\theta_n') = -D_{P_n,j}(\theta_n')$ for $j = J_1 + 1, \ldots, J_1 + J_2$, and:

$$g_{P_n,j}(\theta_n') = \begin{cases} c - \mathbb{G}_{P_n,j}(\theta_n') & j = J_1 + 1, \ldots, J_1 + J_2, \\ c + \mathbb{G}_{P_n, j-J_2}(\theta_n') & j = J_1 + J_2 + 1, \ldots, J. \end{cases} \tag{B.125}$$

For each ν^t, (B.125) implies

$$\sum_{j=J_1+1}^{J_1+2J_2} \nu_j^t g_{P_n,j}(\theta_n') = c \sum_{j=J_1+1}^{J_1+2J_2} \nu_j^t + \sum_{j=J_1+1}^{J_1+J_2} (\nu_j^t - \nu_{j+J_2}^t) \mathbb{G}_{P_n,j}(\theta_n'). \tag{B.126}$$

For each $j = 1, \cdots, J_1 + J_2$, define

$$\tilde{\nu}_j^t \equiv \begin{cases} \nu_j^t & j = 1, \cdots, J_1 \\ \nu_j^t - \nu_{j+J_2}^t & j = J_1 + 1, \cdots, J_1 + J_2. \end{cases}. \tag{B.127}$$

We then let $\tilde{\nu}^t \equiv (\tilde{\nu}_1^t, \cdots, \tilde{\nu}_{J_1+J_2}^t)'$ and have

$$\nu^{t\prime} g_{P_n}(\theta_n') = \sum_{j=1}^{J_1+J_2} \tilde{\nu}_j^t \mathbb{G}_{P_n,j}(\theta_n') + c \sum_{j=1}^{J} \nu_j^t + \sum_{j=1}^{J_1} \nu_j^t \pi_{1,j}^* + \rho \sum_{j=J+1}^{J+2d} \nu_j^t. \tag{B.128}$$

Case 2-a. Suppose $\tilde{\nu}^t \neq 0$. Then, by (B.128), $\frac{\nu^{t\prime} g_{P_n}(\theta_n')}{\nu^{t\prime}\tau}$ is a normal random variable with variance $(\tilde{\nu}^{t\prime}\tau)^{-2} \tilde{\nu}^{\prime t} \Omega_{P_n}(\theta_n') \tilde{\nu}^t$. By Assumption 3.3 (or Assumption 3.3′), there exists a constant $\omega > 0$ that does not depend on θ_n' such that the smallest eigenvalue of $\Omega_{P_n}(\theta_n')$ is bounded from below by ω for

all θ_n'. Hence, letting $\|\cdot\|_p$ denote the p-norm in $\mathbb{R}^{J+2d+2}$, we have

$$\frac{\tilde{\nu}'^t \Omega_{P_n}(\theta_n')\tilde{\nu}^t}{(\tilde{\nu}^{t\prime}\tau)^2} \geq \frac{\omega\|\tilde{\nu}^t\|_2^2}{(J+2d+2)^2\|\tilde{\nu}^t\|_2^2} \geq \frac{\omega}{(J+2d+2)^2}. \tag{B.129}$$

Therefore, the variance of the normal random variable in (B.124) is uniformly bounded away from 0, which in turn allows one to find $\delta > 0$ such that $P_n(0 \leq \frac{\nu^{t\prime} g_{P_n}(\theta_n')}{\nu^{t\prime}\tau} < \delta) \leq \eta/T$.

Case 2-b. Next, consider the case $\tilde{\nu}^t = 0$. Because we are in the case that ν^t assigns positive weight also to constraints in $\{1,\ldots,J\}$, this must be because $\nu_j^t = 0$ for all $j = 1,\cdots,J_1$ and $\nu_j^t = \nu_{j+J_2}^t$ for all $j = J_1+1,\cdots,J_1+J_2$, while $\nu_j^t \neq 0$ for some $j = J_1+1,\cdots,J_1+J_2$. Then we have $\sum_{j=1}^J \nu_j^t g_{P_n,j}(\theta_n') \geq 0$, and $\sum_{j=1}^J \nu_j^t \tau_j = 0$ because $\tau_j = 0$ for each $j = J_1+1,\ldots,J$. Hence, the argument for the case that ν^t assigns positive weight only to constraints in $\{J+1,\ldots,J+2d+2\}$ applies and again $P_n\left(0 \leq \nu^{t\prime} g_{P_n}(\theta_n') < \delta\nu^{t\prime}\tau\right) = 0$. This establishes equation (B.110).

To see why equation (B.111) holds, observe that the bootstrap distribution is conditional on $X_1,\ldots,X_n$, and therefore $\hat{K}_n$ can be treated as non-stochastic, where $\hat{K}_n$ is the matrix in equation (B.18) with $\hat{D}_n$ replacing D_{P_n}. This implies that the set $\hat{\mathcal{M}}_n(\theta_n')$ can also be treated as non-stochastic, where $\hat{\mathcal{M}}_n$ is the set in equation (B.120) with $\hat{K}_n$ replacing K_{P_n}. The result then follows because by Lemma D.2.8 in Bugni, Canay, and Shi (2015), $\mathbb{G}_n^b \xrightarrow{d} \mathbb{G}_{P_n}$ in $l^\infty(\Theta)$ uniformly in $\mathcal{P}$ conditional on $\{X_1,\cdots,X_n\}$.

The results in (B.114)-(B.115) follow by similar arguments, with proper redefinition of τ in equation (B.117). □

LEMMA B.7: *Let (P_n,θ_n) have the almost sure representations given in Lemma B.1, let $\mathcal{J}^*$ be defined as in (B.37), and assume that $\mathcal{J}^* \neq \emptyset$. Let $\widetilde{\mathcal{C}}$ collect all size d subsets C of $\{1,...,J+2d+2\}$ ordered lexicographically by their smallest, then second smallest, etc. elements. Let θ_n^ϵ be as defined in (B.21). Let the random variable $\mathcal{C}(\theta_n^\epsilon)$ equal the first element of $\widetilde{\mathcal{C}}$ s.t. $\det\left(K_P^C(\theta_n^\epsilon)\right) \neq 0$ and $\lambda^C = \left(K_P^C(\theta_n^\epsilon)\right)^{-1} g_P^C(\theta_n^\epsilon) \in W^{*,-\delta}(\theta_n^\epsilon, 0)$ if such an element exists; else, let $\mathcal{C}(\theta_n^\epsilon) = \{J+1,...,J+d\}$. Here $K_P^C(\theta_n^\epsilon)$, $g_P^C(\theta_n^\epsilon)$ and $W^{*,-\delta}(\theta_n^\epsilon,0)$ are as defined in Lemma B.2. Then for any $\eta > 0$, there exist $\alpha_\eta > 0$ and N s.t. $n \geq N$ implies*

$$\mathbf{P}\left\{W^{*,-\delta}(\theta_n^\epsilon,0) \neq \emptyset, \left|\det K_{P_n}^{\mathcal{C}(\theta_n^\epsilon)}(\theta_n^\epsilon)\right| \leq \alpha_\eta\right\} \leq \eta. \tag{B.130}$$

Proof. We establish (B.130) as corollary of the following statement: For each $\eta > 0$, there exist $\alpha_\eta > 0$ and N s.t. $n \geq N$ implies

$$\mathbf{P}\left\{W^{*,-\delta}(\theta_n',0) \neq \emptyset, \left|\det K_{P_n}^{\mathcal{C}(\theta_n')}(\theta_n')\right| \leq \alpha_\eta\right\} \leq \eta$$

for all $\theta_n' \in \theta_n + \frac{\rho}{\sqrt{n}}B^d$. To show this, write

$$\begin{aligned}
& \mathbf{P}\left\{W^{*,-\delta}(\theta_n',0) \neq \emptyset, \left|\det K_{P_n}^{\mathcal{C}(\theta_n')}(\theta_n')\right| \leq \alpha_\eta\right\} \\
\leq\ & \mathbf{P}\left\{\exists C \in \widetilde{\mathcal{C}} : \lambda^C \in \rho B^d, \left|\det K_{P_n}^{\mathcal{C}(\theta_n')}(\theta_n')\right| \leq \alpha_\eta\right\} \\
\leq\ & \sum_{C\in\widetilde{\mathcal{C}}:\left|\det K_{P_n}^C(\theta_n')\right|\leq\alpha_\eta} \mathbf{P}\left(\lambda^C \in \rho B^d\right).
\end{aligned}$$

Here, the first inequality holds because $W^{*,-\delta}(\theta_n',0) \subseteq \rho B^d$ and so the event in the first probability implies the event in the next one; the second inequality is Boolean algebra. Noting that $\widetilde{\mathcal{C}}$ has $\binom{J+2d+2}{d}$

elements, it suffices to show that

$$\left|\det K^C_{P_n}(\theta'_n)\right| \leq \alpha_\eta \Longrightarrow \mathbf{P}\left(\lambda^C \in \rho B^d\right) \leq \overline{\eta} \equiv \frac{\eta}{\binom{J+2d+2}{d}}.$$

Thus, fix $\theta'_n \in \theta_n + \frac{\rho}{\sqrt{n}} B^d$ and $C \in \widetilde{\mathcal{C}}$. To simplify expressions, omit dependence on θ'_n in the remainder of this proof. Let q^C denote the eigenvector associated with the smallest eigenvalue of $K^C_{P_n} K^{C\prime}_{P_n}$ and recall that because $K^C_{P_n} K^{C\prime}_{P_n}$ is symmetric, $\left\|q^C\right\| = 1$. Thus the claim is equivalent to:

$$|q^{C\prime} K^C_{P_n} K^{C\prime}_{P_n} q^C| < \alpha_\eta \implies \mathbf{P}((K^C_{P_n})^{-1} g^C_{P_n} \in \rho B^d) \leq \overline{\eta}. \tag{B.131}$$

Now, if $|q^{C\prime} K^C_{P_n} K^{C\prime}_{P_n} q^C| < \alpha_\eta$ and $(K^C_{P_n})^{-1} g^C_{P_n} \in \rho B^d$, then the Cauchy-Schwarz inequality yields

$$\left|q^{C\prime} g^C_{P_n}\right| = \left|q^{C\prime} K^C_{P_n} \left(K^C_{P_n}\right)^{-1} g^C_{P_n}\right| < \rho\sqrt{d\alpha_\eta}, \tag{B.132}$$

hence

$$\mathbf{P}((K^C_{P_n})^{-1} g^C_{P_n} \in \rho B^d) \leq \mathbf{P}\Big(|q^{C\prime} g^C_{P_n}| < \rho\sqrt{d\alpha_\eta}\Big). \tag{B.133}$$

If q^C assigns non-zero weight only to non-stochastic constraints, then the result follows immediately. Hence, suppose that q^C assigns non-zero weight also to stochastic constraints. Assumptions 3.3 (iii) (or 3.3′ (iii)) and 3.5 (iii) yield that there exists $N \in \mathbb{N}$ and $\omega > 0$ such that for all $n \geq N$ and $\theta'_n \in \theta_n + \frac{\rho}{\sqrt{n}} B^d$,

$$\begin{aligned} & \operatorname{eig}(\tilde{\Omega}_{P_n}) \geq \omega \\ \Longrightarrow\ & Var_{\mathbf{P}}(q^{C\prime} g^C_{P_n}) \geq \omega \\ \Longrightarrow\ & \mathbf{P}\left(|q^{C\prime} g^C_{P_n}| < \rho\sqrt{\alpha_\eta}\right) = \mathbf{P}\left(-\rho\sqrt{\alpha_\eta} < q^{C\prime} g^C_{P_n} < \rho\sqrt{\alpha_\eta}\right) < \frac{2\rho\sqrt{\alpha_\eta}}{\sqrt{2\omega\pi}}, \end{aligned} \tag{B.134}$$

where the result in (B.134) uses that the density of a normal r.v. is maximized at the expected value. The result follows by choosing

$$\alpha_\eta = \frac{\overline{\eta}^2 \omega \pi}{2\rho^2}.$$

□

LEMMA B.8: *If $J_2 \geq d$, then $\exists \underline{c} > 0$ s.t. $\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(\hat{c}_n(\theta) \geq \underline{c}) = 1$.*

Proof. Fix any $c \geq 0$ and restrict attention to constraints $\{J_1{+}1, ..., J_1{+}d, J_1{+}J_2{+}1, ..., J_1{+}J_2{+}d\}$, i.e. the inequalities that jointly correspond to the first d equalities. We separately analyze the case when (i) the corresponding estimated gradients $\{\hat{D}_{n,j}(\theta) : j = J_1 + 1, ..., J_1 + d\}$ are linearly independent and (ii) they are not. If $\{\hat{D}_{n,j}(\theta) : j = J_1 + 1, ..., J_1 + d\}$ converge to linearly independent limits, then only the former case occurs infinitely often; else, both may occur infinitely often, and we conduct the argument along two separate subsequences if necessary.

For the remainder of this proof, because the sequence $\{\theta_n\}$ is fixed and plays no direct role in the proof, we suppress dependence of $\hat{D}_{n,j}(\theta)$ and $\mathbb{G}^b_{n,j}(\theta)$ on θ. Also, if C is an index set picking certain constraints, then $\hat{D}^C_n$ is the matrix collecting the corresponding estimated gradients, and similarly for $\mathbb{G}^{b,C}_n$.

Suppose now case (i), then there exists an index set $\bar{C} \subset \{J_1{+}1, ..., J_1{+}d, J_1{+}J_2{+}1, \ldots, J_1{+}J_2{+}d\}$ picking one direction of each constraint s.t. p is a positive linear combination of the rows of $\hat{D}^{\bar{C}}_P$. (This choice ensures that a Karush-Kuhn-Tucker condition holds, justifying the step from (B.136) to (B.137)

below.) Then the bootstrap coverage probability induced by c is asymptotically bounded above by

$$P^*\left(\sup\left\{p'\lambda : \hat{D}_{n,j}\lambda \leq c - \mathbb{G}^b_{n,j}, j \in \mathcal{J}^*\right\} \geq 0\right) \tag{B.135}$$

$$\leq P^*\left(\sup\left\{p'\lambda : \hat{D}_{n,j}\lambda \leq c - \mathbb{G}^b_{n,j}, j \in \bar{C}\right\} \geq 0\right) \tag{B.136}$$

$$= P^*\left(\sup\left\{p'\lambda : \hat{D}_{n,j}\lambda = c - \mathbb{G}^b_{n,j}, j \in \bar{C}\right\} \geq 0\right) \tag{B.137}$$

$$= P^*\left(p'(\hat{D}^{\bar{C}}_n)^{-1}(c\mathbf{1}_d - \mathbb{G}^{b,\bar{C}}_n) \geq 0\right) \tag{B.138}$$

$$= P^*\left(\frac{p'(\hat{D}^{\bar{C}}_n)^{-1}(c\mathbf{1}_d - \mathbb{G}^{b,\bar{C}}_n)}{\sqrt{p'(\hat{D}^{\bar{C}}_n)^{-1}\Omega^C_P(\hat{D}^{\bar{C}}_n)^{-1}p}} \geq 0\right) \tag{B.139}$$

$$= P^*\left(\frac{p'adj(\hat{D}^{\bar{C}}_n)(c\mathbf{1}_d - \mathbb{G}^{b,\bar{C}}_n)}{\sqrt{p'(adj(\hat{D}^{\bar{C}}_n)\Omega^C_P adj(\hat{D}^{\bar{C}}_n)p}} \geq 0\right) \tag{B.140}$$

$$= \Phi\left(\frac{p'adj(\hat{D}^{\bar{C}}_n)c\mathbf{1}_d}{\sqrt{p'(adj(\hat{D}^{\bar{C}}_n)\Omega^C_P adj(\hat{D}^{\bar{C}}_n)p}}\right) + o_{\mathcal{P}}(1) \tag{B.141}$$

$$\leq \Phi\left(d\omega^{-1/2}c\right) + o_{\mathcal{P}}(1). \tag{B.142}$$

Here, (B.136) removes constraints and hence enlarges the feasible set; (B.137) uses that by construction, the remaining problem is solved at the intersection of its constraints; (B.138) solves in closed form; (B.139) divides through by a positive scalar; (B.140) eliminates the determinant of $\hat{D}^{\bar{C}}_n$, using that rows of $\hat{D}^{\bar{C}}_n$ can always be rearranged so that the determinant is positive; (B.141) follows by Assumption 3.5, using that the term multiplying $\mathbb{G}^{b,\bar{C}}_n$ is $O_{\mathcal{P}}(1)$; and (B.142) uses that by Assumption 3.3 (iii) (or Assumption 3.3′ (iii-2)), there exists a constant $\omega > 0$ that does not depend on θ such that the smallest eigenvalue of Ω_P is bounded from below by ω. The result follows for any choice of $\underline{c} \in (0, \Phi^{-1}(1-\alpha) \times \omega^{1/2}/d)$.

In case (ii), there exists an index set $\bar{C} \subset \{J_1+2, ..., J_1+d, J_1+J_2+2, ..., J_1+J_2+d\}$ collecting $d-1$ or fewer linearly independent constraints s.t. $\hat{D}_{n,J_1+1}$ is a positive linear combination of the rows of $\hat{D}^{\bar{C}}_P$. (Note that $\bar{C}$ cannot contain J_1+1 or J_1+J_2+1.) One can then write

$$P^*\left(\sup\left\{p'\lambda : \hat{D}_{n,j}\lambda \leq c - \mathbb{G}^b_{n,j}, j \in \bar{C} \cup \{J_1+J_2+1\}\right\} \geq 0\right) \tag{B.143}$$

$$\leq P^*\left(\exists\lambda : \hat{D}_{n,j}\lambda \leq c - \mathbb{G}^b_{n,j}, j \in \bar{C} \cup \{J_1+J_2+1\}\right) \tag{B.144}$$

$$\leq P^*\left(\sup\left\{\hat{D}_{n,J_1+1}\lambda : \hat{D}_{n,j}\lambda \leq c - \mathbb{G}^b_{n,j}, j \in \bar{C}\right\} \geq \inf\left\{\hat{D}_{n,J_1+1}\lambda : \hat{D}_{n,J_1+J_2+1}\lambda \leq c - \mathbb{G}^b_{n,J_1+J_2+1}\right\}\right) \tag{B.145}$$

$$= P^*\left(\hat{D}_{n,J_1+1}\hat{D}^{\bar{C}\prime}_n(\hat{D}^{\bar{C}}_n\hat{D}^{\bar{C}\prime}_n)^{-1}(c\mathbf{1}_{\bar{d}} - \mathbb{G}^{b,\bar{C}}_n) \geq -c + \mathbb{G}^b_{n,J_1+J_2+1}\right). \tag{B.146}$$

Here, the reasoning from (B.143) to (B.145) holds because we evaluate the probability of increasingly larger events; in particular, if the event in (B.145) fails, then the constraint sets corresponding to the sup and inf can be separated by a hyperplane with gradient $\hat{D}_{n,J_1+1}$ and so cannot intersect. The last step solves the optimization problems in closed form, using (for the sup) that a Karush-Kuhn-Tucker condition again holds by construction and (for the inf) that $\hat{D}_{n,J_1+J_2+1} = -\hat{D}_{n,J_1+1}$. Expression (B.146) resembles (B.139), and the argument can be concluded in analogy to (B.140)-(B.142). □

LEMMA B.9: *Suppose that both $\pi_{1,j}$ and $\pi_{1,j+J_{11}}$ are finite, with $\pi_{1,j}$, $j = 1, \ldots, J$, defined in*

(A.4). *Then for any* $\theta'_n \in \theta_n + \frac{\rho}{\sqrt{n}} B^d$,

(1) $\sigma^2_{P_n,j}(\theta'_n)/\sigma^2_{P_n,j+J_{11}}(\theta'_n) \to 1$ *for* $j = 1, \cdots, J_{11}$.

(2) $Corr_{P_n}(m_j(X_i,\theta'_n), m_{j+J_{11}}(X_i,\theta'_n)) \to -1$ *for* $j = 1, \cdots, J_{11}$.

(3) $\mathbb{G}^*_j(\theta'_n) + \mathbb{G}^*_{j+J_{11}}(\theta'_n) \to 0$ *almost surely.*

(4) $\|D_{P_n,j+J_{11}}(\theta'_n) + D_{P_n,j}(\theta'_n)\| \to 0$.

Proof. By Lemma B.5, for each j, $\lim_{n\to\infty} \kappa_n^{-1} \frac{\sqrt{n} E_{P_n}[m_j(X_i,\theta'_n)]}{\sigma_{P_n,j}(\theta'_n)} = \pi_{1,j}$, and hence the condition that $\pi_{1,j}, \pi_{1,j+J_{11}}$ are finite is inherited by the limit of the corresponding sequences $\kappa_n^{-1} \frac{\sqrt{n} E_{P_n}[m_j(X_i,\theta'_n)]}{\sigma_{P_n,j}(\theta'_n)}$ and $\kappa_n^{-1} \frac{\sqrt{n} E_{P_n}[m_{j+J11}(X_i,\theta'_n)]}{\sigma_{P_n,j+J11}(\theta'_n)}$.

We first establish Claims 1 and 2. We consider two cases.

Case 1.

$$\lim_{n\to\infty} \frac{\kappa_n}{\sqrt{n}} \sigma_{P_n,j}(\theta'_n) > 0, \tag{B.147}$$

which implies that $\sigma_{P_n,j}(\theta'_n) \to \infty$ at rate $\sqrt{n}/\kappa_n$ or faster. Claim 1 then holds because

$$\frac{\sigma^2_{P_n,j+J_{11}}(\theta'_n)}{\sigma^2_{P_n,j}(\theta'_n)} = \frac{\sigma^2_{P_n,j}(\theta'_n) + Var_{P_n}(t_j(X_i,\theta'_n)) + 2Cov_{P_n}(m_j(X_i,\theta'_n), t_j(X_i,\theta'_n))}{\sigma^2_{P_n,j}(\theta'_n)} \to 1, \tag{B.148}$$

where the convergence follows because $Var_{P_n}(t_j(X_i,\theta'_n))$ is bounded due to Assumption 3.3′ (iii-1),

$$|Cov_{P_n}(m_j(X_i,\theta'_n), t_j(X_i,\theta'_n))/\sigma^2_{P_n,j}(\theta'_n)| \le (Var_{P_n}(t_j(X_i,\theta'_n)))^{1/2}/\sigma_{P_n,j}(\theta'_n),$$

and the fact that $\sigma_{P_n,j}(\theta'_n) \to \infty$. A similar argument yields Claim 2.

Case 2.

$$\lim_{n\to\infty} \frac{\kappa_n}{\sqrt{n}} \sigma_{P_n,j}(\theta'_n) = 0. \tag{B.149}$$

In this case, $\pi_{1,j}$ being finite implies that $E_{P_n} m_j(X_i,\theta'_n) \to 0$. Again using the upper bound on $t_j(X_i,\theta'_n)$ similarly to (B.148), it also follows that

$$\lim_{n\to\infty} \frac{\kappa_n}{\sqrt{n}} \sigma_{P_n,j+J_{11}}(\theta'_n) = 0, \tag{B.150}$$

and hence that $E_{P_n}(t_j(X_i,\theta'_n)) \to 0$. We then have, using Assumption 3.3′ (iii-1) again,

$$\begin{aligned} Var_{P_n}(t_j(X_i,\theta'_n)) &= \int t_j(x,\theta'_n)^2 dP_n(x) - E_{P_n}[t_j(X_i,\theta'_n)]^2 \\ &\le M \int t_j(x,\theta'_n) dP_n(x) - E_{P_n}[t_j(X_i,\theta'_n)]^2 \to 0. \end{aligned} \tag{B.151}$$

Hence,

$$\begin{aligned} \frac{\sigma^2_{P_n,j+J_{11}}(\theta'_n)}{\sigma^2_{P_n,j}(\theta'_n)} &= \frac{\sigma^2_{P_n,j}(\theta'_n) + Var_{P_n}(t_j(X_i,\theta'_n)) + 2Cov_{P_n}(m_j(X_i,\theta'_n), t_j(X_i,\theta'_n))}{\sigma^2_{P_n,j}(\theta'_n)} \\ &\le \frac{\sigma^2_{P_n,j}(\theta'_n) + Var_{P_n}(t_j(X_i,\theta'_n))}{\sigma^2_{P_n,j}(\theta'_n)} + \frac{2(Var_{P_n}(t_j(X_i,\theta'_n)))^{1/2}}{\sigma_{P_n,j}(\theta'_n)} \\ &\to 1, \end{aligned} \tag{B.152}$$

and the first claim follows.

To obtain claim 2, note that

$$Corr_{P_n}(m_j(X_i,\theta'_n), m_{j+J_{11}}(X_i,\theta'_n)) = \frac{-\sigma^2_{P_n,j}(\theta'_n) - Cov_{P_n}(m_j(X_i,\theta'_n), t_j(X_i,\theta'_n))}{\sigma_{P_n,j}(\theta'_n)\sigma_{P_n,j+J_{11}}(\theta'_n)}$$
$$\to -1, \tag{B.153}$$

where the result follows from (B.151) and (B.152).

To establish Claim 3, consider $\mathbb{G}_n$ below. Note that, for $j = 1, \cdots, J_{11}$,

$$\begin{bmatrix} \mathbb{G}_{n,j}(\theta'_n) \\ \mathbb{G}_{n,j+J_{11}}(\theta'_n) \end{bmatrix} \tag{B.154}$$
$$= \begin{bmatrix} \frac{1}{\sqrt{n}} \frac{\sum_{i=1}^n (m_j(X_i,\theta'_n) - E_{P_n}[m_j(X_i,\theta'_n)])}{\sigma_{P_n,j}(\theta'_n)} \\ -\frac{1}{\sqrt{n}} \frac{\sum_{i=1}^n (m_j(X_i,\theta'_n) - E_{P_n}[m_j(X_i,\theta'_n)]) + \frac{1}{\sqrt{n}} \sum_{i=1}^n (t_j(X_i,\theta'_n) - E_{P_n}[t_j(X_i,\theta'_n)])}{\sigma_{P_n,j+J_{11}}(\theta'_n)} \end{bmatrix}. \tag{B.155}$$

Under the conditions of Case 1 above, we immediately obtain

$$|\mathbb{G}_{n,j}(\theta'_n) + \mathbb{G}_{n,j+J_{11}}(\theta'_n)| \xrightarrow{p} 0. \tag{B.156}$$

Under the conditions in Case 2 above, $\frac{1}{\sqrt{n}} \sum_{i=1}^n (t_j(X_i,\theta'_n) - E_{P_n}[t_j(X_i,\theta'_n)] = o_{\mathcal{P}}(1)$ due to the variance of this term being equal to $Var_{P_n}(t_j(X_i,\theta'_n)) \to 0$ and Chebyshev's inequality. Therefore, (B.156) obtains again.

Note that $\mathbb{G}_n$ has an asymptotic almost sure representation such that $\mathbb{G}^*_n \xrightarrow{a.s.} \mathbb{G}^*$ in $\ell^\infty(\Theta)$. This therefore implies

$$|\mathbb{G}^*_j(\theta'_n) + \mathbb{G}^*_{j+J_{11}}(\theta'_n)| \le |\mathbb{G}^*_j(\theta'_n) - \mathbb{G}^*_{n,j}(\theta'_n)|$$
$$+ |\mathbb{G}^*_{n,j}(\theta'_n) + \mathbb{G}^*_{n,j+J_{11}}(\theta'_n)| + |\mathbb{G}^*_{j+J_{11}}(\theta'_n) - \mathbb{G}^*_{n,j+J_{11}}(\theta'_n)| \to 0, \tag{B.157}$$

with probability 1 (under $\mathbf{P}$) where the convergence is due to $\mathbb{G}^*_n \xrightarrow{a.s.} \mathbb{G}^*$ and $|\mathbb{G}^*_{n,j}(\theta'_n) + \mathbb{G}^*_{n,j+J_{11}}(\theta'_n)| \to 0$ with probability 1 implied by (B.156) and $\mathbb{G}^*_n \stackrel{d}{=} \mathbb{G}_n$.

To establish Claim 4, finiteness of $\pi_{1,j}$ and $\pi_{1,j+J_{11}}$ implies that

$$E_{P_n}\left(\frac{m_j(X,\theta'_n)}{\sigma_{P_n,j}(\theta'_n)} + \frac{m_{j+J_{11}}(X,\theta'_n)}{\sigma_{P_n,j+J_{11}}(\theta'_n)}\right) = O_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}}). \tag{B.158}$$

Suppose by contradiction that

$$D_{P_n,j+J_{11}}(\theta'_n) + D_{P_n,j}(\theta'_n) \to q \neq 0. \tag{B.159}$$

Write

$$\tilde{r} = \arg \max_{s:\|s\|=1} qs, \tag{B.160}$$

yielding $q\tilde{r} > 0$. Let

$$r_n = \tilde{r}\kappa_n^2/\sqrt{n}. \tag{B.161}$$

Using a mean value expansion (where $\bar{\theta}_n$ and $\tilde{\theta}_n$ in the expressions below are two potentially different

vectors that lie component-wise between θ_n' and $\theta_n' + r_n$) we obtain

$$
\begin{aligned}
&E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+J_{11}}(X,\theta_n'+r_n)}{\sigma_{P_n,j+J_{11}}(\theta_n'+r_n)}\right)\\
=&E_{P_n}\left(\frac{m_j(X,\theta_n')}{\sigma_{P_n,j}(\theta_n')}+\frac{m_{j+J_{11}}(X,\theta_n')}{\sigma_{P_n,j+J_{11}}(\theta_n')}\right)+\left(D_{P_n,j}(\bar\theta_n)+D_{P_n,j+J_{11}}(\tilde\theta_n)\right)r_n\\
=&O_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}})+\left(D_{P_n,j}(\theta_n')+D_{P_n,j+J_{11}}(\theta_n')\right)r_n+\left(D_{P_n,j}(\bar\theta_n)-D_{P_n,j}(\theta_n')\right)r_n+\left(D_{P_n,j+J_{11}}(\tilde\theta_n)-D_{P_n,j+J_{11}}(\theta_n')\right)r_n\\
=&O_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}})+q\tilde r\frac{\kappa_n^2}{\sqrt{n}}+O_{\mathcal{P}}(\frac{\kappa_n^4}{n}).
\end{aligned}
\tag{B.162}
$$

It then follows that there exists $N \in \mathbb{N}$ such that for all $n \geq N$, the right hand side in (B.162) is strictly greater than zero.

Next, observe that

$$
\begin{aligned}
&E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+J_{11}}(X,\theta_n'+r_n)}{\sigma_{P_n,j+J_{11}}(\theta_n'+r_n)}\right)\\
=&E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+J_{11}}(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}\right)-\left(\frac{\sigma_{P_n,j+J_{11}}(\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}-1\right)\frac{E_{P_n}(m_{j+J_{11}}(X,\theta_n'+r_n))}{\sigma_{P_n,j+J_{11}}(\theta_n'+r_n)}\\
=&E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+J_{11}}(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}\right)-o_{\mathcal{P}}(\frac{\kappa_n^2}{\sqrt{n}}).
\end{aligned}
\tag{B.163}
$$

Here, the last step is established as follows. First, using that $\sigma_{P_n,j}(\theta_n'+r_n)$ is bounded away from zero for n large enough by the continuity of $\sigma(\cdot)$ and Assumption 3.3′, we have

$$
\frac{\sigma_{P_n,j+J_{11}}(\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}-1=\frac{\sigma_{P_n,j+J_{11}}(\theta_n')}{\sigma_{P_n,j}(\theta_n')}-1+o_{\mathcal{P}}(1)=o_{\mathcal{P}}(1),
\tag{B.164}
$$

where we used Claim 1. Second, using Assumption 3.4, we have that

$$
\frac{E_{P_n}(m_{j+J_{11}}(X,\theta_n'+r_n))}{\sigma_{P_n,j+J_{11}}(\theta_n'+r_n)}=\frac{E_{P_n}(m_{j+J_{11}}(X,\theta_n'))}{\sigma_{P_n,j+J_{11}}(\theta_n')}+D_{P_n,j+J_{11}}(\tilde\theta_n)r_n=o_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}})+O_{\mathcal{P}}(\frac{\kappa_n^2}{\sqrt{n}}).
\tag{B.165}
$$

The product of (B.164) and (B.165) is therefore $o_{\mathcal{P}}(\frac{\kappa_n^2}{\sqrt{n}})$ and (B.163) follows.

To conclude the argument, note that for n large enough, $m_{j+J_{11}}(X,\theta_n'+r_n) \leq -m_j(X,\theta_n'+r_n)$ *a.s.* because for any $\theta_n' \in \Theta_I(P_n)+\rho/\sqrt{n}B^d$ by Assumption 3.3′ (i) for n large enough, $\theta_n'+r_n \in \Theta$ and Assumption 3.3′ (iii-1) applies. Therefore, there exists $N \in \mathbb{N}$ such that for all $n \geq N$, the left hand side in (B.162) is strictly less than the right hand side, yielding a contradiction. □

Appendix C Auxiliary Lemmas

LEMMA C.1: *The event*

$$
\max_{\lambda\in\Lambda_n^b(\theta,\rho,c)} p'\lambda \geq 0 \geq \min_{\lambda\in\Lambda_n^b(\theta,\rho,c)} p'\lambda
$$

with $\Lambda_n^b(\theta,\rho,c)$ *defined in equation* (2.7)*, is equivalent to the event*

$$
\Lambda_n^b(\theta,\rho,c)\cap\{p'\lambda=0\}\neq\emptyset.
\tag{C.1}
$$

Proof. "If" is immediate. To see "only if," note that if the first event obtains, then there exist $\underline{\lambda},\bar\lambda \in \Lambda_n^b(\theta,\rho,c)$ with $p'\bar\lambda \geq 0 \geq p'\underline{\lambda}$. If either $p'\bar\lambda = 0$ or $p'\underline{\lambda} = 0$, the result follows. Consider the

case that both are different from zero. As $\Lambda_n^b(\theta,\rho,c)$ is convex, it follows that

$$\frac{-p'\underline{\lambda}}{p'\bar{\lambda}-p'\underline{\lambda}}\bar{\lambda}+\frac{p'\bar{\lambda}}{p'\bar{\lambda}-p'\underline{\lambda}}\underline{\lambda}\in\Lambda_n^b(\theta,\rho,c)$$

and hence the claim. □

LEMMA C.2: *Fix $\theta\in\Theta$, $P\in\mathcal{P}$ and ρ. Suppose Assumptions 3.1, 3.2, 3.3 or 3.3′, 3.4 and 3.5 hold and also that $\varphi_j(x)\leq 0$ for all x and j. Let $0<\delta<\rho$. With a modification of notation, explicitly highlight $\hat{c}_n(\theta)$'s dependence on ρ through the notation $\hat{c}_n(\theta,\rho)$. Then*

$$|\hat{c}_n(\theta,\rho)-\hat{c}_n(\theta,\rho-\delta)|\xrightarrow{p}0 \tag{C.2}$$

if and only if $D_{P,j}(\theta)/\|D_{P,j}(\theta)\|\in\{p,-p\}$ for all $j\in\mathcal{J}^*(\theta)\equiv\{j:E_P[m_j(X_i,\theta)]\geq 0\}$.

REMARK C.1: The lemma applies to any increase or decrease of ρ. The claims about $\hat{c}_n^{AS}(\theta)$ are implied because in the lemma's notation, $\hat{c}_n^{AS}(\theta)=\hat{c}_n(\theta,0)$.

REMARK C.2: For θ such that $\mathcal{J}^*(\theta)=\emptyset$, we have $\hat{c}_n(\theta,\rho)\xrightarrow{p}0$ but also $\hat{c}_n^{AS}(\theta)\xrightarrow{p}0$. This is consistent with Lemma C.2 because the condition on gradients vacuously holds in this case.

Proof. Recall that θ and P are fixed, i.e. we assume a pointwise perspective. Then

$$\hat{c}_n(\theta,\rho)\xrightarrow{p}\inf\{c\geq 0:P(\{\lambda\in\rho B^d:\mathbb{G}_{P,j}(\theta)+D_{P,j}(\theta)\lambda\leq c,j\in\mathcal{J}^*(\theta)\}\cap\{p'\lambda=0\}\neq\emptyset)\geq 1-\alpha\}. \tag{C.3}$$

Here, we used convergence of $\mathbb{G}_j^b(\theta)$ to $\mathbb{G}_{P,j}(\theta)$ and of $\hat{D}_j(\theta)$ to $D_{P,j}(\theta)$, boundedness of gradients, and the fact that

$$\varphi_j(\kappa_n^{-1}\sqrt{n}\bar{m}_j(X_i,\theta)/\sigma_{P,j}(\theta))\xrightarrow{p}\begin{cases}0 & \text{if } j\in\mathcal{J}^*(\theta)\\ -\infty & \text{otherwise,}\end{cases} \tag{C.4}$$

where the first of those cases uses nonpositivity of φ_j. It therefore suffices to show that the right hand side of C.3 strictly decreases in ρ if and only if the conditions of the Lemma hold.

To simplify notation, henceforth omit dependence of $\mathbb{G}_{P,j}(\theta)$, $D_P(\theta)$, and $\mathcal{J}^*(\theta)$ on P and θ. Define the J vector e to have elements $e_j=c-\mathbb{G}_j,\ j=1,\ldots,J$. Suppose for simplicity that $\mathcal{J}^*$ contains the first J^* inequality constraints. Let $e^{[1:J^*]}$ denote the subvector of e that only contains elements corresponding to $j\in\mathcal{J}^*$, define $D^{[1:J^*,:]}$ correspondingly, and write

$$K=\begin{bmatrix}D^{[1:J^*,:]}\\ I_d\\ -I_d\\ p'\\ -p'\end{bmatrix},\ g=\begin{bmatrix}e^{[1:J^*]}\\ \rho\cdot\mathbf{1}_d\\ \rho\cdot\mathbf{1}_d\\ 0\\ 0\end{bmatrix},\ \tau=\begin{bmatrix}0\cdot\mathbf{1}_{J^*}\\ \mathbf{1}_d\\ \mathbf{1}_d\\ 0\\ 0\end{bmatrix}.$$

where I_d denotes the $d\times d$ identity matrix. By Farkas' Lemma (Rockafellar, 1970, Theorem 22.1), the linear system $K\lambda\leq g$ has a solution if and only if for all $\mu\in\mathbb{R}_+^{J^*+2d+2}$,

$$\mu'K=0\Rightarrow\mu'g\geq 0. \tag{C.5}$$

To further simplify expressions, fix $p=[1\ 0\ \ldots\ 0]$. Let $\mathcal{M}=\{\mu\in\mathbb{R}_+^{J^*+2d+2}:\mu'K=0\}$.

Step 1. This step shows that

$$\begin{aligned} & P(\{\lambda \in \rho B^d : \mathbb{G}_{P,j} + D_{P,j}\lambda \leq c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \emptyset) \\ > \; & P(\{\lambda \in (\rho - \delta) B^d : \mathbb{G}_{P,j} + D_{P,j}\lambda \leq c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \emptyset) \end{aligned} \tag{C.6}$$

if and only if the condition on gradients holds. This is done by showing that

$$P(\{\mu' g \geq 0 \; \forall \mu \in \mathcal{M}\} \cap \{\mu' g - \delta\tau < 0 \; \exists \mu \in \mathcal{M}\}) > 0. \tag{C.7}$$

under that same condition. The event $\{\mu' g \geq 0 \; \forall \mu \in \mathcal{M}\}$ obtains if and only if

$$\min_{\mu \in \mathbb{R}_+^{J^*+2d+2}} \{\mu' g : \; \mu' K = 0\} \geq 0 \tag{C.8}$$

and analogously for $\mu'(g - \delta\tau) \geq 0$. The values of these programs are not affected by adding a constraint as follows:

$$\min_{\mu \in \mathbb{R}_+^{J^*+2d+2}} \left\{ \mu' g : \; \mu' K = 0, \; \mu \in \arg \min_{\tilde{\mu} \in \mathbb{R}_+^{J^*+2d+2}} (\tilde{\mu}' g : \; \tilde{\mu}^{[1:J^*]} = \mu^{[1:J^*]}, \tilde{\mu}' K = 0) \right\}, \tag{C.9}$$

That is, we can restrict attention to a concentrated out subset of vectors μ, where the last $(2d+2)$ components of any μ minimize the objective function among all vectors that agree with μ in the first J^* components. The inner minimization problem in equation (C.9) can be written as

$$\min_{\tilde{\mu}^{[J^*+1:J^*+2d+2]} \in \mathbb{R}_+^{2d+2}} \rho \sum_{j=J^*+1}^{J^*+2d} \tilde{\mu}_j \quad \text{s.t.} \quad \begin{bmatrix} \tilde{\mu}_{J^*+1} - \tilde{\mu}_{J^*+d+1} + \tilde{\mu}_{J^*+2d+1} - \tilde{\mu}_{J^*+2d+2} \\ \tilde{\mu}_{J^*+2} - \tilde{\mu}_{J^*+d+2} \\ \vdots \\ \tilde{\mu}_{J^*+d} - \tilde{\mu}_{J^*+2d} \end{bmatrix} = -\mu^{[1:J^*]\prime} D^{[1:J^*,:]}. \tag{C.10}$$

Thus, the solution of the problem is uniquely pinned down as

$$\mu^{[J^*+1:J^*+2d+2]} = \begin{bmatrix} 0 \\ -\left[D^{[1:J^*,2:d]\prime} \mu^{[1:J^*]} \wedge 0 \cdot \mathbf{1}_{d-1}\right] \\ 0 \\ D^{[1:J^*,2:d]\prime} \mu^{[1:J^*]} \vee 0 \cdot \mathbf{1}_{d-1} \\ -\left[D^{[1:J^*,1]\prime} \mu^{[1:J^*]} \wedge 0\right] \\ D^{[1:J^*,1]\prime} \mu^{[1:J^*]} \vee 0 \end{bmatrix}, \tag{C.11}$$

where $D^{[1:J^*,2:d]\prime} \mu^{[1:J^*]} \vee 0 \cdot \mathbf{1}_{d-1}$ indicates a component-wise comparison. Now we consider the following case distinction:

Case (i). If $D_j / \|D_j\| \in \{p, -p\}$ for all $j \in \mathcal{J}^*$, then $\mu^{[1:J^*]\prime} D = (\mu^{[1:J^*]\prime} D^{[1:J^*,1]}, 0, ..., 0)'$ and therefore all but the last two entries of $\mu^{[J^*+1:J^*+2d+2]}$ equal zero. One can, therefore, restrict attention to vectors μ with $\mu^{[J^*+1:J^*+2d]} = 0$. But for these vectors, $\mu'\tau = 0$ and so the programs we compare necessarily have the same value. The probability in equation (C.7) is therefore zero.

Case (ii). Suppose that at least one row of D, say its first row (though it can be one direction of an equality constraint), is not collinear with p, so that $\|D^{[1,2:d]}\| \neq 0$.

Let

$$\varpi = \begin{bmatrix} 1 \\ 0 \cdot \mathbf{1}_{J^*-1} \\ 0 \\ -\left[(D^{[1,2:d]\prime}) \wedge 0 \cdot \mathbf{1}_{d-1}\right] \\ 0 \\ (D^{[1,2:d]\prime}) \vee 0 \cdot \mathbf{1}_{d-1} \\ -\left[(D^{[1,1]}) \wedge 0\right] \\ (D^{[1,1]}) \vee 0 \end{bmatrix} \tag{C.12}$$

and note that $\varpi^{[J^*+1:J^*+2d]} \neq 0$, hence $\varpi'\tau > 0$.

As in the proof of Lemma B.6, the set $\mathcal{M}$ can be expressed as positive span of a finite, nonstochastic set of affinely independent vectors $\nu^t \in \mathbb{R}_+^{J^*+2d+2}$ that are determined only up to multiplication by a positive scalar. All of these vectors have the "concentrated out structure" in equation (C.11). But then ϖ must be one of them because it is the unique concentrated out vector with $\varpi^{[1:J^*]} = (1, 0, ..., 0)'$, and $(1, 0, ..., 0)'$ cannot be spanned by nonnegative J^*-vectors other than positive multiples of itself.

We now establish positive probability of the event

$$\begin{aligned} \nu^{t\prime} g &\geq 0, \text{ all } \nu^t \\ \nu^{t\prime}(g - \delta\tau) &< 0, \text{ some } \nu^t \end{aligned}$$

by observing that if we define

$$\iota_k = \begin{bmatrix} -\rho \cdot \sum_{i=2}^{d} \left|D^{[1,i]}\right| \\ k \cdot \mathbf{1}_{J^*-1} \\ \rho \cdot \mathbf{1}_d \\ \rho \cdot \mathbf{1}_d \\ 0 \\ 0 \end{bmatrix}, \tag{C.13}$$

then we have

$$0 = \varpi'\iota_k = \min_t \nu^{t\prime}\iota_k.$$

Any other spanning vector ν^t will not have $\varpi^{[2:J^*]} = 0$ and so for any such vector, $\nu^{t\prime}\iota_k$ strictly increases in k. As there are finitely many spanning vectors, all of them have strictly positive inner product with ι_k if k is chosen large enough.

A realization of $g = \iota_k$ would, therefore, yield

$$\nu^{t\prime} g \geq 0 \ \forall \nu^t \in \mathcal{M}, \text{ and } \varpi^{t\prime}(g - \delta\tau) < -\epsilon, \tag{C.14}$$

for some $\epsilon > 0$. Let

$$\Gamma_k = \{\iota : \iota = \iota_k + \epsilon/2b, \ \|b\| \leq 1 \text{ and } \varpi' b > 0\}. \tag{C.15}$$

Then

$$\nu^{t\prime}\iota \geq 0 \ \forall \nu^t \in \mathcal{M}, \text{ and } \varpi^{t\prime}(\iota - \delta\tau) < -\epsilon/2, \ \forall \iota \in \Gamma_k. \tag{C.16}$$

The probability in equation (C.7) is therefore strictly positive.

Step 2. Next, we argue that

$$P(\{\lambda \in \rho B^d : \ \mathbb{G}_j + D_j\lambda \leq c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \emptyset) \tag{C.17}$$

strictly continuously increases in c. The rigorous argument is very similar to the use of Farkas' Lemma in step 1 and in Lemma B.6. We leave it at an intuition: As c increases, the set of vectors g fulfilling the right hand side of (C.5) strictly increases, hence the set of realizations of $\mathbb{G}_j$ that render the program feasible strictly increases, and $\mathbb{G}_j$ has full support.

Step 3. Steps 1 and 2 imply that

$$\begin{aligned} & \inf_{c\geq 0}\{P(\{\lambda \in \rho B^d : \mathbb{G}_j + D_j\lambda \leq c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \emptyset) \geq 1-\alpha\} \\ > \; & \inf_{c\geq 0}\{P(\{\lambda \in (\rho-\delta) B^d : \mathbb{G}_j + D_j\lambda \leq c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \emptyset) \geq 1-\alpha\} \end{aligned} \tag{C.18}$$

and hence the result. □

References

Andrews, D. W. K., and P. J. Barwick (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80(6), 28052826.

Andrews, D. W. K., S. T. Berry, and P. Jia (2004): "Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location," mimeo.

Andrews, D. W. K., and X. Cheng (2012): "Estimation and Inference with Weak, Semi-Strong, and Strong Identification," *Econometrica*, 80, 2153–2211.

Andrews, D. W. K., and P. Guggenberger (2009): "Validity of Subsampling and 'Plug-In Asymptotic' Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25(3), 669–709.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Bahadur, R. R., and L. J. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 27(4), 1115–1122.

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

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set Identified Linear Models," *Econometrica*, 80, 1129–1155.

Bugni, F. A. (2009): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78(2), 735–753.

Bugni, F. A., I. A. Canay, and X. Shi (2014): "Inference for Functions of Partially Identified Parameters in Moment Inequality Models," Working Paper CWP22/14, CeMMAP.

——— (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185(1), 259–282.

Canay, I. (2010): "EL inference for partially identified models: large deviations optimality and bootstrap validity," *Journal of Econometrics*, 156(2), 408–425.

Chen, X., E. Tamer, and A. Torgovitsky (2011): "Sensitivity Analysis in Semiparametric Likelihood Models," Working paper.

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

Ciliberto, F., and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77, 1791–1828.

Demuynck, T. (2015): "Bounding average treatment effects: A linear programming approach," *Economics Letters*, 137, 75 – 77.

Freyberger, J., and J. L. Horowitz (2015): "Identification and shape restrictions in nonparametric instrumental variables estimation," *Journal of Econometrics*, 189, 41–53.

Gafarov, B., and J. L. Montiel-Olea (2015): "On the maximum and minimum response to an impulse in SVARs," mimeo.

Hirano, K., and J. Porter (2012): "Impossibility results for nondifferentiable functionals," *Econometrica*, 80(4), 1769–1790.

Horn, R. A., and C. R. Johnson (1985): *Matrix Analysis*. Cambdridge University Press.

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

Jones, D. R. (2001): "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *Journal of Global Optimization*, 21(4), 345–383.

Jones, D. R., M. Schonlau, and W. J. Welch (1998): "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, 13(4), 455–492.

Kaido, H. (2012): "A Dual Approach to Inference for Partially Identified Econometric Models," Working Paper.

Kaido, H., and A. Santos (2014): "Asymptotically efficient estimation of models defined by convex moment inequalities," *Econometrica*, 82(1), 387–413.

Kitagawa, T. (2012): "Inference and Decision for Set Identified Parameters Using Posterior Lower Probabilities," CeMMAP Working Paper.

Kline, B., and E. Tamer (2015): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, forthcoming.

Leeb, H., and B. M. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

Mikusheva, A. (2007): "Uniform inference in autoregressive models," *Econometrica*, 75, 1411–1452.

Mohapatra, D., and C. Chatterjee (2015): "Price Control and Access to Drugs: The Case of India's Malaria Market," Working Paper. Cornell University.

Molchanov, I. (2005): *Theory of Random Sets*. Springer, London.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2011): "Moment Inequalities and Their Application," Discussion Paper, Harvard University.

——— (2015): "Moment Inequalities and Their Application," *Econometrica*, 83, 315334.

Pata, V. (2014): "Fixed Point Theorems and Applications," Mimeo.

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

Rockafellar, R. T., and R. J.-B. Wets (2005): *Variational Analysis, Second Edition*. Springer-Verlag, Berlin.

Romano, J. P., A. Shaikh, and M. Wolf (2014): "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 82, 1979–2002.

Romano, J. P., and A. M. Shaikh (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, 138, 2786–2807.

——— (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169–211.

Rosen, A. M. (2008): "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146(1), 107 – 117.

Stoye, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299–1315.

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *Review of Economic Studies*, 70, 147–165.

van der Vaart, A., and J. Wellner (2000): *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, Berlin.

Wan, Y. (2013): "An Integration-based Approach to Moment Inequality Models," Working Paper.