cemmap
centre for microdata methods and practice

# Confidence intervals for projections of partially identified parameters

Hiroaki Kaido
Francesca Molinari
Jörg Stoye

The Institute for Fiscal Studies
Department of Economics, UCL

**cemmap** working paper CWP49/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

# Confidence Intervals for Projections of Partially Identified Parameters*

Hiroaki Kaido† Francesca Molinari‡ Jörg Stoye§

October 25, 2017

**Abstract**

We propose a bootstrap-based *calibrated projection* procedure to build confidence intervals for single components and for smooth functions of a partially identified parameter vector in moment (in)equality models. The method controls asymptotic coverage uniformly over a large class of data generating processes.

The extreme points of the calibrated projection confidence interval are obtained by extremizing the value of the component (or function) of interest subject to a proper relaxation of studentized sample analogs of the moment (in)equality conditions. The degree of relaxation, or critical level, is calibrated so that the component (or function) of $\theta$, not $\theta$ itself, is uniformly asymptotically covered with prespecified probability. This calibration is based on repeatedly checking feasibility of linear programming problems, rendering it computationally attractive.

Nonetheless, the program defining an extreme point of the confidence interval is generally nonlinear and potentially intricate. We provide an algorithm, based on the response surface method for global optimization, that approximates the solution rapidly and accurately. The algorithm is of independent interest for inference on optimal values of stochastic nonlinear programs. We establish its convergence under conditions satisfied by canonical examples in the moment (in)equalities literature.

Our assumptions and those used in the leading alternative approach (a profiling based method) are not nested. An extensive Monte Carlo analysis confirms the accuracy of the solution algorithm and the good statistical as well as computational performance of calibrated projection, including in comparison to other methods.

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

*We are grateful to Elie Tamer and three anonymous reviewers for very useful suggestions that substantially improved the paper. We thank for their comments Ivan Canay and seminar and conference participants at Bonn, BC/BU joint workshop, Brown, Cambridge, Chicago, Columbia, Cornell, CREST, Kobe, Maryland, Michigan, Michigan State, NYU, Penn State, Royal Holloway, Syracuse, Toronto, UCLA, UCSD, UPenn, Vanderbilt, Vienna, Yale, Wisconsin, CEME, ES-NAWM 2015, Frontiers of Theoretical Econometrics Conference, ES-World Congress 2015, ES-Asia Meeting 2016, KEA-KAEA International Conference, Verein für Socialpolitik Ausschuss für Ökonometrie, and ES-ESM 2017. We are grateful to Zhonghao Fu, Debi Mohapatra, Sida Peng, Talal Rahim, Matthew Thirkettle, and Yi Zhang for excellent research assistance. A MATLAB package implementing the method proposed in this paper, Kaido, Molinari, Stoye, and Thirkettle (2017), is available at https://molinari.economics.cornell.edu/programs/KMSportable_V3.zip. We are especially grateful to Matthew Thirkettle for his contributions to this package. Finally, we gratefully acknowledge financial support through NSF grants SES-1230071 (Kaido), SES-0922330 (Molinari), and SES-1260980 (Stoye).

†Department of Economics, Boston University, hkaido@bu.edu.

‡Department of Economics, Cornell University, fm72@cornell.edu.

§Departments of Economics, Cornell University and University of Bonn, stoye@cornell.edu.

# 1 Introduction

This paper provides theoretically and computationally attractive confidence intervals for projections and smooth functions of a parameter vector $\theta \in \Theta \subset \mathbb{R}^d$, $d < \infty$, that is partially or point identified through a finite number of moment (in)equalities. The values of $\theta$ that satisfy these (in)equalities constitute the *identification region* $\Theta_I$.

Until recently, the rich literature on inference in this class of models focused on confidence sets for the entire vector $\theta$, usually obtained by test inversion as

$$\mathcal{C}_n(c_{1-\alpha}) \equiv \{\theta \in \Theta : T_n(\theta) \leqslant c_{1-\alpha}(\theta)\}, \tag{1.1}$$

where $T_n(\theta)$ is a test statistic that aggregates violations of the sample analog of the moment (in)equalities, and $c_{1-\alpha}(\theta)$ is a critical value that controls asymptotic coverage, often uniformly over a large class of data generating processes (DGPs). In point identified moment equality models, this would be akin to building confidence ellipsoids for $\theta$ by inversion of the $F$-test statistic proposed by Anderson and Rubin (1949).

However, applied researchers are frequently primarily interested in a specific component (or function) of $\theta$, e.g., the returns to education. Even if not, they may simply want to report separate confidence intervals for components of a vector, as is standard practice in other contexts. Thus, consider the projection $p'\theta$, where $p$ is a known unit vector. To date, it has been common to report as confidence interval for $p'\theta$ the projection of $\mathcal{C}_n(c_{1-\alpha})$:

$$CI_n^{proj} = \left[\inf_{\theta \in \mathcal{C}_n(c_{1-\alpha})} p'\theta, \sup_{\theta \in \mathcal{C}_n(c_{1-\alpha})} p'\theta\right], \tag{1.2}$$

where $n$ denotes sample size; see for example Ciliberto and Tamer (2009), Grieco (2014) and Dickstein and Morales (2016). Such projection is asymptotically valid, but typically yields conservative and therefore needlessly large confidence intervals. The potential severity of this effect is easily appreciated in a point identified example. Given a $\sqrt{n}$-consistent estimator $\hat{\theta}_n \in \mathbb{R}^d$ with limiting covariance matrix equal to the identity matrix, a 95% confidence interval for $\theta_k$ is obtained as $\hat{\theta}_{n,k} \pm 1.96$, $k = 1, \ldots, d$. In contrast, if $d = 10$, then projection of a 95% Wald confidence ellipsoid yields $\hat{\theta}_{n,k} \pm 4.28$ with true coverage of essentially 1. We refer to this problem as *projection conservatism*.

Our first contribution is to provide a bootstrap-based *calibrated projection* method that largely anticipates and corrects for projection conservatism. For each candidate $\theta$, $\hat{c}_n(\theta)$ is calibrated so that across bootstrap repetitions the projection of $\theta$ is covered with at least some pre-specified probability. Computationally, this bootstrap is relatively attractive because we linearize all constraints around $\theta$, so that coverage of $p'\theta$ corresponds to the projection of a

stochastic linear constraint set covering zero. We then propose the confidence interval

$$CI_n \equiv \left[ \inf_{\theta \in \mathcal{C}_n(\hat{c}_n)} p'\theta, \sup_{\theta \in \mathcal{C}_n(\hat{c}_n)} p'\theta \right]. \tag{1.3}$$

We prove that $CI_n$ asymptotically covers $p'\theta$ with probability at least $1-\alpha$ uniformly over a large class of DGPs and that it is weakly shorter than (1.2) for each $n$.[1] We also provide simple conditions under which it is asymptotically strictly shorter.

Our second contribution is a general method to accurately and rapidly compute projection-based confidence intervals. These can be our calibrated projection confidence intervals, but they can also correspond to projection of many other methods for inference on either $\theta$ or its identified set $\Theta_I$. Examples include Chernozhukov, Hong, and Tamer (2007), Andrews and Soares (2010), or (for conditional moment inequalities) Andrews and Shi (2013). Projection-based inference extends well beyond its application in partial identification, hence our computational method proves useful more broadly. For example, Freyberger and Reeves (2017a,b, Section S.3) use it to construct uniform confidence bands for an unknown function of interest under (nonparametric) shape restrictions.

We propose an algorithm that is based on the response surface method, thus it resembles an *expected improvement algorithm* (see e.g. Jones, 2001; Jones, Schonlau, and Welch, 1998, and references therein). Bull (2011) established convergence of the expected improvement algorithm for unconstrained optimization problems where the objective is a "black box" function. Building on his results, we show convergence of our algorithm for constrained optimization problems in which the constraint functions are "black box" functions, assuming that they are sufficiently smooth. We then verify this smoothness condition for canonical examples in the moment (in)equalities literature. Our extensive Monte Carlo experiments confirm that the algorithm is fast and accurate.[2]

Previous implementations of projection-based inference were based on approximating the set $\mathcal{C}_n(c_{1-\alpha}) \subset \mathbb{R}^d$ by searching for vectors $\theta \in \Theta$ such that $T_n(\theta) \leqslant c_{1-\alpha}(\theta)$ (using, e.g., grid-search or simulated annealing with no cooling), and reporting the smallest and largest value of $p'\theta$ among parameter values that were "guessed and verified" to belong to $\mathcal{C}_n(c_{1-\alpha})$. This becomes computationally cumbersome as $d$ increases because it typically requires a number of evaluation points that grows exponentially with $d$. In contrast, our method typically requires a number of evaluation points that grows linearly with $d$.

The main alternative inference prodedure for projections was introduced in Romano and Shaikh (2008) and significantly advanced in Bugni, Canay, and Shi (2017, BCS henceforth). It is based on profiling out a test statistic. The classes of DGPs for which our procedure and

[1] This comparison is based on projection of the confidence set of Andrews and Soares (2010) and holds the choice of tuning parameters and criterion function in (1.2) and (1.3) constant across methods.

[2] Freyberger and Reeves (2017b, Section S.3) similarly find our method to be accurate and to considerably reduce computational time.

the profiling-based method of BCS (BCS-profiling henceforth) can be shown to be uniformly valid are non-nested. We show that in well behaved cases, calibrated projection and BCS-profiling are asymptotically equivalent. We also provide conditions under which calibrated projection has lower probability of false coverage, thereby establishing that the two methods' power properties are non-ranked. Computationally, calibrated projection has the advantage that the bootstrap iterates over linear as opposed to nonlinear programming problems. While the "outer" optimization problems in (1.3) are potentially intricate, our algorithm is geared toward them. Our Monte Carlo simulations suggest that these two factors give calibrated projection a considerable computational edge over BCS-profiling, with an average speed gain of about 78-times.

In an influential paper, Pakes, Porter, Ho, and Ishii (2011) also use linearization but, subject to this approximation, directly bootstrap the sample projection.[3] This is valid only under stringent conditions, and we show that calibrated projection can be much simplified under those conditions. 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 (2016), 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. Chen, Christensen, and Tamer (2017) establish uniform validity of MCMC-based confidence intervals for projections, but these are aimed at covering the entire set $\{p'\theta : \theta \in \Theta_I(P)\}$, whereas we aim at covering the projection of $\theta$. Finally, Gafarov, Meier, and Montiel-Olea (2016) have used our insight in the context of set identified spatial VARs.

**Structure of the paper.** Section 2 sets up notation and describes our approach in detail. Section 3 describes computational implementation of the method and establishes convergence of our proposed algorithm. Section 4 lays out our assumptions and, under these assumptions, establishes uniform validity of calibrated projection for inference on projections and smooth functions of $\theta$. It also shows that more stringent conditions allow for several simplifications to the method, including that it can suffice to evaluate $\hat{c}_n$ at only two values of $\theta$ and that one can dispense with a tuning parameter. The section closes with a formal comparison of calibrated projection and BCS-profiling. Section 5 reports the results of Monte Carlo simulations. Section 6 draws conclusions. The proof of convergence of our algorithm is in Appendix A. All other proofs, background material for our algorithm, and additional results are in the Online Appendix.[4]

[3] The published version, i.e. Pakes, Porter, Ho, and Ishii (2015), does not contain the inference part.

[4] Section B provides convergence-related results and background material for our algorithm and describes how to compute $\hat{c}_n(\theta)$. Section C verifies, for a number of canonical moment (in)equality models, the assumptions that we invoke to show validity of our inference procedure and for our algorithm. Section D contains proofs of the Theorems in this paper's Section 4. Section E collects Lemmas supporting the preceding proofs. Section F provides further comparisons with the profiling method of Bugni, Canay, and Shi (2017), including an example where calibrated projection has higher power in finite sample. Section G provides comparisons with "uncalibrated" projection of the confidence region in Andrews and Soares (2010), including simple conditions

## 2 Detailed Explanation of the Method

Let $X_i \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ 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 $\theta$ is assumed to satisfy the moment inequality and equality restrictions

$$E_P[m_j(X_i, \theta)] \leqslant 0, \;\; j = 1, \cdots, J_1 \tag{2.1}$$

$$E_P[m_j(X_i, \theta)] = 0, \;\; j = J_1 + 1, \cdots, J_1 + J_2. \tag{2.2}$$

The identification region $\Theta_I(P)$ is the set of parameter values in $\Theta$ satisfying (2.1)-(2.2). For a random sample $\{X_i, i = 1, \cdots, n\}$ of observations drawn from $P$, we write

$$\bar{m}_{n,j}(\theta) \equiv n^{-1} \textstyle\sum_{i=1}^n m_j(X_i, \theta), \quad j = 1, \cdots, J_1 + J_2 \tag{2.3}$$

$$\hat{\sigma}_{n,j} \equiv (n^{-1} \textstyle\sum_{i=1}^n [m_j(X_i, \theta)]^2 - [\bar{m}_{n,j}(\theta)]^2)^{1/2}, \quad j = 1, \cdots, J_1 + J_2 \tag{2.4}$$

for the sample moments and the analog estimators of the population moment functions' standard deviations $\sigma_{P,j}$.[5]

The confidence interval in (1.3) then becomes $CI_n = [-s(-p, \mathcal{C}_n(\hat{c}_n)), s(p, \mathcal{C}_n(\hat{c}_n))]$, where

$$s(p, \mathcal{C}_n(\hat{c}_n)) \equiv \sup_{\theta \in \Theta} \; p'\theta \;\; s.t. \;\; \sqrt{n} \frac{\bar{m}_{n,j}(\theta)}{\hat{\sigma}_{n,j}(\theta)} \leqslant \hat{c}_n(\theta), \;\; j = 1, \cdots, J \tag{2.5}$$

and similarly for $(-p)$. Here, we define $J \equiv J_1 + 2J_2$ moments, where $\bar{m}_{n,J_1+J_2+k}(\theta) = -\bar{m}_{J_1+k}(\theta)$ for $k = 1, \cdots, J_2$. That is, we split moment equality constraints into two opposing inequality constraints and relax them separately.[6]

For a class of DGPs $\mathcal{P}$ that we specify below, define the asymptotic size of $CI_n$ by

$$\liminf_{n \to \infty} \inf_{P \in \mathcal{P}} \inf_{\theta \in \Theta_I(P)} P(p'\theta \in CI_n). \tag{2.6}$$

Our goal is to calibrate $\hat{c}_n$ so that (2.6) is at least equal to a prespecified level $1 - \alpha \geqslant 1/2$ while anticipating projection conservatism. To build intuition, fix $(\theta, P)$ s.t. $\theta \in \Theta_I(P), P \in$

under which $CI_n$ is asymptotically strictly shorter than $CI_n^{proj}$.

[5] Under Assumption 4.3-(II), in equation (2.5) instead of $\hat{\sigma}_{n,j}$ we use the estimator $\hat{\sigma}_{n,j}^M$ specified in (E.188) in Lemma E.10 p.50 of the Online Appendix for $j = 1, \ldots, 2R_1$ (with $R_1 \leqslant J_1/2$ defined in the assumption). In equation (3.2) we use $\hat{\sigma}_{n,j}$ for all $j = 1, \ldots, J$. To ease notation, we distinguish the two only where needed.

[6] 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 $\theta$ 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 $\sup_\theta \{p'\theta \;\; \text{s.t.} \;\; -\hat{c}_n \leqslant \sqrt{n}(\bar{X} - \theta)/\hat{\sigma}_n \leqslant \hat{c}_n\}$, with $p = 1$, and similarly for the lower endpoint.

$\mathcal{P}$. The projection of $\theta$ is covered when

$$
\begin{aligned}
&-s(-p, \mathcal{C}_n(\hat{c}_n)) \leqslant p'\theta \leqslant s(p, \mathcal{C}_n(\hat{c}_n)) \\
\Leftrightarrow &\left\{ \begin{array}{l} \inf_\vartheta p'\vartheta \\ s.t.\ \vartheta \in \Theta, \quad \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leqslant \hat{c}_n(\vartheta), \forall j \end{array} \right\} \leqslant p'\theta \leqslant \left\{ \begin{array}{l} \sup_\vartheta p'\vartheta \\ s.t.\ \vartheta \in \Theta, \quad \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leqslant \hat{c}_n(\vartheta), \forall j \end{array} \right\} \\
\Leftrightarrow &\left\{ \begin{array}{c} \inf_{\lambda \in \sqrt{n}(\Theta-\theta)} p'\lambda \\ s.t.\ \frac{\sqrt{n}\bar{m}_{n,j}\left(\theta+\frac{\lambda}{\sqrt{n}}\right)}{\hat{\sigma}_{n,j}\left(\theta+\frac{\lambda}{\sqrt{n}}\right)} \leqslant \hat{c}_n\left(\theta+\frac{\lambda}{\sqrt{n}}\right), \forall j \end{array} \right\} \leqslant 0 \leqslant \left\{ \begin{array}{c} \sup_{\lambda \in \sqrt{n}(\Theta-\theta)} p'\lambda \\ s.t.\ \frac{\sqrt{n}\bar{m}_{n,j}\left(\theta+\frac{\lambda}{\sqrt{n}}\right)}{\hat{\sigma}_{n,j}\left(\theta+\frac{\lambda}{\sqrt{n}}\right)} \leqslant \hat{c}_n\left(\theta+\frac{\lambda}{\sqrt{n}}\right), \forall j \end{array} \right\},
\end{aligned}
\tag{2.7}
$$

where the second equivalence follows from substituting $\vartheta = \theta + \lambda/\sqrt{n}$ and taking $\lambda$ to be the choice parameter. (Intuitively, we localize around $\theta$ at rate $1/\sqrt{n}$.)

We control asymptotic size by finding $\hat{c}_n$ such that 0 asymptotically lies within the optimal values of the NLPs in (2.7) with probability $1-\alpha$. To reduce computational burden, we approximate the event in equation (2.7) through linear expansion in $\lambda$ of the constraint set. To each constraint $j$, we add and subtract $\sqrt{n}E_P[m_j(X_i, \theta + \lambda/\sqrt{n})]/\hat{\sigma}_{n,j}(\theta + \lambda/\sqrt{n})$ and apply the mean value theorem to obtain

$$
\frac{\sqrt{n}\bar{m}_{n,j}\left(\theta+\frac{\lambda}{\sqrt{n}}\right)}{\hat{\sigma}_{n,j}\left(\theta+\frac{\lambda}{\sqrt{n}}\right)} = \left\{\mathbb{G}_{n,j}\left(\theta+\frac{\lambda}{\sqrt{n}}\right) + D_{P,j}(\bar{\theta})\lambda + \sqrt{n}\gamma_{1,P,j}(\theta)\right\} \frac{\sigma_{P,j}\left(\theta+\frac{\lambda}{\sqrt{n}}\right)}{\hat{\sigma}_{n,j}\left(\theta+\frac{\lambda}{\sqrt{n}}\right)}. \tag{2.8}
$$

Here $\mathbb{G}_{n,j}(\cdot) \equiv \sqrt{n}(\bar{m}_{n,j}(\cdot) - E_P[m_j(X_i, \cdot)])/\sigma_{P,j}(\cdot)$ 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, $\gamma_{1,P,j}(\cdot) \equiv E_P(m_j(X_i, \cdot))/\sigma_{P,j}(\cdot)$ is the studentized population moment, and the mean value $\bar{\theta}$ lies componentwise between $\theta$ and $\theta + \lambda/\sqrt{n}$.[7]

Calibration of $\hat{c}_n$ requires careful analysis of the local behavior of the moment restrictions at each point in the identification region. This is because the extent of projection conservatism depends on (i) the asymptotic behavior of the sample moments entering the inequality restrictions, which can change discontinuously depending on whether they bind at $\theta$ ($\gamma_{1,P,j}(\theta) = 0$) or not, and (ii) the local geometry of the identification region at $\theta$, i.e. the shape of the constraint set formed by the moment restrictions, and its relation to the level set of the objective function $p'\theta$. Features (i) and (ii) can be quite different at different points in $\Theta_I(P)$, making uniform inference for the projection challenging. In particular, (ii) does not arise if one only considers inference for the entire parameter vector, and hence is a new challenge requiring new methods.[8] This is where this paper's core theoretical innovation lies.

[7] The mean value $\bar{\theta}$ changes with $j$ but we omit the dependence to ease notation.

[8] This is perhaps best expressed in the testing framework: Inference for projections entails a null hypothesis specifying the value of a single component (or a function) of $\theta$. The components not under test become additional nuisance parameters, and dealing with them presents challenges that one does not face when testing hypotheses that specify the value of the entire vector $\theta$.

An important component of this innovation is to add to (2.7) the constraint that $\lambda \in \rho B^d$, where $B^d = [-1,1]^d$ and $\rho > 0$ a tuning parameter. This is slightly conservative but regularizes the effect of the local geometry of $\Theta_I(P)$ at $\theta$ on the inference problem. See Section 4.3 for further discussion. We show that the probability of the event in (2.7), with $\lambda$ restricted to be in $\rho B^d$, is asymptotically approximated by the probability that 0 lies between the optimal values of two programs that are linear in $\lambda$. The constraint sets of these programs are characterized by (i) a Gaussian process $\mathbb{G}_{P,j}(\theta)$ evaluated at $\theta$ (that we can approximate through a simple nonparametric bootstrap), (ii) a gradient $D_{P,j}(\theta)$ (that we can uniformly consistently estimate[9] on compact sets), and (iii) the parameter $\gamma_{1,P,j}(\theta)$ that measures the extent to which each moment inequality is binding (that we can conservatively estimate using insights from Andrews and Soares (2010)). This suggests a computationally attractive bootstrap procedure based on linear programs.

# 3 Computing Calibrated Projection Confidence Intervals

## 3.1 Computing the Critical Level

For a given $\theta \in \Theta$, we calibrate $\hat{c}_n(\theta)$ through a bootstrap procedure that iterates over linear programs.[10] Define

$$\Lambda_n^b(\theta, \rho, c) = \{\lambda \in \sqrt{n}(\Theta - \theta) \cap \rho B^d : \mathbb{G}_{n,j}^b(\theta) + \hat{D}_{n,j}(\theta)\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta)) \leqslant c, j = 1, \ldots, J\}, \tag{3.1}$$

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 bootstrap analog of $\mathbb{G}_{P,j}$,[11] $\hat{D}_{n,j}(\cdot)$ is a consistent estimator of $D_{P,j}(\cdot)$, $\rho > 0$ is a constant chosen by the researcher (see Section 4.3), $B^d = [-1,1]^d$, and $\hat{\xi}_{n,j}$ is defined by

$$\hat{\xi}_{n,j}(\theta) \equiv \begin{cases} \kappa_n^{-1}\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta) & j = 1, \ldots, J_1 \\ 0 & j = J_1 + 1, \ldots, J, \end{cases} \tag{3.2}$$

where $\kappa_n$ is a user-specified thresholding sequence such that $\kappa_n \to \infty$, $\varphi : \mathbb{R}^J_{[\pm\infty]} \to \mathbb{R}^J_{[\pm\infty]}$ is one of the generalized moment selection (GMS) functions proposed by Andrews and Soares (2010), and $\mathbb{R}_{[\pm\infty]} = \mathbb{R} \cup \{\pm\infty\}$. A common choice of $\varphi$ is given component-wise by

$$\varphi_j(x) = \begin{cases} 0 & \text{if } x \geqslant -1 \\ -\infty & \text{if } x < -1. \end{cases} \tag{3.3}$$

Restrictions on $\varphi$ and the rate at which $\kappa_n$ diverges are imposed in Assumption 4.2.

[9] See Online Appendix C for proposal of such estimators in some canonical moment (in)equality examples.

[10] If $\Theta$ is defined through smooth convex (in)equalities, these can be linearized too.

[11] Bugni, Canay, and Shi (2017) approximate the stochastic process $\mathbb{G}_{P,j}$ using $n^{-1/2}\sum_{i=1}^n [(m_j(X_i, \cdot) - \bar{m}_{n,j}(\cdot))/\hat{\sigma}_{n,j}(\cdot)]\chi_i$ 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.

REMARK 3.1: For concreteness, in (3.3) we write out the "hard thresholding" GMS function. As we establish below, our results apply to all but one of the GMS functions in Andrews and Soares (2010).[12]

Heuristically, the random convex polyhedral set $\Lambda_n^b(\theta, \rho, c)$ in (3.1) is a local (to $\theta$) linearized bootstrap approximation to the random constraint set in (2.7). To see this, note that the bootstrapped empirical process and the estimator of the gradient approximate the first two terms in the constraint in (2.7) as linearized in (2.8). Next, for $\theta \in \Theta_I(P)$, the GMS function conservatively approximates the local slackness parameter $\sqrt{n}\gamma_{1,P,j}(\theta)$. This is needed because the scaling of $\sqrt{n}\gamma_{1,P,j}(\theta)$ precludes consistent estimation. The problem is resolved 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 $\lambda$ to the "$\rho$-box" $\rho B^d$ 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 more detail in Section 4.3.

The critical level $\hat{c}_n(\theta)$ to be used in (1.3) is the smallest value of $c$ that makes the bootstrap probability of the event

$$\min_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \leqslant 0 \leqslant \max_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \tag{3.4}$$

at least $1-\alpha$. Because $\Lambda_n^b(\theta, \rho, c)$ is convex, we have

$$\left\{ \min_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \leqslant 0 \leqslant \max_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \right\} \Longleftrightarrow \left\{ \Lambda_n^b(\theta,\rho,c) \cap \{p'\lambda = 0\} \neq \varnothing \right\},$$

so that we can equivalently define

$$\hat{c}_n(\theta) \equiv \inf\{c \in \mathbb{R}_+ : P^*(\Lambda_n^b(\theta,\rho,c) \cap \{p'\lambda = 0\} \neq \varnothing) \geqslant 1-\alpha\}, \tag{3.5}$$

where $P^*$ denotes the law of the random set $\Lambda_n^b(\theta, \rho, c)$ induced by the bootstrap sampling process, i.e. by the distribution of $(X_1^b, \ldots, X_n^b)$, conditional on the data. Importantly, $P^*$ can be assessed by repeatedly checking feasibility of a linear program.[13] We describe in detail in Online Appendix B.4 how we compute $\hat{c}_n(\theta)$ through a root-finding algorithm.

[12] These are $\varphi^1 - \varphi^4$ in Andrews and Soares (2010), 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 $\varphi^5$ in Andrews and Soares (2010), which depends also on the covariance matrix of the moment functions.

[13] We implement a program in $\mathbb{R}^d$ for simplicity but, because $p'\lambda = 0$ defines a linear subspace, one could reduce this to $\mathbb{R}^{d-1}$.

### 3.2 Computation of the Outer Maximization Problem

Projection based methods as in (1.2) and (1.3) have nonlinear constraints involving a critical value which in general is an unknown function of $\theta$. Moreover, in all methods, including ours and Andrews and Soares (2010), the gradients of the critical values with respect to $\theta$ are not available in closed form. When the dimension of the parameter vector is large, directly solving optimization problems with such constraints can be expensive even if evaluating the critical value at each $\theta$ is cheap.

To mitigate this issue, we provide an algorithm that is a contribution to the moment (in)equalities literature in its own right and that can be helpful for implementing other approaches.[14] We apply it to constrained optimization problems of the following form:

$$\begin{aligned} p'\theta^* \equiv \sup_{\theta\in\Theta} \; & p'\theta \\ s.t. \; & g_j(\theta) \leqslant c(\theta), \; j = 1, \cdots, J, \end{aligned} \tag{3.6}$$

where $\theta^*$ is an optimal solution of the problem, $g_j, j = 1, \ldots, J$ are known functions, and $c$ is a function that requires a higher computational cost. In our context, $g_j(\theta) = \sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)$ and, for calibrated projection, $c(\theta) = \hat{c}_n(\theta)$. Conditional on the data $\{X_1, \cdots, X_n\}$, these functions are considered deterministic. A key feature of the problem is that the function $c(\cdot)$ is relatively costly to evaluate.[15] Our algorithm evaluates $c(\cdot)$ on finitely many values of $\theta$. For other values, it imposes a probabilistic model that gets updated as specific values are computed and that is used to determine the next evaluation point. Under reasonable conditions, the resulting sequence of approximate optimal values converges to $p'\theta^*$.

Specifically, after drawing an initial set of evaluation points that grows linearly with the dimensionality of parameter space, the algorithm has three steps called E, A, and M below.

**Initialization-step:** Draw randomly (uniformly) over $\Theta$ a set $(\theta^{(1)}, \cdots, \theta^{(k)})$ of initial evaluation points. We suggest setting $k = 10d + 1$.

**E-step: (Evaluation)** Evaluate $c(\theta^{(\ell)})$ for $\ell = 1, \cdots, L$, where $L \geqslant k$. Set $\Upsilon^{(\ell)} = c(\theta^{(\ell)}), \ell = 1, \cdots, L$. The current estimate $p'\theta^{*,L}$ of the optimal value can be computed using

$$\theta^{*,L} \in \text{argmax}_{\theta\in\mathcal{C}^L} p'\theta, \tag{3.7}$$

where $\mathcal{C}^L \equiv \{\theta^{(\ell)} : \ell \in \{1, \cdots, L\}, g_j(\theta^{(\ell)}) \leqslant c(\theta^{(\ell)}), j = 1, \cdots, J\}$ is the set of feasible evaluation points.

[14] This algorithm is based on the response surface method used in the optimization literature; see Jones (2001), Jones, Schonlau, and Welch (1998), and references therein.

[15] Here we assume that computing the sample moments is less expensive than computing the critical value. When computation of moments is also very expensive, our proposed algorithm can be used to approximate these too.

**A-step: (Approximation)** Approximate $\theta \mapsto c(\theta)$ by a flexible auxiliary model. We use a Gaussian-process regression model (or kriging), which for a mean-zero Gaussian process $\epsilon(\cdot)$ indexed by $\theta$ and with constant variance $\varsigma^2$ specifies

$$\Upsilon^{(\ell)} = \mu + \epsilon(\theta^{(\ell)}),\ \ell = 1, \cdots, L \tag{3.8}$$

$$Corr(\epsilon(\theta), \epsilon(\theta')) = K_\beta(\theta - \theta'),\ \theta, \theta' \in \Theta, \tag{3.9}$$

where $K_\beta$ is a kernel with parameter vector $\beta \in \times_{k=1}^d [\underline{\beta}_k, \overline{\beta}_k] \subset \mathbb{R}^d_{++}$, e.g. $K_\beta(\theta - \theta') = \exp(-\sum_{k=1}^d |\theta_k - \theta'_k|^2/\beta_k)$. The unknown parameters $(\mu, \varsigma^2)$ 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 $\beta$ 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

$$c_L(\theta) = \hat{\mu} + \mathbf{r}_L(\theta)' \mathbf{R}_L^{-1} (\boldsymbol{\Upsilon} - \hat{\mu}\mathbf{1}), \tag{3.10}$$

$$\nabla_\theta c_L(\theta) = \hat{\mu} + \mathbf{Q}_L(\theta) \mathbf{R}_L^{-1} (\boldsymbol{\Upsilon} - \hat{\mu}\mathbf{1}), \tag{3.11}$$

where $\mathbf{r}_L(\theta)$ is a vector whose $\ell$-th component is $Corr(\epsilon(\theta), \epsilon(\theta^{(\ell)}))$ as given above with estimated parameters, $\mathbf{Q}_L(\theta) = \nabla_\theta \mathbf{r}_L(\theta)'$, and $\mathbf{R}_L$ is an $L$-by-$L$ matrix whose $(\ell, \ell')$ entry is $Corr(\epsilon(\theta^{(\ell)}), \epsilon(\theta^{(\ell')}))$ with estimated parameters. This approximating (or surrogate) model has the property that its predictor satisfies $c_L(\theta^{(\ell)}) = c(\theta^{(\ell)}), \ell = 1, \cdots, L$. Hence, it provides an analytical interpolation to the evaluated critical values together with an analytical gradient.[16] Further, the amount of uncertainty left in $c(\theta)$ (at an arbitrary point) is captured by the following variance:

$$\hat{\varsigma}^2 s_L^2(\theta) = \hat{\varsigma}^2 \Big( 1 - \mathbf{r}_L(\theta)' \mathbf{R}_L^{-1} \mathbf{r}_L(\theta) + \frac{(1 - \mathbf{1}' \mathbf{R}_L^{-1} \mathbf{r}_L(\theta))^2}{\mathbf{1}' \mathbf{R}_L^{-1} \mathbf{1}} \Big). \tag{3.12}$$

**M-step: (Maximization):** With probability $1 - \epsilon$, maximize the expected improvement function $\mathbb{EI}_L$ to obtain the next evaluation point, with:

$$\theta^{(L+1)} \equiv \arg\max_{\theta \in \Theta} \mathbb{EI}_L(\theta) = \arg\max_{\theta \in \Theta} (p'\theta - p'\theta^{*,L})_+ \Big( 1 - \Phi\Big( \frac{\bar{g}(\theta) - c_L(\theta)}{\hat{\varsigma} s_L(\theta)} \Big) \Big), \tag{3.13}$$

where $\bar{g}(\theta) = \max_{j=1,\cdots,J} g_j(\theta)$. This step can be implemented by standard nonlinear optimization solvers, e.g. Matlab's `fmincon` or `KNITRO` (see Appendix B.3 for details). With probability $\epsilon$, draw $\theta^{(L+1)}$ randomly from a uniform distribution over $\Theta$.

Once the next evaluation point $\theta^{(L+1)}$ is determined, one adds it to the set of evaluation

[16] 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 our Monte Carlo experiments.

points and iterates the E-A-M steps. This yields an increasing sequence of approximate optimal values $p'\theta^{*,L}, L = k+1, k+2, \cdots$. Once a convergence criterion is met, the value $p'\theta^{*,L}$ is reported as the end point of $CI_n$. We discuss convergence criteria in Section 5.

REMARK 3.2: The advantages of E-A-M are as follows. First, we control the number of points at which we evaluate the critical value. Since the evaluation of the critical value is the relatively expensive step, controlling the number of evaluations is important. One should also note that the E-step with the initial $k$ evaluation points can easily be parallelized. For any additional E-step (i.e. $L > k$), one needs to evaluate $c(\cdot)$ only at a single point $\theta^{(L+1)}$. The M-step is crucial for reducing the number of additional evaluation points. To determine the next evaluation point, one needs to take into account the trade-off between "exploitation" (i.e. the benefit of drawing a point at which the optimal value is high) and "exploration" (i.e. the benefit of drawing a point in a region in which the approximation error of $c$ is currently large). The expected improvement function in (3.13) quantifies this trade-off, and draws a point only in an area where one can expect the largest improvement in the optimal value, yielding substantial computational savings.[17]

Second, the proposed algorithm simplifies the M-step by providing constraints and their gradients for program (3.13) in closed form. Availability of analytical gradients greatly aids fast and stable numerical optimization. The price is the additional approximation step. In the numerical exercises of Section 5, this price turns out to be low.

### 3.3 Convergence of the E-A-M Algorithm

We now provide formal conditions under which $p'\theta^{*,L}$ converges to the true end point of $CI_n$ as $L \to \infty$.[18] Our convergence result recognizes that the parameters of the Gaussian process prior in (3.8) are estimated for each iteration of the A-step using the "observations" $\{\theta^\ell, c(\theta^\ell)\}_{\ell=1}^L$, and hence change with $L$. Because of this, a requirement for convergence is that $c(\theta)$ is a sufficiently smooth function of $\theta$. We show that a high-level condition guaranteeing this level of smoothness ensures a general convergence result for the E-A-M algorithm. This is a novel contribution to the literature on response surface methods for constrained optimization.

In the statement of Theorem 3.1 below, $\mathcal{H}_\beta(\Theta)$ is the reproducing kernel Hilbert space (RKHS) on $\Theta \subseteq \mathbb{R}^d$ determined by the kernel used to define the correlation functional in (3.9). The norm on this space is $\|\cdot\|_{\mathcal{H}_\beta}$; see Online Appendix B.2 for details. Also, the expectation $E_{\mathbb{Q}}$ is taken with respect to the law of $(\theta^{(1)}, \cdots, \theta^{(L)})$ determined by the Initialization-step and the M-step, holding the sample fixed. See Appendix A for a precise definition of $E_{\mathbb{Q}}$ and a proof of the theorem.

[17] It is also possible to draw multiple points in each iteration. See Schonlau, Welch, and Jones (1998).

[18] We build on Bull (2011), who proves a convergence result for the algorithm proposed by Jones, Schonlau, and Welch (1998) applied to an unconstrained optimization problem in which the objective function is unknown outside the evaluation points.

THEOREM 3.1: *Suppose $\Theta \subset \mathbb{R}^d$ is a compact hyperrectangle with nonempty interior and that $\|p\| = 1$. Let the evaluation points $(\theta^{(1)}, \cdots, \theta^{(L)})$ be drawn according to the Initialization and the M steps. Let $K_\beta$ in (3.9) be a Matérn kernel with index $\nu \in (0, \infty)$ and $\nu \notin \mathbb{N}$. Let $c : \Theta \mapsto \mathbb{R}$ satisfy $\|c\|_{\mathcal{H}_{\bar{\beta}}} \leqslant R$ for some $R > 0$, where $\bar{\beta} = (\bar{\beta}_1, \cdots, \bar{\beta}_d)'$. Then*

$$E_{\mathbb{Q}}\left[p'\theta^* - p'\theta^{*,L+1}\right] \to 0 \quad as \quad L \to \infty. \tag{3.14}$$

REMARK 3.3: The requirement that $\Theta$ is a compact hyperrectangle with nonempty interior can be replaced by a requirement that $\Theta$ belongs to the interior of a closed hyperrectangle in $\mathbb{R}^d$ such that $c$ satisfies the smoothness requirement in Theorem 3.1 on that rectangle.

In order to apply Theorem 3.1 to calibrated projection, we provide low level conditions (Assumption B.1 in Online Appendix B.1.1) under which the map $\theta \mapsto \hat{c}_n(\theta)$ uniformly stochastically satisfies a Lipschitz-type condition. To get smoothness, we work with a mollified version of $\hat{c}_n$, denoted $\hat{c}_{n,\tau_n}$ and provided in equation (B.1), with $\tau_n = o(n^{-1/2})$.[19] Theorem B.1 in the Online Appendix shows that $\hat{c}_n$ and $\hat{c}_{n,\tau_n}$ can be made uniformly arbitrarily close to each other and that $\hat{c}_{n,\tau_n}$ yields valid inference in the sense of equation (2.6). In practice, one may therefore directly apply the E-A-M steps to $\hat{c}_n$.

REMARK 3.4: The key condition imposed in Theorem B.1 is Assumption B.1. It requires that the GMS function used is Lipschitz in its argument, and that the standardized moment functions are Lipschitz in $\theta$. In Online Appendix C.1 we establish that the latter condition is satisfied by some canonical examples in the moment (in)equality literature, namely the mean with missing data, linear regression and best linear prediction with interval data (and discrete covariates), and entry games with multiple equilibria (and discrete covariates).[20]

# 4 Asymptotic Validity of Inference

## 4.1 Assumptions

We posit that $P$, the distribution of the observed data, belongs to a class of distributions denoted by $\mathcal{P}$. We write stochastic order relations that hold uniformly over $P \in \mathcal{P}$ using the notations $o_{\mathcal{P}}$ and $O_{\mathcal{P}}$; see Online Appendix D.1 for the formal definitions. Below, $\epsilon$, $\varepsilon$, $\delta$, $\omega$, $\underline{\sigma}$, $M$, $\bar{M}$ denote generic constants which may be different in different appearances but cannot depend on $P$. Given a square matrix $A$, we write $\text{eig}(A)$ for its smallest eigenvalue.

ASSUMPTION 4.1: *(a) $\Theta \subset \mathbb{R}^d$ is a compact hyperrectangle with nonempty interior. (b) All distributions $P \in \mathcal{P}$ satisfy the following:*

[19] For a discussion of mollification, see e.g. Rockafellar and Wets (2005, Example 7.19)

[20] It can also be shown to hold in semi-parametric binary regression models with discrete or interval valued covariates under the assumptions of Magnac and Maurin (2008).

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

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

*(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* $j$, $E_P[\sup_{\theta\in\Theta} |m_j(X_i,\theta)/\sigma_{P,j}(\theta)|^{2+\delta}] \leqslant M$.

ASSUMPTION 4.2: *The function* $\varphi_j$ *is continuous at all* $x \geqslant 0$ *and* $\varphi_j(0) = 0$*;* $\kappa_n \to \infty$ *and* $\kappa_n = o(n^{1/2})$*. If Assumption 4.3-(II) is imposed,* $\kappa_n = o(n^{1/4})$.

Assumption 4.1-(a) requires that $\Theta$ is a hyperrectangle, but can be replaced with the assumption that $\theta$ is defined through a finite number of nonstochastic inequality constraints smooth in $\theta$ and such that $\Theta$ is convex. Compactness is a standard assumption on $\Theta$ for extremum estimation. We additionally require convexity as we use mean value expansions of $E_P[m_j(X_i,\theta)]/\sigma_{P,j}(\theta)$ in $\theta$; see (2.8). Assumption 4.1-(b) defines our moment (in)equalities model. Assumption 4.2 constrains the GMS function and the rate at which its tuning parameter diverges. Both 4.1-(b) and 4.2 are based on Andrews and Soares (2010) and are standard in the literature,[21] although typically with $\kappa_n = o(n^{1/2})$. The slower rate $\kappa_n = o(n^{1/4})$ is satisfied for the popular choice, recommended by Andrews and Soares (2010), of $\kappa_n = \sqrt{\ln n}$.

Next, and unlike some other papers in the literature, we impose restrictions on the correlation matrix of the moment functions. These conditions can be easily verified in practice because they are implied when the correlation matrix of the moment equality functions and the moment inequality functions specified below have a determinant larger than a predefined constant for any $\theta \in \Theta$.

ASSUMPTION 4.3: *All distributions* $P \in \mathcal{P}$ *satisfy* ***one*** *of the following two conditions for some constants* $\omega > 0, \underline{\sigma} > 0, \epsilon > 0, \varepsilon > 0, M < \infty$*:*

*(I) Let* $\mathcal{J}(P,\theta;\varepsilon) \equiv \{j \in \{1,\cdots,J_1\} : E_P[m_j(X_i,\theta)]/\sigma_{P,j}(\theta) \geqslant -\varepsilon\}$*. Denote*

$$\tilde{m}(X_i,\theta) \equiv \left(\{m_j(X_i,\theta)\}_{j\in\mathcal{J}(P,\theta;\varepsilon)}, m_{J_1+1}(X_i,\theta),\cdots,m_{J_1+J_2}(X_i,\theta)\right)',$$
$$\tilde{\Omega}_P(\theta) \equiv Corr_P(\tilde{m}(X_i,\theta)).$$

*Then* $\inf_{\theta\in\Theta_I(P)} \text{eig}(\tilde{\Omega}_P(\theta)) \geqslant \omega$.

*(II) The functions* $m_j(X_i,\theta)$ *are defined on* $\Theta^\epsilon = \{\theta \in \mathbb{R}^d : d(\theta,\Theta) \leqslant \epsilon\}$*. There exists* $R_1 \in \mathbb{N}$, $1 \leqslant R_1 \leqslant J_1/2$*, and measurable functions* $t_j : \mathcal{X} \times \Theta^\epsilon \to [0,M]$, $j \in \mathcal{R}_1 \equiv \{1,\ldots,R_1\}$*, such that for each* $j \in \mathcal{R}_1$*,*

$$m_{j+R_1}(X_i,\theta) = -m_j(X_i,\theta) - t_j(X_i,\theta). \tag{4.1}$$

[21] Continuity of $\varphi_j$ for $x \geqslant 0$ is restrictive only for GMS function $\varphi^{(2)}$ in Andrews and Soares (2010).

*For each* $j \in \mathcal{R}_1 \cap \mathcal{J}(P,\theta;\varepsilon)$ *and any choice* $\ddot{m}_j(X_i,\theta) \in \{m_j(X_i,\theta), m_{j+R_1}(X_i,\theta)\}$, *denoting* $\tilde{\Omega}_P(\theta) \equiv Corr_P(\tilde{m}(X_i,\theta))$, *where*

$$\tilde{m}(X_i,\theta) \equiv \Big(\{\ddot{m}_j(X_i,\theta)\}_{j\in\mathcal{R}_1\cap\mathcal{J}(P,\theta;\varepsilon)}, \\ \{m_j(X_i,\theta)\}_{j\in\mathcal{J}(P,\theta;\varepsilon)\setminus\{1,\ldots,2R_1\}}, m_{J_1+1}(X_i,\theta),\cdots,m_{J_1+J_2}(X_i,\theta)\Big)',$$

*one has*

$$\inf_{\theta\in\Theta_I(P)} \text{eig}(\tilde{\Omega}_P(\theta)) \geqslant \omega. \tag{4.2}$$

*Finally,*

$$\inf_{\theta\in\Theta_I(P)} \sigma_{P,j}(\theta) > \underline{\sigma} \text{ for } j = 1,\ldots,R_1. \tag{4.3}$$

Assumption 4.3-(I) requires that the correlation matrix of the moment functions corresponding to close-to-binding moment conditions has eigenvalues uniformly bounded from below. This assumption holds in many applications of interest, including: (i) instances when the data is collected by intervals with minimum width;[22] (ii) in treatment effect models with (uniform) overlap; (iii) in static complete information entry games under weak solution concepts, e.g. rationality of level 1, see Aradillas-Lopez and Tamer (2008).

We are aware of two examples in which Assumption 4.3-(I) may fail. One are missing data scenarios, e.g. scalar mean, linear regression, and best linear prediction, with a vanishing probability of missing data. The other example, which is extensively simulated in Section 5, is the Ciliberto and Tamer (2009) entry game model when the solution concept is pure strategy Nash equilibrium. We show in Online Appendix C.2 that these examples satisfy Assumption 4.3-(II).

REMARK 4.1: Assumption 4.3-(II) weakens 4.3-(I) by allowing for (drifting to) perfect correlation among moment inequalities that cannot cross. This assumption is often satisfied in moment conditions that are separable in data and parameters, i.e. for each $j = 1,\ldots,J$,

$$E_P[m_j(X_i,\theta)] = E_P[h_j(X_i)] - v_j(\theta), \tag{4.4}$$

for some measurable functions $h_j : \mathcal{X} \to \mathbb{R}$ and $v_j : \Theta \to \mathbb{R}$. Models like the one in Ciliberto and Tamer (2009) fall in this category, and we verify Assumption 4.3-(II) for them in Online Appendix C.2. The argument can be generalized to other separable models.

[22] Empirically relevant examples are that of: (a) the Occupational Employment Statistics (OES) program at the Bureau of Labor Statistics, which collects wage data from employers as intervals of positive width, and uses these data to construct estimates for wage and salary workers in 22 major occupational groups and 801 detailed occupations; and (b) when, due to concerns for privacy, data is reported as the number of individuals who belong to each of a finite number of cells (for example, in public use tax data).

In Online Appendix C.2, we also verify Assumption 4.3-(II) for some models that are not separable in the sense of equation (4.4), for example best linear prediction with interval outcome data. The proof can be extended to cover (again non-separable) binary models with discrete or interval valued covariates under the assumptions of Magnac and Maurin (2008).

In what follows, we refer to pairs of inequality constraints indexed by $\{j, j+R_1\}$ and satisfying (4.1) as "paired inequalities." Their presence requires a modification of the bootstrap procedure. This modification exclusively concerns the definition of $\Lambda_n^b(\theta, \rho, c)$ in equation (3.1). We explain it here for the case that the GMS function $\varphi_j$ is the hard-thresholding one in (3.3), and refer to Online Appendix E equations (E.12)-(E.13) for the general case. If

$$\varphi_j(\hat{\xi}_{n,j}(\theta)) = 0 = \varphi_j(\hat{\xi}_{n,j+R_1}(\theta)),$$

we replace $\mathbb{G}^b_{n,j+R_1}(\theta)$ with $-\mathbb{G}^b_{n,j}(\theta)$ and $\hat{D}_{n,j+R_1}(\theta)$ with $-\hat{D}_{n,j}(\theta)$, so that inequality $\mathbb{G}^b_{n,j+R_1}(\theta) + \hat{D}_{n,j+R_1}(\theta)\lambda \leqslant c$ is replaced with $-\mathbb{G}^b_{n,j}(\theta) - \hat{D}_{n,j}(\theta)\lambda \leqslant c$ in equation (3.1). 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. In the proof of Theorem 4.1, we show that this tightens the stochastic program.[23] The rest of the procedure is unchanged.

Instead of Assumption 4.3, BCS (Assumption 2) impose the following high-level condition: (a) The limit distribution of their profiled test statistic is continuous at its $1-\alpha$ quantile if this quantile is positive; (b) else, their test is asymptotically valid with a critical value of zero. In Online Appendix D.2.3, we show that we can replace Assumption 4.3 with a weaker high level condition (Assumption D.1-(I)) that resembles the BCS assumption but constrains the limiting coverage probability. (We do not claim that the conditions are equivalent.) The substantial amount of work required for us to show that Assumption 4.3 implies Assumption D.1-(I) is suggestive of how difficult these high-level conditions can be to verify.[24] Moreover, in Online Appendix F.2 we provide a simple example that violates Assumption 4.3 and in which all of calibrated projection, BCS-profiling, and the boosttrap procedure in Pakes, Porter, Ho, and Ishii (2011) fail. The example leverages the fact that when binding constraints are near-perfectly correlated, the projection may be estimated superconsistently, invalidating the simple nonparametric bootstrap.[25]

Together with imposition of the $\rho$-box constraints, Assumption 4.3 allows us to dispense with restrictions on the local geometry of the set $\Theta_I(P)$. Restrictions of this type, which are akin to constraint qualification conditions, are imposed by BCS (Assumption A.3-(a)),

[23] When paired inequalities are present, in equation (2.5) instead of $\hat{\sigma}_{n,j}$ we use the estimator $\hat{\sigma}^M_{n,j}$ specified in (E.188) in Lemma E.10 p.50 of the Online Appendix for $\sigma_{P,j}, j = 1, \ldots, 2R_1$ (with $R_1 \leqslant J_1/2$ defined in the assumption). In equation (3.2) we use $\hat{\sigma}_{n,j}$ for all $j = 1, \ldots, J$. To ease notation, we do not distinguish the two unless it is needed.

[24] Assumption 4.3 is used exclusively to obtain the conclusions of Lemma E.6, E.7 and E.8, hence any alternative assumption that delivers such results can be used.

[25] The example we provide satisfies all assumptions explicitly stated in Pakes, Porter, Ho, and Ishii (2011), illustrating an oversight in their Theorem 2.

Pakes, Porter, Ho, and Ishii (2011, Assumptions A.3-A.4), Chernozhukov, Hong, and Tamer (2007, Condition C.2), and elsewhere. In practice, they can be hard to verify or pre-test for. We study this matter in detail in Kaido, Molinari, and Stoye (2017).

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

ASSUMPTION 4.4: *All distributions $P \in \mathcal{P}$ satisfy the following conditions:*

*(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 $\sup_{\theta\in\Theta^\epsilon} \|\hat{D}_{n,j}(\theta) - D_{P,j}(\theta)\| = o_{\mathcal{P}}(1)$.*

*(ii) There exist $M, \bar{M} < \infty$ such that for all $\theta, \tilde{\theta} \in \Theta^\epsilon$ $\max_{j=1,\cdots,J} \|D_{P,j}(\theta) - D_{P,j}(\tilde{\theta})\| \leqslant M\|\theta - \tilde{\theta}\|$ and $\max_{j=1,\cdots,J} \sup_{\theta\in\Theta_I(P)} \|D_{P,j}(\theta)\| \leqslant \bar{M}$.*

Assumption 4.4 requires that each of the $J$ normalized population moments is differentiable, that its derivative is Lipschitz continuous, and that this derivative can be consistently estimated uniformly in $\theta$ and $P$.[26] We require these conditions because we use a linear expansion of the population moments to obtain a first-order approximation to the nonlinear programs defining $CI_n$, and because our bootstrap procedure requires an estimator of $D_P$.

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

$$\varrho_P(\theta,\tilde{\theta}) \equiv \left\| \left\{ \left[ Var_P\left(\sigma_{P,j}^{-1}(\theta)m_j(X,\theta) - \sigma_{P,j}^{-1}(\tilde{\theta})m_j(X,\tilde{\theta})\right)\right]^{1/2} \right\}_{j=1}^{J} \right\|. \tag{4.5}$$

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 $m_k(X_i,\tilde{\theta}))/\sigma_{P,k}(\tilde{\theta})$.

ASSUMPTION 4.5: *All distributions $P \in \mathcal{P}$ satisfy the following conditions:*

*(i) The class of functions $\{\sigma_{P,j}^{-1}(\theta)m_j(\cdot,\theta) : \mathcal{X} \to \mathbb{R}, \theta \in \Theta\}$ is measurable for each $j = 1,\cdots,J$.*

*(ii) The empirical process $\mathbb{G}_n$ with $j$-th component $\mathbb{G}_{n,j}$ is uniformly asymptotically $\varrho_P$-equicontinuous. That is, for any $\epsilon > 0$,*

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} P\left( \sup_{\varrho_P(\theta,\tilde{\theta})<\delta} \|\mathbb{G}_n(\theta) - \mathbb{G}_n(\tilde{\theta})\| > \epsilon \right) = 0. \tag{4.6}$$

*(iii) $Q_P$ satisfies*

$$\lim_{\delta\downarrow 0} \sup_{\|(\theta_1,\tilde{\theta}_1)-(\theta_2,\tilde{\theta}_2)\|<\delta} \sup_{P\in\mathcal{P}} \|Q_P(\theta_1,\tilde{\theta}_1) - Q_P(\theta_2,\tilde{\theta}_2)\| = 0. \tag{4.7}$$

[26] The requirements are imposed on $\Theta^\epsilon$. Under Assumption 4.3-(I) it suffices they hold on $\Theta$.

Under this assumption, the class of normalized moment functions is uniformly Donsker (Bugni, Canay, and Shi, 2015). We use this fact to show validity of our method.

## 4.2 Theoretical Results

**First set of results: Uniform asymptotic validity in the general case**.

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 $s(p, \mathcal{C}_n(\hat{c}_n))$ was defined in equation (2.5) and $\hat{c}_n$ in (3.5).

THEOREM 4.1: *Suppose Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. Let* $0 < \alpha < 1/2$. *Then*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(p'\theta \in CI_n) \geqslant 1-\alpha. \tag{4.8}$$

A simple corollary to Theorem 4.1, whose proof is omitted, is that we can provide joint confidence regions for several projections, in particular confidence hyperrectangles for subvectors. Thus, let $p^1, \ldots, p^k$ denote unit vectors in $\mathbb{R}^d$, $k \leqslant d$. Then:

COROLLARY 4.1: *Suppose Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. Let* $0 < \alpha < 1/2$. *Then,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(p^{\ell\prime}\theta \in CI_{n,\ell}, \ell = 1, \ldots, k) \geqslant 1-\alpha, \tag{4.9}$$

*where* $CI_{n,\ell} = \left[\inf_{\theta\in\mathcal{C}_n(\hat{c}_n^k)} p^{\ell\prime}\theta, \sup_{\theta\in\mathcal{C}_n(\hat{c}_n^k)} p^{\ell\prime}\theta\right]$ *and* $\hat{c}_n^k(\theta) \equiv \inf\{c \in \mathbb{R}_+ : P^*(\Lambda_n^b(\theta, \rho, c) \cap \{\cap_{\ell=1}^k \{p^{\ell\prime}\lambda = 0\}\} \neq \varnothing) \geqslant 1-\alpha\}$.

The difference in this Corollary compared to Theorem 4.1 is that $\hat{c}_n^k$ is calibrated so that (3.4) holds for all $p^1, \ldots, p^k$ simultaneously.

In applications, a researcher might wish to obtain a confidence interval for a known nonlinear function $f : \Theta \mapsto \mathbb{R}$. Examples include policy analysis and counterfactual estimation in the presence of partial identification, or demand extrapolation subject to rationality constraints. It is possible to extend our results to uniformly continuously differentiable functions $f$. Because the function $f$ is known, the conditions on its gradient required below can be easily verified in practice (especially if the first one is strengthened to hold over $\Theta$).

THEOREM 4.2: *Let* $CI_n^f$ *be a confidence interval whose lower and upper points are obtained solving*

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

*where* $\hat{c}_n^f(\theta) \equiv \inf\{c \geqslant 0 : P^*(\Lambda_n^b(\theta,\rho,c) \cap \{\|\nabla_\theta f(\theta)\|^{-1}\nabla_\theta f(\theta)\lambda = 0\} \neq \varnothing) \geqslant 1-\alpha\}$. *Suppose Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. Suppose that there exist* $\varpi > 0$ *and* $M < \infty$ *such that* $\inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} \|\nabla f(\theta)\| \geqslant \varpi$ *and* $\sup_{\theta,\bar{\theta}\in\Theta} \|\nabla f(\theta) - \nabla f(\bar{\theta})\| \leqslant M\|\theta - \bar{\theta}\|$, *where* $\nabla_\theta f(\theta)$ *is the gradient of* $f(\theta)$. *Let* $0 < \alpha < 1/2$. *Then,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(f(\theta) \in CI_n^f) \geqslant 1-\alpha. \tag{4.10}$$

**Second set of results: Simplifications for special cases**.

We now consider more restrictive assumptions on the model, defining a subset of DGPs $\mathcal{Q} \subset \mathcal{P}$; across theorems below, the set $\mathcal{Q}$ differs based on which assumptions are maintained. If $P \in \mathcal{Q}$, a number of simplifications to the method, including dropping the $\rho$-box constraints, are possible. Here we state the formal results and then we give a heuristic explanation of the conditions needed for these simplifications. Online Appendix D.3.1 contains the exact assumptions and Online Appendix D.3.2 the proofs. We remark that all of the additional assumptions are implied by assumptions in Pakes, Porter, Ho, and Ishii (2011), hence under their conditions Theorem 4.3 applies in its entirety.

THEOREM 4.3: *Suppose Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. Let* $0 < \alpha < 1/2$.

*(I) If Assumption D.2-(1) holds for either* $p$ *or* $-p$ *(or both), then setting*

$$CI_n = \Big[\inf_{\theta\in\mathcal{C}_n(\hat{c}_{n,-p})} p'\theta, \sup_{\theta\in\mathcal{C}_n(\hat{c}_{n,p})} p'\theta\Big], \tag{4.11}$$

$$\hat{c}_{n,q}(\theta) = \inf\{c \in \mathbb{R}_+ : P^*(\Lambda_n^b(\theta,\rho,c) \cap \{q'\lambda \geqslant 0\} \neq \varnothing) \geqslant 1-\alpha\},\ q \in \{p,-p\}, \tag{4.12}$$

*we have*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{Q}} \inf_{\theta\in\Theta_I(P)} P(p'\theta \in CI_n) \geqslant 1-\alpha. \tag{4.13}$$

*(II) If Assumptions D.2-(1) (for either* $p$ *or* $-p$ *or both), D.3 and D.4 hold, then (4.13) continues to be satisfied with* $CI_n$ *as defined in (4.11) and evaluated at* $\hat{c}_{n,q}(\theta) = \hat{c}_{n,q}(\hat{\theta}_q)$ *for* $q \in \{-p,p\}$ *and for all* $\theta \in \Theta$ *in (4.12), where* $\hat{\theta}_q \in \arg\max_{\theta\in\hat{\Theta}_I} q'\theta$ *and* $\hat{\Theta}_I = \{\theta \in \Theta : \bar{m}_{n,j}(\theta) \leqslant 0,\ j = 1,\ldots,J\}$.

*(III) If Assumptions D.2-(2) (for either* $p$ *or* $-p$ *or both) and D.5 hold, then setting* $\rho = +\infty$ *to obtain* $\hat{c}_{n,q}(\hat{\theta}_q)$ *in (4.12) and using these values for* $q \in \{-p,p\}$ *for each* $\theta \in \Theta$ *in computing* $CI_n$ *as defined in (4.11), we have that (4.13) continues to be satisfied.*

REMARK 4.2: If Theorem 4.3-(II) applies and the standardized moment conditions in (2.5) are linear in $\theta$, then $CI_n$ can be computed by solving just two linear programs.

Assumption D.2-(1) in Theorem 4.3-(I) ensures that some point in $\{p'\theta, \theta \in \Theta_I(P)\}$ is covered with probability approaching 1. Hence, the inference problem is effectively one-sided at the projection's end points and degenerate in between. It then suffices to intersect two one-sided $(1-\alpha)$-confidence intervals. Under Assumptions 4.1-4.5, Assumption D.2 is implied both by a "degeneracy condition"in Chernozhukov, Hong, and Tamer (2007) and by an assumption in Pakes, Porter, Ho, and Ishii (2011). A simple sufficient condition is that there exists a parameter value at which all population constraints hold with slack.

Assumptions D.3 and D.4 in Theorem 4.3-(II) are logically independent "polynomial minorant" conditions imposed in Chernozhukov, Hong, and Tamer (2007) and Bugni, Canay, and Shi (2017). Jointly, they assure that the sample support set $H(p, \hat{\Theta}_I)$ is an "inner consistent" estimator of the population support set $H(p, \Theta_I)$.[27] That is, any accumulation point of a selection from $H(p, \hat{\Theta}_I)$ is in $H(p, \Theta_I)$, but $H(p, \hat{\Theta}_I)$ may be much smaller than $H(p, \Theta_I)$. Then for one-sided inference, it suffices to compute $\hat{c}_n(\theta)$ exactly once, namely at one arbitrary selection $\hat{\theta} \in H(p, \hat{\Theta}_I)$, and to set $\hat{c}_n(\theta) = \hat{c}_n(\hat{\theta})$ for all $\theta$. We again remark that these conditions are implied by assumptions in Pakes, Porter, Ho, and Ishii (2011).

Assumptions D.2-(2) and D.5 in Theorem 4.3-(III) yield that the support set is a singleton and the tangent cone at the support set is pointy (in a uniform sense). We show that, in this case, the $\rho$-box constraints can be entirely dropped. This assumption is directly imposed by Pakes, Porter, Ho, and Ishii (2011), but we weaken it by showing that it is only needed in a local sense; hence, it suffices that the support set consists of distinct extreme points and all corresponding tangent cones are pointy.

**Result 3: A comparison with BCS-profiling**. We finally compare calibrated projection to BCS-profiling in well behaved cases. Suppose that Theorem 4.3 applies. Then $CI_n$ is the intersection of two one-sided confidence intervals and we can set $\rho = +\infty$. Hence, a scalar $s$ is in the one-sided (unbounded from below) confidence interval for $p'\theta$ if

$$\min_{p'\theta = s} T_n(\theta) \leqslant \hat{c}_n(\hat{\theta}_p), \tag{4.14}$$

$$T_n(\theta) \equiv \sqrt{n} \max_j \bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta). \tag{4.15}$$

While it was not originally constructed in this manner, this simplified confidence interval is the lower contour set of a profiled test statistic.[28] Indeed, up to an inconsequential squaring, $T_n$ is a special case of the statistic used in Bugni, Canay, and Shi (2017). This raises the question of how the tests compare. In the especially regular case where all parts of Theorem 4.3 apply, and assuming that calibrated projection is implemented with the corresponding simplifications, the answer is as follows:

[27] For a given unit vector $p$ and compact set $A \subset \mathbb{R}^d$, the *support set* of $A$ is $H(p, A) \equiv \arg\max_{a \in A} p'a$.

[28] By contrast, the corresponding expression without Theorem 4.3-(II) is $\min_{p'\theta = s}\{T_n(\theta) - \hat{c}_n(\theta)\} \leqslant 0$, which is not usefully interpreted as test inversion.

THEOREM 4.4: *Suppose Assumptions 4.1, 4.2, 4.3, 4.4, 4.5, D.2, D.3, D.4, D.5, and D.6 hold. Let BCS-profiling be implemented with the criterion function in equation (4.15) and GMS function $\varphi(x) = \min\{0, x\}$.*[29] *Let calibrated projection be implemented using the simplifications from Theorem 4.3, including setting $\rho = +\infty$. If both methods furthermore use the same $\kappa_n$, they are uniformly asymptotically equivalent:*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{Q}} \inf_{s\in[\min_{\theta\in\Theta} p'\theta, \max_{\theta\in\Theta} p'\theta]} P\left(\mathbf{1}\{s \in CI_n\} = \mathbf{1}\{s \in CI_n^{prof}\}\right) \to 1,$$

*where $CI_n^{prof}$ denotes the confidence interval resulting from the BCS-profiling method.*

Thus there is strong agreement between methods in extremely well-behaved cases.[30] We also show in Online Appendix F.1 that, in a further specialization of the above setting, finite sample power is higher with calibrated projection. This effect is due to a conservative distortion of order $1/\kappa_n$ in Bugni, Canay, and Shi (2017) and therefore vanishes asymptotically; however, due to the slow rate at which $\kappa_n$ diverges, it can be large in samples of considerable size. In sum, the approaches are not ranked in terms of power in empirically relevant examples.

### 4.3 Role of the $\rho$-box Constraints and Heuristics for Choosing $\rho$

When we use the bootstrap to calibrate $\hat{c}_n(\cdot)$, we restrict the localization vector $\lambda$ to lie in a $\rho$-box; see equation (3.1). This restriction has a crucial regularization effect. Comparing (2.7) and (3.4), it is apparent that we estimate coverage probabilities by replacing a nonlinear program with a linear one. It is intuitive that a Karush-Kuhn-Tucker condition (with uniformly bounded Lagrange multipliers) is needed for this to work (uniformly), and also that the linearization in (2.8) should be uniformly valid. But direct imposition of a Karush-Kuhn-Tucker condition would amount to a hard-to-verify constraint qualification. Rather than doing this, we show that Assumption 4.3 and imposition of the $\rho$-box constraints jointly yield such constraint qualification conditions on the set $\Lambda_n^b(\theta, \rho, c)$ (defined in (3.1)) with arbitrarily high probability for $n$ large enough, as well as uniform validity of the linearization. If one knows (or assumes) a priori that the population (limit) counterpart of the constraint set in (2.7) is contained in a ball with a radius bounded in probability (see Assumption D.1-(II) in Online Appendix D.2.2), then $\rho$ can be set equal to $+\infty$. The assumptions in Theorem 4.3-(III) are sufficient for this condition to hold.[31]

In practice, the choice of $\rho$ requires trading off how much conservative bias one is willing to bear in well-behaved cases against how much finite-sample size distortion one is willing

[29] The restriction on the GMS function is needed only because the "penalized resampling" approximation in BCS employs a specific "slackness function" equal to $\hat{\xi}_{n,j}$.

[30] This is not true for Pakes, Porter, Ho, and Ishii (2011) because they do not studentize the moment inequalities.

[31] See Online Appendix D.1 for proofs of these statements.

to bear in ill-behaved cases.[32] We propose a heuristic approach to calibrate $\rho$ focusing on conservative bias in the well behaved cases just considered, i.e. cases such as those characterized in Assumptions D.2, D.3, D.4, D.5 and D.6, in which the $\rho$-box could be dropped. In these cases, the optimal value of each of the two programs in equation (3.4) 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 $\eta = 1 - [1 - 2\Phi(-\rho)]^d$. The parameter $\rho$ can therefore be calibrated to achieve a conservative bias of at most $\eta$. When $J_1 + J_2 > d$, we propose to calibrate $\rho$ using the benchmark

$$\eta = 1 - [1 - 2\Phi(-\rho)]^{d\binom{J_1+J_2}{d}}, \tag{4.16}$$

again achieving a target conservative bias (in well-behaved cases) of $\eta$. For a few numerical examples, set $\eta = 0.01$: then $J_1 + J_2 = 10$ and $d = 3$ imply $\rho = 4.2$, whereas $J_1 + J_2 = 100$ and $d = 10$ imply $\rho = 8.4$. In the Monte Carlo experiments of Section 5, we investigate sensitivity of calibrated projection to the choice of $\rho$.

# 5 Monte Carlo Simulations

We evaluate the statistical and numerical performance of calibrated projection and EAM in two sets of Monte Carlo experiments run on a server with two Intel Xeon X5680 processors rated at 3.33GHz with 6 cores each and with a memory capacity of 24Gb rated at 1333MHz.[33] Both simulate a two-player entry game. The first experiment compares calibrated projection and BCS-profiling in the Monte Carlo exercise of BCS, using their code.[34] The other experiments feature a considerably more involved entry model with and without correlated unobservables. We were unable to numerically implement BCS-profiling for this model.[35]

## 5.1 The General Entry Game Model

We consider a two player entry game based on Ciliberto and Tamer (2009):

| | $Y_2 = 0$ | $Y_2 = 1$ |
|---|---|---|
| $Y_1 = 0$ | $0, 0$ | $0, Z_2'\zeta_1 + u_2$ |
| $Y_1 = 1$ | $Z_1'\zeta_1 + u_1, 0$ | $Z_1'(\zeta_1 + \Delta_1) + u_1, Z_2'(\zeta_2 + \Delta_2) + u_2$ |

Here, $Y_\ell$, $Z_\ell$, and $u_\ell$ denote player $\ell$'s binary action, observed characteristics, and unobserved characteristics. The strategic interaction effects $Z_\ell'\Delta_\ell \leqslant 0$ measure the impact of the opponent's entry into the market. We let $X \equiv (Y_1, Y_2, Z_1', Z_2')'$. We generate $Z = (Z_1, Z_2)$ as

[32] In Kaido, Molinari, and Stoye (2017) we provide examples of well-behaved and ill-behaved cases.

[33] To run the more than 120 distinct simulations reported here, we employed multiple servers. We benched the relative speed of each and report average computation time normalized to the server just described.

[34] See http://qeconomics.org/ojs/index.php/qe/article/downloadSuppFile/431/1411.

[35] For implementations of calibrated projection with real-world data, we refer the reader to Mohapatra and Chatterjee (2015), where $d = 5$, $J_1 = 44$, and $J_2 = 0$.

an i.i.d. random vector taking values in a finite set whose distribution $p_z = P(Z = z)$ is known. We let $u = (u_1, u_2)$ be independent of $Z$ and such that $Corr(u_1, u_2) \equiv r \in [0, 1]$ and $Var(u_\ell) = 1, \ell = 1, 2$. We let $\theta \equiv (\zeta_1', \zeta_2', \Delta_1', \Delta_2', r)'$. For a given set $A \subset \mathbb{R}^2$, we define $G_r(A) \equiv P(u \in A)$. We choose $G_r$ so that the c.d.f. of $u$ is continuous, differentiable, and has a bounded p.d.f. The outcome $Y = (Y_1, Y_2)$ results from pure strategy Nash equilibrium play. For some value of $Z$ and $u$, the model predicts monopoly outcomes $Y = (0, 1)$ and $(1, 0)$ as multiple equilibria. When this occurs, we select outcome $(0, 1)$ by independent Bernoulli trials with parameter $\mu \in [0, 1]$. This gives rise to the following restrictions:

$$E[1\{Y = (0,0)\}1\{Z = z\}] - G_r((-\infty, -z_1'\zeta_1) \times (-\infty, -z_2'\zeta_2))p_z = 0 \tag{5.1}$$

$$E[1\{Y = (1,1)\}1\{Z = z\}] - G_r([-z_1'(\zeta_1 + \Delta_1), +\infty) \times [-z_2'(\zeta_2 + \Delta_2), +\infty))p_z = 0 \tag{5.2}$$

$$E[1\{Y = (0,1)\}1\{Z = z\}] - G_r((-\infty, -z_1'(\zeta_1 + \Delta_1)) \times [-z_2'\zeta_2, +\infty))p_z \leqslant 0 \tag{5.3}$$

$$\begin{aligned} -E[1\{Y = (0,1)\}1\{Z = z\}] + \Big[ & G_r((-\infty, -z_1'(\zeta_1 + \Delta_1)) \times [-z_2'\zeta_2, +\infty) \\ & - G_r([-z_1'\zeta_1, -z_1'(\zeta_1 + \Delta_1)) \times [-z_2'\zeta_2, -z_2'(\zeta_2 + \Delta_2))\Big] p_z \leqslant 0. \end{aligned} \tag{5.4}$$

We show in Online Appendix C that this model satisfies Assumptions B.1 and 4.3-(II).[36] Throughout, we analytically compute the moments' gradients and studentize them using sample analogs of their standard deviations.

## 5.2 Specific Implementations and Results

**Set 1: A comparison with BCS-Profiling**

BCS specialize this model as follows. First, $u_1, u_2$ are independently uniformly distributed on $[0, 1]$ and the researcher knows $r = 0$. Equality (5.1) disappears because $(0, 0)$ is never an equilibrium. Next, $Z_1 = Z_2 = [1; \{W_k\}_{k=0}^{d_W}]$, where $W_k$ are observed market type indicators, $\Delta_\ell = [\delta_\ell; 0_{d_W}]$ for $\ell = 1, 2$, and $\zeta_1 = \zeta_2 = \zeta = [0; \{\zeta^{[k]}\}_{k=0}^{d_W}]$.[37] The parameter vector is $\theta = [\delta_1; \delta_2; \zeta]$ with parameter space $\Theta = \{\theta \in \mathbb{R}^{2+d_W} : (\delta_1, \delta_2) \in [0, 1]^2, \ \zeta_k \in [0, \min\{\delta_1, \delta_2\}], \ k = 1, \ldots, d_W\}$. This leaves 4 moment equalities and 8 moment inequalities (so $J = 16$); compare equation (5.1) in BCS. We set $d_W = 3$, $P(W_k = 1) = 1/4, k = 0, 1, 2, 3$, $\theta = [0.4; 0.6; 0.1; 0.2; 0.3]$, and $\mu = 0.6$. The implied true bounds on parameters are $\delta_1 \in [0.3872, 0.4239]$, $\delta_2 \in [0.5834, 0.6084]$, $\zeta^{[1]} \in [0.0996, 0.1006]$, $\zeta^{[2]} \in [0.1994, 0.2010]$, and $\zeta^{[3]} \in [0.2992, 0.3014]$.

The BCS-profiling confidence interval $CI_n^{prof}$ inverts a test of $H_0 : p'\theta = s_0$ over a grid for $s_0$. We do not in practice exhaust the grid but search inward from the extreme points of $\Theta$ in directions $\pm p$. At each $s_0$ that is visited, we compute (the square of) a profiled test statistic

[36] The specialization in which we compare to BCS also fulfils their assumptions. The assumptions in Pakes, Porter, Ho, and Ishii (2011) exclude any DGP that has moment equalities.

[37] This allows for market-type homogeneous fixed effects but not for player-specific covariates nor for observed heterogeneity in interaction effects.

$\min_{p'\theta=s_0} T_n(\theta)$; see equations (4.14)-(4.15) above. The corresponding critical value $\hat{c}_n^{prof}(s_0)$ is a quantile of the minimum of two distinct bootstrap approximations, each of which solves a nonlinear program for each bootstrap draw. Computational cost quickly increases with grid resolution, bootstrap size, and the number of starting points used to solve the nonlinear programs.

Calibrated projection computes $\hat{c}_n(\theta)$ by solving a series of linear programs for each bootstrap draw.[38] It computes the extreme points of $CI_n$ by solving NLP (2.5) twice, a task that is much accelerated by the E-A-M algorithm. Projection of Andrews and Soares (2010) operates very similarly but computes its critical value $\hat{c}_n^{proj}(\theta)$ through bootstrap simulation without any optimization.

We align grid resolution in BCS-profiling with the E-A-M algorithm's convergence threshold of 0.005.[39] We run all methods with $B = 301$ bootstrap draws, and calibrated and "uncalibrated" (i.e., based on Andrews and Soares (2010)) projection also with $B = 1001$.[40] Some other choices differ: BCS-profiling is implemented with their own choice to multi-start the nonlinear programs at 3 oracle starting points, i.e. using knowledge of the true DGP; our implementation of both other methods multi-starts the nonlinear programs from 30 data dependent random points (see Kaido, Molinari, Stoye, and Thirkettle (2017) for details).

Table 1 displays results for $(\delta_1, \delta_2)$ and for 300 Monte Carlo repetitions of all three methods. All confidence intervals are conservative, reflecting the effect of GMS. As expected, uncalibrated projection is most conservative, with coverage of essentially 1. Also, BCS-profiling is more conservative than calibrated projection. We suspect this relates to the conservative effect highlighted in Online Appendix F.1. The most striking contrast is in computational effort, where uncalibrated projection is fastest but calibrated projection also beats BCS-profiling by a factor of about 78. There are two effects at work here: First, because the calibrated projection bootstrap iterates over linear programs, it is much faster than the BCS-profiling one. Second, both uncalibrated projection and calibrated projection confidence intervals were computed using the E-A-M algorithm. Indeed, the computation times reported for uncalibrated projection indicate that, in contrast to received wisdom, this procedure is computationally somewhat easy. This is due to the E-A-M algorithm and therefore part of this paper's contribution.

Table 2 extends the analysis to all components of $\theta$ and to 1000 Monte Carlo repetitions. We were unable to compute this or any of the next tables for BCS-profiling.

**Set 2: Heterogeneous interaction effects and potentially correlated errors**

[38] We implement this step using the high-speed solver CVXGEN, available from http://cvxgen.com and described in Mattingley and Boyd (2012).

[39] This is only one of several individually necessary stopping criteria. Others include that the current optimum $\theta^{*,L}$ and the expected improvement maximizer $\theta^{L+1}$ (see equation (3.13)) satisfy $|p'(\theta^{L+1} - \theta^{*,L})| \leqslant 0.005$. See Kaido, Molinari, Stoye, and Thirkettle (2017) for the full list of convergence requirements.

[40] Based on some trial runs of BCS-profiling for $\delta_1$, we estimate that running it with $B = 1001$ throughout would take 3.14-times longer than the computation times reported in Table 1. By comparison, calibrated projection takes only 1.75-times longer when implemented with $B = 1001$ instead of $B = 301$.

In our second set of experiments, we let $u = (u_1, u_2)$ be bivariate Normal with (nondegenerate) correlation $r$, so all outcomes have positive probability. We let $Z$ include a constant and a player specific, binary covariate, so $Z_1 \in \{(1,-1),(1,1)\}$ and $Z_2 \in \{(1,-1),(1,1)\}$. This implies $J_1 = J_2 = 8$, hence $J = 24$. The marginal distribution of $(Z_1^{[2]}, Z_2^{[2]})$ is multinomial with weights $(0.1, 0.2, 0.3, 0.4)$ on $((-1,-1),(-1,1),(1,-1),(1,1))$.

In our Set 2-DGP1, we set $\zeta_1 = (.5, .25)'$, $\Delta_1 = (-1,-1)'$, and $r = 0$. Set 2-DGP2 differs by setting $\Delta_1 = (-1,-.75)'$. In both cases, $(\zeta_2, \Delta_2) = (\zeta_1, \Delta_1)$ and $\mu = 0.5$; we only report results for $(\zeta_1, \Delta_1)$. Although parameter values are similar, there is a qualitative difference: In DGP1, parameters are point identified; in DGP2, they are not but the true bounds ($\zeta_1^{[1]} \in [0.405, 0.589]$, $\zeta_1^{[2]} \in [0.236, 0.266]$, $\Delta_1^{[1]} \in [-1.158, -0.832]$, $\Delta_1^{[2]} \in [-0.790, -0.716]$) are not wide compared to sampling uncertainty. We therefore expect all methods that use GMS to be conservative in DGP2.[41] In both Set 2-DGP1& DGP2 we use knowledge that $r = 0$, so that $d = 8$. Our Set 2-DGP3 preserves the same payoff parameters values as in Set 2-DGP2 but sets $r = 0.5$ and this parameter is also unknown, so that $d = 9$.

Within Set 2-DGP2, we also experiment with the sensitivity of coverage probability and length of $CI_n$ to the choice of $\rho$ and $\kappa_n$. We consider choices of $\rho$ that are (1) very large or "liberal", so that in well behaved cases the $\rho$-box constraints induce an amount $\eta$ of overcoverage in $CI_n$ smaller than machine precision (see equation (4.16)); (2) "default", so that $\eta = 0.01$; (3) small or "conservative", so that $\eta = 0.025$. For $\kappa_n$, we have experimented with a "conservative" choice $\kappa_n = n^{1/7}$, and a "liberal" choice $\kappa_n = \sqrt{\ln \ln n}$, while out "default" is $\kappa_n = \sqrt{\ln n}$.

Results are reported in Tables 3 through 7. An interesting feature of Table 3 is that in this (point identified) DGP, calibrated projection is not conservative at all. This presumably reflects an absence of near-binding inequalities. Conservative bias is larger in the partially identified Set 2-DGP2 in Table 4. For these two tables, we do note the increased computational advantage of uncalibrated projection over calibrated projection. This advantage is bound to increase as DGP's, and therefore the linear programs iterated over in the bootstrap, become more complex. Table 5 shows that allowing for correlation of the errors does not change the results much in terms of the confidence intervals' length and coverage probabilities. However, due to the repeated evaluation of the bivariate normal CDFs, both calibrated and uncalibrated projection have higher computational time than the case with $r = 0$. Another feature to note is that both confidence intervals for $r$ tend to be wide although the projection of $\Theta_I$ is short, which suggests that this component may be weakly identified.

Table 6 examines the effect of varying the tuning parameter $\rho$. Increasing $\rho$ necessarily (weakly) decreases length and also coverage of intervals, and this effect is evident in the table but is arguably small. This is even more the case for the GMS tuning parameter $\kappa_n$. Numerically, for $n = 4000$, the values explored in the table are rather different at

[41] We also note that this is a case where non-uniform methods may severely undercover in finite sample.

$4000^{1/7} \approx 3.27$ and $\sqrt{\ln(\ln(4000))} \approx 1.45$, but the effect on inference is very limited, see Table 7. Indeed, differences in coverage are so small that reported results are occasionally slightly nonmonotonic, reflecting numerical and simulation noise.

# 6 Conclusions

This paper introduces a computationally attractive confidence interval for linear functions of parameter vectors that are partially identified through finitely many moment (in)equalities. The extreme points of our *calibrated projection* confidence interval are obtained by minimizing and maximizing $p'\theta$ subject to properly relaxed sample analogs of the moment conditions. The relaxation amount, or critical level, is computed to insure uniform asymptotic coverage of $p'\theta$ rather than $\theta$ itself. Its calibration is computationally attractive because it is based on repeatedly checking feasibility of (bootstrap) 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 is of independent interest in the partial identification literature. Indeed, a key contribution of the paper is to establish convergence of this algorithm.

Our Monte Carlo analysis shows that, in the DGPs that we considered, calibrated projection is fast and accurate: Computation of the confidence intervals is orders of magnitude faster than for the main alternative to our method, a profiling-based procedure due to Bugni, Canay, and Shi (2017). The class of DGPs over which we can establish uniform validity of our procedure is non-nested with corresponding class for the alternative method. Important cases covered here but not elsewhere include linear functions of best linear predictor parameters with interval valued outcomes and discrete covariates. The price to pay for this generality is the use of one additional (non-drifting) tuning parameter. We provide conditions under which this parameter can be eliminated and compare the power properties of calibrated projection and BCS-profiling. The false coverage properties of the two methods are non-ranked but are asymptotically the same in very well-behaved cases. We establish considerable finite sample advantage in a specific case.

Similarly to confidence regions proposed in Andrews and Soares (2010), Bugni, Canay, and Shi (2017), Stoye (2009), and elsewhere, our confidence interval can be empty, namely if sample violations of moment inequalities exceed $\hat{c}_n(\theta)$ at each $\theta$. This event can be interpreted as rejection of maintained assumptions. See Stoye (2009) and especially Andrews and Soares (2010) for further discussion and Bugni, Canay, and Shi (2015) for a paper that focuses on this interpretation and improves on $\hat{c}_n^{proj}$ for the purpose of specification testing. We leave a detailed analysis of our implied specification test to future research.

# A Convergence of the E-A-M Algorithm

In this appendix, we provide details on the algorithm used to solve the outer maximization problem as described in Section 3.2. Below, let $(\Omega, \mathcal{F})$ be a measurable space and $\omega$ a generic element of $\Omega$. Let $L \in \mathbb{N}$ and let $(\theta^{(1)}, \cdots, \theta^{(L)})$ be a measurable map on $(\Omega, \mathcal{F})$ whose law is specified below. The value of the function $c$ in (3.6) is unknown ex ante. Once the evaluation points $\theta^{(\ell)}, \ell = 1, \cdots, L$ realize, the corresponding values of $c$, i.e. $\Upsilon^{(\ell)} \equiv c(\theta^{(\ell)}), \ell = 1, \cdots, L$, are known. We may therefore define the information set

$$\mathcal{F}_L \equiv \sigma(\theta^{(\ell)}, \Upsilon^{(\ell)}, \ell = 1, \cdots, L). \tag{A.1}$$

We note that $\theta^{*,L} \equiv \text{argmax}_{\theta \in \mathcal{C}^L} p'\theta$ is measurable with respect to $\mathcal{F}_L$.

Our algorithm iteratively determines evaluation points based on the *expected improvement* (Jones, Schonlau, and Welch, 1998). For this, we formally introduce a model that describes the uncertainty associated with the values of $c$ outside the current evaluation points. Specifically, the unknown function $c$ is modeled as a Gaussian process such that[42]

$$\mathbb{E}[c(\theta)] = \mu, \ \mathbb{C}\text{ov}(c(\theta), c(\theta')) = \varsigma^2 K_\beta(\theta - \theta'), \tag{A.2}$$

where $\beta = (\beta_1, \cdots, \beta_d) \in \mathbb{R}^d$ controls the length-scales of the process. Two values $c(\theta)$ and $c(\theta')$ are highly correlated when $\theta_k - \theta'_k$ is small relative to $\beta_k$. Throughout, we assume $\underline{\beta}_k \leqslant \beta_k \leqslant \overline{\beta}_k$ for some $0 < \underline{\beta}_k < \overline{\beta}_k < \infty$ for $k = 1, \cdots, d$. We let $\bar{\beta} = (\bar{\beta}_1, \cdots, \bar{\beta}_d)' \in \mathbb{R}^d$. Specific suggestions on the forms of $K_\beta$ are given in Appendix B.2.

For a given $(\mu, \varsigma, \beta)$, the posterior distribution of $c$ given $\mathcal{F}_L$ is then another Gaussian process whose mean $c_L(\cdot)$ and variance $\varsigma^2 s_L^2(\cdot)$ are given as follows (Santner, Williams, and Notz, 2013, Section 4.1.3):

$$c_L(\theta) = \mu + \mathbf{r}_L(\theta)' \mathbf{R}_L^{-1}(\boldsymbol{\Upsilon} - \mu \mathbf{1}) \tag{A.3}$$

$$\varsigma^2 s_L^2(\theta) = \varsigma^2 \Big(1 - \mathbf{r}_L(\theta)' \mathbf{R}_L^{-1} \mathbf{r}_L(\theta) + \frac{(1 - \mathbf{1}' \mathbf{R}_L^{-1} \mathbf{r}_L(\theta))^2}{\mathbf{1}' \mathbf{R}_L^{-1} \mathbf{1}}\Big). \tag{A.4}$$

Given this, the expected improvement function can be written as

$$\begin{aligned}
\mathbb{E}\mathbb{I}_L(\theta) &\equiv \mathbb{E}[(p'\theta - p'\theta^{*,L})_+ 1\{\bar{g}(\theta) \leqslant c(\theta)\} | \mathcal{F}_L] \\
&= (p'\theta - p'\theta^{*,L})_+ \mathbb{P}(c(\theta) \geqslant \max_{j=1,\cdots,J} g_j(\theta) | \mathcal{F}_L) \\
&= (p'\theta - p'\theta^{*,L})_+ \mathbb{P}\left( \frac{c(\theta) - c_L(\theta)}{\varsigma s_L(\theta)} \geqslant \frac{\max_{j=1,\cdots,J} g_j(\theta) - c_L(\theta)}{\varsigma s_L(\theta)} \right) \\
&= (p'\theta - p'\theta^{*,L})_+ \Big(1 - \Phi\Big( \frac{\bar{g}(\theta) - c_L(\theta)}{\varsigma s_L(\theta)} \Big)\Big),
\end{aligned} \tag{A.5}$$

The evaluation points $(\theta^{(1)}, \cdots, \theta^{(L)})$ are then generated according to the following algorithm (**M-step** in Section 3.2).

[42] We use $\mathbb{P}$ and $\mathbb{E}$ to denote the probability and expectation for the prior and posterior distributions of $c$ to distinguish them from $P$ and $E$ used for the sampling uncertainty for $X_i$.

ALGORITHM A.1: *Let $k \in \mathbb{N}$.*
*Step 1: Initial evaluation points $\theta^{(1)}, \cdots, \theta^{(k)}$ are drawn randomly independent of $c$.*
*Step 2: For $L \geqslant k$, with probability $1-\epsilon$, let $\theta^{(L+1)} = argmax_{\theta \in \Theta} \mathbb{EI}_L(\theta)$. With probability $\epsilon$, draw $\theta^{(L+1)}$ uniformly at random from $\Theta$.*

Below, we use $\mathbb{Q}$ to denote the law of $(\theta^{(1)}, \cdots, \theta^{(L)})$ determined by the algorithm above. We also note that $\theta^{*,L+1} = \arg\max_{\theta \in \mathcal{C}^{L+1}} p'\theta$ is a function of the evaluation points and therefore is a random variable whose law is governed by $\mathbb{Q}$.

## A.1 Proof of Theorem 3.1

*Proof.* We adopt the method used in the proof of Theorem 5 in Bull (2011), who proves a convergence result for an unconstrained optimization problem in which the objective function is unknown outside the evaluation points.

Below, we let $L \geqslant 2k$. Let $0 < \nu < \infty$. Let $0 < \eta < \epsilon$ and $A_L \in \mathcal{F}$ be the event that at least $\lfloor \eta L \rfloor$ of the points $\theta^{(k+1)}, \cdots, \theta^{(L)}$ are drawn independently from a uniform distribution on $\Theta$. Let $B_L \in \mathcal{F}$ be the event that one of the points $\theta^{(L+1)}, \cdots, \theta^{(2L)}$ is chosen by maximizing the expected improvement. For each $L$, define the mesh norm:

$$h_L \equiv \sup_{\theta \in \Theta} \min_{\ell = 1, \cdots L} \|\theta - \theta^{(\ell)}\|. \tag{A.6}$$

For a given $\bar{M} > 0$, let $C_L \in \mathcal{F}$ be the event that $h_L \leqslant \bar{M}(L/\ln L)^{-1/d}$. We then let

$$D_L \equiv A_L \cap B_L \cap C_L. \tag{A.7}$$

On $D_L$, the following results hold. First, let $\beta_L$ be the estimated parameter. Noting that there are $\lfloor \eta L \rfloor$ uniformly sampled points and arguing as in (A.24)-(A.25), it follows that

$$\sup_{\theta \in \Theta} s_L(\theta; \beta_L) \leqslant M r_L, \tag{A.8}$$

for some constant $M > 0$ by $\omega \in C_L$, and $r_L$ is defined by

$$r_L \equiv (L/\ln L)^{-\nu/d}. \tag{A.9}$$

For later use, we note that, for any $L \geqslant 2$,

$$r_{L-1}/r_L = \Big(\frac{L}{L-1}\Big)^{\nu/d} \Big(\frac{\ln(L-1)}{\ln L}\Big)^{\nu/d} \leqslant 2^{\nu/d}. \tag{A.10}$$

Second, by $\omega \in B_L$, there is $\ell$ such that $L \leqslant \ell \leqslant 2L$ and $\theta^{(\ell)}$ is chosen by maximizing the expected improvement. For $\theta \in \Theta$ and $L \in \mathbb{N}$, let $I_L(\theta) \equiv (p'\theta - p'\theta^{*,L})_+ 1\{\bar{g}(\theta) \leqslant c(\theta)\}$. Recall that $\theta^*$ is an

optimal solution to (3.6). Then,

$$
\begin{aligned}
p'\theta^* - p'\theta^{*,\ell-1} &\overset{(1)}{=} I_{\ell-1}(\theta^*) \\
&\overset{(2)}{\leqslant} \mathbb{E}\mathbb{I}_{\ell-1}(\theta^*)\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1} \\
&\overset{(3)}{\leqslant} \mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)})\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1} \\
&\overset{(4)}{\leqslant} \Big(I_{\ell-1}(\theta^{(\ell)}) + M_1 s_{\ell-1}(\theta^{(\ell)})\exp(-M_2 s_{\ell-1}(\theta^{(\ell)})^{-2})\Big)\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1} \\
&\overset{(5)}{\leqslant} \Big(I_{\ell-1}(\theta^{(\ell)}) + MM_1 r_{\ell-1}\exp(-M^{-2}M_2 r_{\ell-1}^{-2})\Big)\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1} \\
&\overset{(6)}{\leqslant} \Big(I_{\ell-1}(\theta^{(\ell)}) + 2^{\nu/d}MM_1 r_{\ell}\exp(-(2^{\nu/d}M)^{-2}M_2 r_{\ell}^{-2})\Big)\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1} \\
&= \Big((p'\theta^{(\ell)} - p'\theta^{*,\ell-1})1\{\bar{g}(\theta^{(\ell)}) \leqslant c(\theta^{(\ell)})\} + 2^{\nu/d}MM_1 r_{\ell}\exp(-(2^{\nu/d}M)^{-2}M_2 r_{\ell}^{-2})\Big)\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1} \\
&\overset{(7)}{\leqslant} \Big((p'\theta^{*,\ell} - p'\theta^{*,\ell-1}) + 2^{\nu/d}MM_1 r_{\ell}\exp(-(2^{\nu/d}M)^{-2}M_2 r_{\ell}^{-2})\Big)\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1} \\
&\overset{(8)}{\leqslant} \Big(h_{\ell} + 2^{\nu/d}MM_1 r_{\ell}\exp(-(2^{\nu/d}M)^{-2}M_2 r_{\ell}^{-2})\Big)\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1}, \qquad \text{(A.11)}
\end{aligned}
$$

where (1) follows by construction, (2) follows from Lemma A.1 (ii), (3) follows from $\theta^{(\ell)}$ being the maximizer of the expected improvement, (4) follows from Lemma A.1 (i), (5) follows from (A.8), (6) follows from $r_{\ell-1} \leqslant 2^{\nu/d} r_{\ell}$ for $\ell \geqslant 2$ by (A.10), (7) follows from $\theta^{*,\ell} = \operatorname{argmax}_{\theta\in\mathcal{C}^{\ell}} p'\theta$, (8) follows from $p'\theta^{*,\ell} - p'\theta^{*,\ell-1}$ being dominated by the mesh-norm. Therefore, by $\omega \in C_L$, there exists a constant $M > 0$ such that

$$
p'\theta^* - p'\theta^{*,\ell-1} \leqslant \Big(M(\ell/\ln\ell)^{-1/d} + Mr_{\ell}\exp(-Mr_{\ell}^{-2})\Big)\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1}. \qquad \text{(A.12)}
$$

Since $L \leqslant \ell \leqslant 2L$, $p'\theta^{*,L}$ is non-decreasing in $L$, and $r_L$ is non-increasing in $L$, we have

$$
\begin{aligned}
p'\theta^* - p'\theta^{*,2L} &\leqslant \Big(M(L/\ln L)^{-1/d} + Mr_L\exp(-Mr_L^{-2})\Big)\big(1-\Phi\big(\frac{R}{\varsigma}\big)\big)^{-1} \\
&= O((2L/\ln 2L)^{-1/d}) + O(r_{2L}\exp(-Mr_{2L}^{-2})), \qquad \text{(A.13)}
\end{aligned}
$$

where the last equality follows from the existence of a positive constant $C$ such that $r_L = Cr_{2L}$ and redefining multiplying constants properly.

Now consider the case $\omega \notin D_L$. By (A.7),

$$
\mathbb{Q}(D_L^c) \leqslant \mathbb{Q}(A_L^c) + \mathbb{Q}(B_L^c) + \mathbb{Q}(C_L^c). \qquad \text{(A.14)}
$$

Let $Z_{\ell}$ be a Bernoulli random variable such that $Z_{\ell} = 1$ if $\theta^{(\ell)}$ is randomly drawn from a uniform distribution. Then, by the Chernoff bounds (see e.g. Boucheron, Lugosi, and Massart, 2013, p.48),

$$
\mathbb{Q}(A_L^c) = \mathbb{Q}\Big(\sum_{\ell=k+1}^{L} Z_{\ell} < \lfloor \eta L \rfloor\Big) \leqslant \exp(-(L-k+1)\epsilon(\epsilon-\eta)^2/2). \qquad \text{(A.15)}
$$

Further, by the definition of $B_L$,

$$\mathbb{Q}(B_L^c) = \epsilon^L, \tag{A.16}$$

and finally by taking $\bar{M}$ large upon defining the event $C_L$ and applying Lemma 4 in Bull (2011), one has

$$\mathbb{Q}(C_L^c) = O((L/\ln L)^{-\gamma}), \tag{A.17}$$

for any $\gamma > 0$. Combining (A.14)-(A.17), for any $\gamma > 0$,

$$\mathbb{Q}(D_L^c) = O((L/\ln L)^{-\gamma}). \tag{A.18}$$

Finally, noting that $p'\theta^* - p'\theta^{*,2L}$ is bounded by some constant $M > 0$ due to the boundedness of $\Theta$, we have

$$\begin{aligned} E_{\mathbb{Q}}\big[p'\theta^* - p'\theta^{*,2L}\big] &= \int_{D_L} p'\theta^* - p'\theta^{*,2L} d\mathbb{Q} + \int_{D_L^c} p'\theta^* - p'\theta^{*,2L} d\mathbb{Q} \\ &= O((2L/\ln 2L)^{-1/d}) + O(r_{2L}\exp(-Mr_{2L}^{-2})) + O((2L/\ln 2L)^{-\gamma}) = o(1), \end{aligned} \tag{A.19}$$

where the second equality follows from (A.13) and (A.18). This completes the proof. □

The following lemma is an analog of Lemma 8 in Bull (2011), which links the expected improvement to the actual improvement achieved by a new evaluation point $\theta$.

LEMMA A.1: *Suppose $\Theta \subset \mathbb{R}^d$ is bounded and $p \in \mathbb{S}^{d-1}$. Suppose the evaluation points $(\theta^{(1)}, \cdots, \theta^{(L)})$ are drawn by Algorithm A.1 and $\|c\|_{\mathcal{H}_{\tilde{g}}} \leqslant R$ for some $R > 0$. For $\theta \in \Theta$ and $L \in \mathbb{N}$, let $I_L(\theta) \equiv (p'\theta - p'\theta^{*,L})_+ 1\{\bar{g}(\theta) \leqslant c(\theta)\}$. Then, (i) there exist constants $M_j > 0, j = 1, 2$ that only depend on $(\varsigma, R)$ and an integer $\bar{L} \in \mathbb{N}$ such that*

$$\mathbb{E}\mathbb{I}_L(\theta) \leqslant I_L(\theta) + M_1 s_L(\theta)\exp(-M_2 s_L^{-2}(\theta)) \tag{A.20}$$

*for all $L \geqslant \bar{L}$. Further, (ii) for any $L \in \mathbb{N}$ and $\theta \in \Theta$,*

$$I_L(\theta) \leqslant \mathbb{E}\mathbb{I}_L(\theta)\Big(1 - \Phi\Big(\frac{R}{\varsigma}\Big)\Big)^{-1}. \tag{A.21}$$

**Proof of Lemma A.1**. (i) If $s_L(\theta) = 0$, then the posterior variance of $c(\theta)$ is zero. Hence, $\mathbb{E}\mathbb{I}_L(\theta) = I_L(\theta)$, and the claim of the lemma holds.

For $s_L(\theta) > 0$, we first show the upper bound. Let $u \equiv (\bar{g}(\theta) - c_L(\theta))/s_L(\theta)$ and $t \equiv (\bar{g}(\theta) -$

$c(\theta))/s_L(\theta)$. By Lemma 6 in Bull (2011), we have $|u-t| \leqslant R$. Since $1-\Phi(\cdot)$ is decreasing, we have

$$\begin{aligned}
\mathbb{E}\mathbb{I}_L(\theta) &= (p'\theta - p'\theta^{*,L})_+\Big(1-\Phi\Big(\frac{u}{\varsigma}\Big)\Big) \\
&\leqslant (p'\theta - p'\theta^{*,L})_+\Big(1-\Phi\Big(\frac{t-R}{\varsigma}\Big)\Big) \\
&= (p'\theta - p'\theta^{*,L})_+(1\{\bar{g}(\theta) \leqslant c(\theta)\} + 1\{\bar{g}(\theta) > c(\theta)\})\Big(1-\Phi\Big(\frac{t-R}{\varsigma}\Big)\Big) \\
&\leqslant I_L(\theta) + (p'\theta - p'\theta^{*,L})_+ 1\{\bar{g}(\theta) > c(\theta)\}\Big(1-\Phi\Big(\frac{t-R}{\varsigma}\Big)\Big), \qquad \text{(A.22)}
\end{aligned}$$

where the last inequality used $1-\Phi(x) \leqslant 1$ for any $x \in \mathbb{R}$. Note that one may write

$$1\{\bar{g}(\theta) > c(\theta)\}\Big(1-\Phi\Big(\frac{t-R}{\varsigma}\Big)\Big) = 1\{\bar{g}(\theta) > c(\theta)\}\Big(1-\Phi\Big(\frac{\bar{g}(\theta)-c(\theta)-s_L(\theta)R}{\varsigma s_L(\theta)}\Big)\Big). \qquad \text{(A.23)}$$

Below we assume $\bar{g}(\theta) > c(\theta)$ because otherwise, the expression above is 0, and the claim holds. To be clear about the parameter value at which we evaluate $s_L$, we will write $s_L(\theta;\beta)$. By the hypothesis that $\|c\|_{\mathcal{H}_{\bar{\beta}}} \leqslant R$ and Lemma 4 in Bull (2011), we have

$$\|c\|_{\mathcal{H}_{\beta_L}} \leqslant S, \qquad \text{(A.24)}$$

where $S = R^2 \prod_{k=1}^d (\overline{\beta}_k/\underline{\beta}_k)$. Note that there are $\lfloor \eta L \rfloor$ uniformly sampled points, and $K_\beta$ is associated with index $\nu \in (0,\infty), \nu \notin \mathbb{N}$. By Corollary 6.4 in Narcowich, Ward, and Wendland (2003),

$$\sup_{\theta\in\Theta} s_L(\theta;\beta) = O(M(\beta)h_L^\nu), \qquad \text{(A.25)}$$

uniformly in $\beta$, where $h_L = \sup_{\theta\in\Theta}\min_{\ell=1,\cdots L}\|\theta-\theta^{(\ell)}\|$ and $\beta \mapsto M(\beta)$ is a continuous function (note that the exponent $\nu$ in our notation matches matches $(k+\nu)/2$ in theirs). Hence, $s_L(\theta) = o(1)$. This, together with $\bar{g}(\theta) > c(\theta)$, implies that there are a constant $M$ and $\bar{L} \in \mathbb{N}$ such that

$$0 < M < (\bar{g}(\theta) - c(\theta) - s_L(\theta)R)/\varsigma, \ \forall L \geqslant \bar{L}. \qquad \text{(A.26)}$$

Therefore, again by the fact that $1-\Phi(\cdot)$ is decreasing, one obtains

$$\begin{aligned}
1\{\bar{g}(\theta) > c(\theta)\}\Big(1-\Phi\Big(\frac{\bar{g}(\theta)-c(\theta)-s_L(\theta)R}{\varsigma s_L(\theta)}\Big)\Big) &\leqslant \Big(1-\Phi\Big(\frac{M}{s_L(\theta)}\Big)\Big) \\
&\leqslant \frac{s_L(\theta)}{M}\phi\Big(\frac{M}{s_L(\theta)}\Big), \qquad \text{(A.27)}
\end{aligned}$$

where $\phi$ is the density of the standard normal distribution, and the last inequality follows from $1-\Phi(x) \leqslant \phi(x)/x$, which is due to Gordon (1941). The claim on the upper bound then follows from (A.22), $(p'\theta - p'\theta^{*,L}) \leqslant M$ for some $M > 0$ due to $\Theta$ being bounded, and (A.27).

(ii) For the lower bound in (A.21), we have

$$\begin{aligned}\mathbb{E}\mathbb{I}_L(\theta) &\geqslant (p'\theta - p'\theta^{*,L})_+\Big(1-\Phi\Big(\frac{t+R}{\varsigma}\Big)\Big)\\ &= (p'\theta - p'\theta^{*,L})_+ 1\{\bar{g}(\theta)\leqslant c(\theta)\}\Big(1-\Phi\Big(\frac{t+R}{\varsigma}\Big)\Big)\\ &\geqslant I_L(\theta)\Big(1-\Phi\Big(\frac{R}{\varsigma}\Big)\Big),\end{aligned} \tag{A.28}$$

where the last inequality follows from $t = (\bar{g}(\theta) - c(\theta))/s_L(\theta) \leqslant 0$ and the fact that $1-\Phi(\cdot)$ is decreasing. □

# Tables

Table 1: Results for Set 1 with $n = 4000$, $MCs = 300$, $B = 301$, $\rho = 5.04$, $\kappa_n = \sqrt{\ln n}$.

| | $1-\alpha$ | Median CI | | | $CI_n^{prof}$ Coverage | | $CI_n$ Coverage | | $CI_n^{proj}$ Coverage | | Average Time | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $CI_n^{prof}$ | $CI_n$ | $CI_n^{proj}$ | Lower | Upper | Lower | Upper | Lower | Upper | $CI_n^{prof}$ | $CI_n$ | $CI_n^{proj}$ |
| $\delta_1 = 0.4$ | 0.95 | [0.330,0.495] | [0.336,0.482] | [0.290,0.557] | 0.997 | 0.990 | 0.993 | 0.973 | 1 | 1 | 1858.42 | 22.86 | 13.82 |
| | 0.90 | [0.340,0.485] | [0.342,0.474] | [0.298,0.543] | 0.990 | 0.980 | 0.980 | 0.963 | 1 | 1 | 1873.23 | 22.26 | 15.81 |
| | 0.85 | [0.345,0.475] | [0.348,0.466] | [0.303,0.536] | 0.970 | 0.970 | 0.960 | 0.937 | 1 | 1 | 1907.84 | 23.00 | 13.98 |
| $\delta_2 = 0.6$ | 0.95 | [0.515,0.655] | [0.518,0.650] | [0.461,0.682] | 0.987 | 0.993 | 0.980 | 0.987 | 1 | 1 | 1753.54 | 23.84 | 19.10 |
| | 0.90 | [0.525,0.647] | [0.533,0.643] | [0.473,0.675] | 0.977 | 0.973 | 0.957 | 0.953 | 1 | 1 | 1782.91 | 24.45 | 17.16 |
| | 0.85 | [0.530,0.640] | [0.540,0.639] | [0.481,0.670] | 0.967 | 0.957 | 0.943 | 0.923 | 1 | 1 | 1809.65 | 23.38 | 17.33 |

Notes: (1) Projections of $\Theta_I$ are: $\delta_1 \in [0.3872, 0.4239]$, $\delta_2 \in [0.5834, 0.6084]$, $\zeta_1 \in [0.0996, 0.1006]$, $\zeta_2 \in [0.1994, 0.2010]$, $\zeta_3 \in [0.2992, 0.3014]$. (2) "Upper" coverage is for $\max_{\theta \in \Theta_I(P)} p'\theta$, and similarly for "Lower". (3) "Average time" is computation time in seconds averaged over MC replications. (4) $CI_n^{prof}$ results from BCS-profiling, $CI_n$ is calibrated projection, and $CI_n^{proj}$ is uncalibrated projection.

[31]

Table 2: Results for Set 1 with $n = 4000$, $MCs = 1000$, $B = 1001$, $\rho = 5.04$, $\kappa_n = \sqrt{\ln n}$.

| | $1-\alpha$ | Median CI | | $CI_n$ Coverage | | $CI_n^{proj}$ Coverage | | Average Time | |
|---|---|---|---|---|---|---|---|---|---|
| | | $CI_n$ | $CI_n^{proj}$ | Lower | Upper | Lower | Upper | $CI_n$ | $CI_n^{proj}$ |
| $\delta_1 = 0.4$ | 0.95 | [0.333,0.479] | [0.288,0.555] | 0.990 | 0.979 | 1 | 1 | 42.35 | 15.79 |
| | 0.90 | [0.342,0.470] | [0.296,0.542] | 0.978 | 0.957 | 1 | 1 | 41.13 | 11.60 |
| | 0.85 | [0.347,0.464] | [0.302,0.534] | 0.960 | 0.942 | 1 | 1 | 39.91 | 15.36 |
| $\delta_2 = 0.6$ | 0.95 | [0.526,0.653] | [0.466,0.683] | 0.969 | 0.978 | 1 | 1 | 41.40 | 24.30 |
| | 0.90 | [0.538,0.646] | [0.478,0.677] | 0.948 | 0.959 | 1 | 0.999 | 41.39 | 32.78 |
| | 0.85 | [0.545,0.642] | [0.485,0.672] | 0.925 | 0.941 | 1 | 1 | 38.49 | 31.55 |
| $\zeta^{[1]} = 0.1$ | 0.95 | [0.054,0.143] | [0.020,0.179] | 0.951 | 0.952 | 1 | 1 | 35.57 | 20.80 |
| | 0.90 | [0.060,0.137] | [0.028,0.171] | 0.916 | 0.916 | 0.998 | 0.998 | 38.42 | 28.07 |
| | 0.85 | [0.064,0.132] | [0.033,0.166] | 0.868 | 0.863 | 0.998 | 0.998 | 38.63 | 28.77 |
| $\zeta^{[2]} = 0.2$ | 0.95 | [0.156,0.245] | [0.120,0.281] | 0.950 | 0.949 | 1 | 1 | 35.99 | 18.07 |
| | 0.90 | [0.162,0.238] | [0.128,0.273] | 0.910 | 0.908 | 0.999 | 0.998 | 33.29 | 23.13 |
| | 0.85 | [0.166,0.235] | [0.133,0.268] | 0.869 | 0.863 | 0.995 | 0.995 | 33.76 | 17.33 |
| $\zeta^{[3]} = 0.3$ | 0.95 | [0.257,0.344] | [0.222,0.379] | 0.945 | 0.944 | 1 | 1 | 39.92 | 31.27 |
| | 0.90 | [0.262,0.337] | [0.230,0.371] | 0.896 | 0.900 | 0.998 | 0.998 | 43.37 | 29.17 |
| | 0.85 | [0.266,0.333] | [0.235,0.366] | 0.866 | 0.863 | 0.995 | 0.995 | 43.60 | 26.99 |

Notes: Same DGP and conventions as in Table 1.

Table 3: Results for Set 2-DGP1, $Corr(u_1, u_2) = 0$, $n = 4000$, $MCs = 1000$, $\rho = 6.02$, $\kappa_n = \sqrt{\ln n}$.

| | $1-\alpha$ | Median CI | | Coverage | | Average Time | |
|---|---|---|---|---|---|---|---|
| | | $CI_n$ | $CI_n^{proj}$ | $CI_n$ | $CI_n^{proj}$ | $CI_n$ | $CI_n^{proj}$ |
| $\zeta_1^{[1]} = 0.50$ | 0.95 | [0.355,0.715] | [0.127,0.938] | 0.948 | 1 | 82.34 | 23.56 |
| | 0.90 | [0.374,0.687] | [0.172,0.902] | 0.902 | 0.999 | 84.33 | 21.61 |
| | 0.85 | [0.387,0.669] | [0.200,0.878] | 0.856 | 0.996 | 87.33 | 22.31 |
| $\zeta_1^{[2]} = 0.25$ | 0.95 | [0.115,0.354] | [0.003,0.488] | 0.954 | 0.998 | 103.58 | 32.63 |
| | 0.90 | [0.132,0.340] | [0.024,0.464] | 0.904 | 0.996 | 106.20 | 26.52 |
| | 0.85 | [0.142,0.330] | [0.040,0.448] | 0.848 | 0.996 | 110.10 | 32.01 |
| $\Delta_1^{[1]} = -1$ | 0.95 | [-1.321,-0.716] | [-1.712,-0.296] | 0.946 | 1 | 88.21 | 22.11 |
| | 0.90 | [-1.284,-0.755] | [-1.647,-0.368] | 0.895 | 0.999 | 94.38 | 22.65 |
| | 0.85 | [-1.259,-0.778] | [-1.611,-0.416] | 0.849 | 0.997 | 92.77 | 27.52 |
| $\Delta_1^{[2]} = -1$ | 0.95 | [-1.179,-0.791] | [-1.443,0.500] | 0.950 | 1 | 96.97 | 27.31 |
| | 0.90 | [-1.153,-0.814] | [-1.398,-0.544] | 0.891 | 0.999 | 98.69 | 25.13 |
| | 0.85 | [-1.136,-0.832] | [-1.370,-0.575] | 0.853 | 0.999 | 102.16 | 25.11 |

Table notes: (1) $\Theta_I$ is a singleton in this DGP. (2) $B = 1001$ bootstrap draws. (3) "Average time" is computation time in seconds averaged over MC replications. (4) $CI_n$ is calibrated projection and $CI_n^{proj}$ is uncalibrated projection.

Table 4: Results for Set 2-DGP2, $Corr(u_1, u_2) = 0$, $n = 4000$, $MCs = 1000$, $\rho = 6.02$, $\kappa_n = \sqrt{\ln n}$.

| | $1-\alpha$ | Median CI | | $CI_n$ Coverage | | $CI_n^{proj}$ Coverage | | Average Time | |
|---|---|---|---|---|---|---|---|---|---|
| | | $CI_n$ | $CI_n^{proj}$ | Lower | Upper | Lower | Upper | $CI_n$ | $CI_n^{proj}$ |
| $\zeta_1^{[1]} = 0.50$ | 0.95 | [0.249,0.790] | [-0.007,1.004] | 0.954 | 0.971 | 0.999 | 1 | 85.76 | 50.10 |
| | 0.90 | [0.271,0.765] | [0.038,0.969] | 0.918 | 0.941 | 0.998 | 1 | 91.47 | 50.51 |
| | 0.85 | [0.287,0.750] | [0.067,0.948] | 0.883 | 0.919 | 0.999 | 1 | 91.39 | 61.10 |
| $\zeta_1^{[2]} = 0.25$ | 0.95 | [0.112,0.376] | [0.009,0.523] | 0.969 | 0.963 | 0.998 | 1 | 94.09 | 36.46 |
| | 0.90 | [0.128,0.359] | [0.025,0.498] | 0.938 | 0.927 | 0.997 | 0.999 | 93.26 | 52.80 |
| | 0.85 | [0.138,0.348] | [0.038,0.489] | 0.909 | 0.891 | 0.998 | 0.996 | 95.68 | 61.25 |
| $\Delta_1^{[1]} = -1$ | 0.95 | [-1.467,-0.497] | [-1.869,-0.003] | 0.960 | 0.967 | 0.999 | 0.999 | 82.54 | 27.25 |
| | 0.90 | [-1.432,-0.544] | [-1.806,-0.091] | 0.932 | 0.939 | 1 | 0.999 | 89.97 | 28.63 |
| | 0.85 | [-1.408,-0.571] | [-1.766,-0.146] | 0.901 | 0.902 | 1 | 0.999 | 91.72 | 28.38 |
| $\Delta_1^{[2]} = -0.75$ | 0.95 | [-0.979,-0.514] | [-1.276,-0.237] | 0.973 | 0.969 | 1 | 1 | 97.75 | 32.09 |
| | 0.90 | [-0.953,-0.539] | [-1.226,-0.282] | 0.941 | 0.940 | 1 | 1 | 95.86 | 27.34 |
| | 0.85 | [-0.936,-0.556] | [-1.194,-0.312] | 0.916 | 0.917 | 1 | 0.999 | 104.52 | 31.15 |

Notes: (1) Projections of $\Theta_I$ are: $\zeta_1^{[1]} \in [0.405, 0.589]$; $\zeta_1^{[2]} \in [0.236, 0.266]$; $\Delta_1^{[1]} \in [-1.158, -0.832]$; $\Delta_1^{[2]} \in [-0.790, -0.716]$. (2) "Upper" coverage refers to coverage of $\max\{p'\theta : \theta \in \Theta_I(P)\}$, and similarly for "Lower". (3) "Average time" is computation time in seconds averaged over MC replications. (4) $B = 1001$ bootstrap draws. (5) $CI_n$ is calibrated projection and $CI_n^{proj}$ is uncalibrated projection.

Table 5: Results for Set 2-DGP3, $Corr(u_1, u_2) = 0.5$, $n = 4000$, $MCs = 1000$, $\rho = 6.02$, $\kappa_n = \sqrt{\ln n}$.

| | $1-\alpha$ | Median CI | | $CI_n$ Coverage | | $CI_n^{proj}$ Coverage | | Average Time | |
|---|---|---|---|---|---|---|---|---|---|
| | | $CI_n$ | $CI_n^{proj}$ | Lower | Upper | Lower | Upper | $CI_n$ | $CI_n^{proj}$ |
| $\zeta_1^{[1]} = 0.50$ | 0.95 | [0.196,0.895] | [-0.043,1.053] | 0.978 | 0.978 | 0.996 | 0.995 | 561.66 | 163.42 |
| | 0.90 | [0.224,0.864] | [-0.009,1.009] | 0.958 | 0.966 | 0.993 | 0.984 | 583.80 | 163.42 |
| | 0.85 | [0.244,0.844] | [0.015,1.000] | 0.945 | 0.945 | 0.989 | 0.972 | 562.05 | 99.90 |
| $\zeta_1^{[2]} = 0.25$ | 0.95 | [0.099,0.436] | [0.001,0.586] | 0.974 | 0.969 | 0.997 | 0.996 | 626.00 | 245.39 |
| | 0.90 | [0.115,0.417] | [0.016,0.583] | 0.951 | 0.950 | 0.997 | 0.997 | 597.29 | 206.35 |
| | 0.85 | [0.126,0.404] | [0.031,0.564] | 0.939 | 0.941 | 0.993 | 0.994 | 681.24 | 234.50 |
| $\Delta_1^{[1]} = -1$ | 0.95 | [-1.664,-0.372] | [-1.956,-0.000] | 0.957 | 0.962 | 0.986 | 0.993 | 578.63 | 156.00 |
| | 0.90 | [-1.609,-0.441] | [-1.929,-0.000] | 0.939 | 0.930 | 0.986 | 0.996 | 594.27 | 145.85 |
| | 0.85 | [-1.568,-0.490] | [-1.912,-0.000] | 0.909 | 0.916 | 0.986 | 0.994 | 638.16 | 132.73 |
| $\Delta_1^{[2]} = -0.75$ | 0.95 | [-1.065,-0.504] | [-1.312,-0.1938] | 0.956 | 0.955 | 0.994 | 0.995 | 559.10 | 214.71 |
| | 0.90 | [-1.037,-0.525] | [-1.286,-0.241] | 0.940 | 0.947 | 0.994 | 0.997 | 553.53 | 128.71 |
| | 0.85 | [-1.021,-0.542] | [-1.276,-0.266] | 0.918 | 0.928 | 0.989 | 0.998 | 645.54 | 129.67 |
| $r = 0.5$ | 0.95 | [0.000,0.830] | [0.000,0.925] | 0.968 | 0.968 | 0.995 | 0.995 | 269.98 | 42.66 |
| | 0.90 | [0.000,0.802] | [0.000,0.925] | 0.935 | 0.935 | 0.994 | 0.995 | 308.58 | 47.55 |
| | 0.85 | [0.042,0.784] | [0.000,0.925] | 0.897 | 0.897 | 0.995 | 0.995 | 334.43 | 49.54 |

Notes: (1) Projections of $\Theta_I$ are: $\zeta_1^{[1]} \in [0.465, 0.533]$; $\zeta_1^{[2]} \in [0.240, 0.261]$; $\Delta_1^{[1]} \in [-1.069, -0.927]$; $\Delta_1^{[2]} \in [-0.782, -0.720]$; $r \in [0.4998, 0.5000]$. (2) "Upper" coverage refers to coverage of $\max\{p'\theta : \theta \in \Theta_I(P)\}$, and similarly for "Lower". (3) "Average time" is computation time in seconds averaged over MC replications. (4) $B = 1001$ bootstrap draws. (5) $CI_n$ is calibrated projection and $CI_n^{proj}$ is uncalibrated projection.

Table 6: Results for Set 2-DGP2, $Corr(u_1, u_2) = 0$, $n = 4000$, $MCs = 1000$, varying $\rho$, $\kappa_n = \sqrt{\ln n}$.

| | | Median $CI_n$ | | $CI_n$ Coverage | | | | Average Time | |
|---|---|---|---|---|---|---|---|---|---|
| | $1-\alpha$ | $\rho = 5.87$ | $\rho = 10$ | $\rho = 5.87$ | | $\rho = 10$ | | $\rho = 5.87$ | $\rho = 10$ |
| | | | | Lower | Upper | Lower | Upper | | |
| $\zeta_1^{[1]} = 0.50$ | 0.95 | [0.248,0.790] | [0.254,0.776] | 0.959 | 0.971 | 0.947 | 0.962 | 116.19 | 104.14 |
| | 0.90 | [0.271,0.766] | [0.275,0.754] | 0.921 | 0.939 | 0.908 | 0.925 | 121.24 | 115.65 |
| | 0.85 | [0.286,0.749] | [0.289,0.738] | 0.888 | 0.916 | 0.868 | 0.895 | 115.41 | 112.38 |
| $\Delta_1^{[1]} = -1$ | 0.95 | [-1.471,-0.498] | [-1.454,-0.512] | 0.964 | 0.965 | 0.955 | 0.959 | 104.34 | 108.77 |
| | 0.90 | [-1.434,-0.543] | [-1.418,-0.555] | 0.933 | 0.940 | 0.927 | 0.924 | 113.63 | 114.74 |
| | 0.85 | [-1.410,-0.571] | [-1.394,-0.583] | 0.904 | 0.905 | 0.887 | 0.895 | 114.23 | 119.55 |

Notes: Same DGP, number of bootstrap draws and conventions as in Table 4. Results are for calibrated projection $CI_n$.

Table 7: Results for Set 2-DGP2, $Corr(u_1, u_2) = 0$, $n = 4000$, $MCs = 1000$, $\rho = 6.02$, varying $\kappa_n$.

| | | Median $CI_n$ | | $CI_n$ Coverage | | | | Average Time | |
|---|---|---|---|---|---|---|---|---|---|
| | $1-\alpha$ | $\kappa_n = n^{1/7}$ | $\kappa_n = \sqrt{\ln \ln n}$ | $\kappa_n = n^{1/7}$ | | $\kappa_n = \sqrt{\ln \ln n}$ | | $\kappa_n = n^{1/7}$ | $\kappa_n = \sqrt{\ln \ln n}$ |
| | | | | Lower | Upper | Lower | Upper | | |
| $\zeta_1^{[1]} = 0.50$ | 0.95 | [0.249,0.790] | [0.250,0.787] | 0.955 | 0.972 | 0.955 | 0.970 | 85.11 | 89.65 |
| | 0.90 | [0.270,0.765] | [0.274,0.763] | 0.922 | 0.943 | 0.914 | 0.936 | 89.12 | 94.49 |
| | 0.85 | [0.286,0.748] | [0.287,0.746] | 0.891 | 0.916 | 0.870 | 0.901 | 89.82 | 92.15 |
| $\Delta_1^{[1]} = -1$ | 0.95 | [-1.469,-0.497] | [-1.464,-0.501] | 0.966 | 0.968 | 0.956 | 0.959 | 80.33 | 81.70 |
| | 0.90 | [-1.432,-0.542] | [-1.426,-0.548] | 0.935 | 0.938 | 0.926 | 0.923 | 85.12 | 88.07 |
| | 0.85 | [-1.408,-0.568] | [-1.402,-0.577] | 0.909 | 0.908 | 0.889 | 0.892 | 86.95 | 89.34 |

Notes: Same DGP, number of bootstrap draws and conventions as in Table 4. Results are for calibrated projection $CI_n$.

## References


Anderson, T. W., and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

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. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Aradillas-Lopez, A., and E. Tamer (2008): "The Identification Power of Equilibrium in Simple Games," *Journal of Business & Economic Statistics*, 26(3), 261–283.

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.

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bugni, F. A., I. A. Canay, and X. Shi (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185(1), 259–282.

——— (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8(1), 1–38.

Bull, A. D. (2011): "Convergence rates of efficient global optimization algorithms," *Journal of Machine Learning Research*, 12(Oct), 2879–2904.

Chen, X., T. M. Christensen, and E. Tamer (2017): "MCMC Confidence Sets for Identified Sets," Discussion paper, https://arxiv.org/abs/1605.00499.

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.

Dickstein, M., and E. Morales (2016): "What do Exporters Know?," Discussion paper, mimeo.

Freyberger, J., and B. Reeves (2017a): "Inference Under Shape Restrictions," mimeo.

——— (2017b): "Supplementary Appendix: Inference Under Shape Restrictions," mimeo.

Gafarov, B., M. Meier, and J. L. Montiel-Olea (2016): "Projection Inference for Set-Identified SVARs," mimeo.

Gordon, R. (1941): "Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument," *The Annals of Mathematical Statistics*.

Grieco, P. L. E. (2014): "Discrete games with flexible information structures: an application to local grocery markets," *The RAND Journal of Economics*, 45(2), 303–340.

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. (2016): "A dual approach to inference for partially identified econometric models," *Journal of Econometrics*, 192(1), 269 – 290.

Kaido, H., F. Molinari, and J. Stoye (2017): "Constraint qualifications in projection inference," Work in progress.

Kaido, H., F. Molinari, J. Stoye, and M. Thirkettle (2017): "Calibrated Projection in MATLAB," Discussion paper, available at https://molinari.economics.cornell.edu/docs/KMST_Manual.pdf.

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.

Magnac, T., and E. Maurin (2008): "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data," *Review of Economic Studies*, 75, 835–864.

Mattingley, J., and S. Boyd (2012): "CVXGEN: a code generator for embedded convex optimization," *Optimization and Engineering*, 13(1), 1–27.

Mohapatra, D., and C. Chatterjee (2015): "Price Control and Access to Drugs: The Case of India's Malaria Market," Working Paper. Cornell University.

Narcowich, F., J. Ward, and H. Wendland (2003): "Refined error estimates for radial basis function interpolation," *Constructive approximation*.

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.

Rockafellar, R. T., and R. J.-B. Wets (2005): *Variational Analysis, Second Edition*. Springer-Verlag, Berlin.

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.

Santner, T. J., B. J. Williams, and W. I. Notz (2013): *The design and analysis of computer experiments*. Springer Science & Business Media.

Schonlau, M., W. J. Welch, and D. R. Jones (1998): "Global versus local search in constrained optimization of computer models," *New Developments and Applications in Experimental Design*, Lecture Notes-Monograph Series, Vol. 34, 11–25.

Stoye, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299–1315.

Wan, Y. (2013): "An Integration-based Approach to Moment Inequality Models," Working Paper.

# Online Appendix: Confidence Intervals for Projections of Partially Identified Parameters

## Contents



## Structure of the Appendix

Section B states and proofs Theorem B.1, which establishes convergence-related results for our E-A-M algorithm. It also provides provides background material for the E-A-M algorithm, and details on the root-finding algorithm

that we use to compute $\hat{c}_n(\theta)$. Section C verifies some of our main assumptions for moment (in)equality models that have received much attention in the literature. Section D summarizes the notation we use and the structure of the proof of Theorem 4.1,[43] and provides a proof of Theorems 4.1 (both under our main assumptions and under a high level assumption replacing Assumption 4.3 and dropping the $\rho$-box constraints), 4.2, 4.3 and 4.4. Section E contains the statements and proofs of the lemmas used to establish Theorems 4.1 and B.1, as well as a rigorous derivation of the almost sure representation result for the bootstrap empirical process that we use in the proof of Theorem 4.1. Section F provides further results comparing our calibrated projection method and the profiling method proposed by Bugni, Canay, and Shi (2017, BCS-profiling henceforth), and gives an example of methods' failure (including calibrated projection, BCS-profiling and the method in Pakes, Porter, Ho, and Ishii (2011)) when some key assumptions are violated. Section G provides a formal comparison of our calibrated projection method and projection of the confidence set of Andrews and Soares (2010, AS henceforth).

Throughout the Appendix we use the convention $\infty \cdot 0 = 0$.

# Appendix B Additional Convergence Results and Background Materials for the E-A-M algorithm and for Computation of $\hat{c}_n(\theta)$

## B.1 Theorem B.1: An Approximating Critical Level Sequence for the E-A-M Algorithm

### B.1.1 Assumption B.1: A Low Level Condition Yielding a Stochastic Lipschitz-Type Property for $\hat{c}_n$

In order to establish convergence of our E-A-M algorithm, we need $\hat{c}_n$ to uniformly stochastically exhibit a Lipschitz-type property so that its mollified counterpart (see equation (B.1)) is sufficiently smooth and yields valid inference. Below we provide a low level condition under which we are able to establish the Lipschitz-type property. In Appendix C.1 we verify the condition for the canonical examples in the moment (in)equality literature.

ASSUMPTION B.1: *The model $\mathcal{P}$ for $P$ satisfies:*

*(i)* $|\sigma_{P,j}(\theta)^{-1}m_j(x,\theta) - \sigma_{P,j}(\theta')^{-1}m_j(x,\theta')| \leqslant \bar{M}(x)\|\theta - \theta'\|$ *with* $E_P[\bar{M}(X)^2] < M$ *for all* $\theta, \theta' \in \Theta$, $x \in \mathcal{X}$, $j = 1, \cdots, J$, *and there exists a function* $F$ *such that* $|\sigma_{P,j}(\theta)^{-1}m_j(\cdot,\theta)| \leqslant F(\cdot)$ *for all* $\theta \in \Theta$ *and* $E_P[|F(X)\bar{M}(X)|^2] < M$.

*(ii)* $\varphi_j$ *is Lipschitz continuous in* $x \in \mathbb{R}$ *for all* $j = 1, \ldots, J$.

### B.1.2 Statement and Proof of Theorem B.1

For all $\tau > 0$ let $\hat{c}_{n,\tau}(\theta)$ be a mollified version of $\hat{c}_n(\theta)$, i.e.:

$$\hat{c}_{n,\tau}(\theta) = \int_{\mathbb{R}^d} \hat{c}_n(\theta - \nu)\phi_\tau(\nu)d\nu = \int_{\mathbb{R}^d} \hat{c}_n(\theta)\phi_\tau(\theta - \nu)d\nu, \tag{B.1}$$

[43] Section D.1 provides in Table D.0 a summary of the notation used throughout, and in Figure D.1 and Table D.1 a flow diagram and heuristic explanation of how each lemma contributes to the proof of Theorem 4.1.

where the family of functions $\phi_\tau$ is a mollifier as defined in Rockafellar and Wets (2005, Example 7.19). Choose it to be a family of bounded, measurable, smooth functions such that $\phi_\tau(z) \geqslant 0\ \forall z \in \mathbb{R}^d$, $\int_{\mathbb{R}^d} \phi_\tau(z)dz = 1$ and with $\mathbb{B}_\tau = \{z : \phi_\tau(z) > 0\} = \{z : \|z\| \leqslant \tau\}$.

THEOREM B.1: *Suppose Assumptions 4.1, 4.2, 4.4, 4.5 and B.1 hold. Let $\tau_n$ be a positive sequence such that $\tau_n = n^{-\zeta}$ with $\zeta > 1/2$. Let $\{\beta_n\}$ be a positive sequence such that $\beta_n = o(1)$ and $\|\hat{D}_n - D_P\|_\infty = O_{\mathcal{P}}(\beta_n)$. Let $\varepsilon_n = \kappa_n^{-1}\sqrt{n}\tau_n \vee \beta_n$. Then,*

1. $$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} P\left( \sup_{\|\theta-\theta'\|\leqslant\tau_n} |\hat{c}_n(\theta) - \hat{c}_n(\theta')| > C\varepsilon_n \right) = 0; \tag{B.2}$$

2. *Let $\hat{c}_{n,\tau_n}$ be defined as in (B.1) with $\tau_n$ replacing $\tau$. Then there exists $C > 0$ such that*
$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\Big( \|\hat{c}_n - \hat{c}_{n,\tau_n}\|_\infty \leqslant C\varepsilon_n \Big) = 1; \tag{B.3}$$

3. *There exists $R > 0$ such that $\|\hat{c}_{n,\tau_n}\|_{\mathcal{H}_\beta} \leqslant R$.*

4. *Let Assumption 4.3 also hold. Let $\{P_n, \theta_n\}$ be a sequence such that $P_n \in \mathcal{P}$ and $\theta_n \in \Theta_I(P_n)$ for all $n$ and $\kappa_n^{-1}\sqrt{n}\gamma_{1,P_n,j}(\theta_n) \to \pi_{1j} \in \mathbb{R}_{[-\infty]}$, $j = 1, \ldots, J$, $\Omega_{P_n} \xrightarrow{u} \Omega$, and $D_{P_n}(\theta_n) \to D$. Let*
$$\hat{c}_{n,\rho,\tau}(\theta) \equiv \inf_{\lambda\in B^d_{n,\rho}} \hat{c}_{n,\tau}(\theta + \frac{\lambda\rho}{\sqrt{n}}). \tag{B.4}$$
*For $c \geqslant 0$, let $U_n(\theta_n, c)$ be defined as in (D.25). Then,*
$$\liminf_{n\to\infty} P_n\left(U_n(\theta_n, \hat{c}_{n,\rho,\tau_n}) \neq \varnothing\right) \geqslant 1 - \alpha. \tag{B.5}$$

*Proof.* We establish each part of the theorem separately.

**Part 1.** Throughout, let $C > 0$ denote a positive constant, which may be different in different appearances. Define the event

$$E_n \equiv \big\{x^\infty \in \mathcal{X}^\infty : \|\hat{D}_n - D_P\|_\infty \leqslant C\beta_n,\ \sup_{\|\theta-\theta'\|\leqslant\tau_n} \|\mathbb{G}_n(\theta) - \mathbb{G}_n(\theta'))\| \leqslant (\ln n)^2\tau_n,$$
$$\sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant C/\sqrt{n},\ \max_{j=1,\cdots,J} \sup_{\|\theta-\theta'\|<\tau_n} |\eta_{n,j}(\theta) - \eta_{n,j}(\theta')| \leqslant C\tau_n\big\}. \tag{B.6}$$

Note that $(\ln n)^2\tau_n/(-\tau_n \ln \tau_n) = (\ln n)^2/\zeta \ln n = \ln n/\zeta$, and hence tends to $\infty$. By Assumption B.1-(i) and arguing as in the proof of Theorem 2 in Andrews (1994), condition (E.216) in Lemma E.11 is satisfied with $v = d$. Also, by Lemma E.13, (E.217) in Lemma E.11 holds with $\gamma = 1$. This therefore ensures the conditions of Lemma E.11.

Similarly, by Assumption B.1-(i) $m_j^2(x,\theta)/\sigma_{P,j}^2(\theta)$ satisfies

$$\Big|\frac{m_j^2(x,\theta)}{\sigma_{P,j}^2(\theta)} - \frac{m_j^2(x,\theta)}{\sigma_{P,j}^2(\theta)}\Big| \leqslant \Big|\frac{m_j(x,\theta)}{\sigma_{P,j}(\theta)} + \frac{m_j(x,\theta')}{\sigma_{P,j}(\theta')}\Big|\Big|\frac{m_j(x,\theta)}{\sigma_{P,j}(\theta)} - \frac{m_j(x,\theta')}{\sigma_{P,j}(\theta')}\Big| \tag{B.7}$$
$$\leqslant 2F(x)\bar{M}(x)\|\theta - \theta'\|. \tag{B.8}$$

Let $\bar{F}(x) \equiv 2F(x)\bar{M}(x)$. By Theorem 2.7.11 in van der Vaart and Wellner (2000),

$$N_{[]}(\epsilon\|\bar{F}\|_{L_P^2}, \mathcal{M}_P^2, \|\cdot\|_{L_P^2}) \leqslant N(\epsilon, \Theta, \|\cdot\|) \leqslant (\mathrm{diam}(\Theta)/\epsilon)^d, \tag{B.9}$$

where $N(\epsilon, \Theta, \|\cdot\|)$ is the covering number of $\Theta$. This ensures

$$\int_0^\infty \sup_{P\in\mathcal{P}} \sqrt{\ln N_{[]}(\epsilon \|\bar{F}\|_{L_P^2}, \mathcal{M}_P^2, \|\cdot\|_{L_P^2})} d\epsilon < \infty. \tag{B.10}$$

Further, for any $C > 0$

$$\begin{aligned} E_P[\bar{F}^2(X) 1\{\bar{F}(X) > C\}] &\leqslant E_P[\bar{F}^2(X)] P(\bar{F}(X) > C) \\ &\leqslant 4E_P[|F(X)M(X)|^2] \frac{\|\bar{F}\|_{L_P^1}}{C} \leqslant \frac{4M^2}{C}, \end{aligned} \tag{B.11}$$

which implies $\lim_{C\to\infty} \sup_{P\in\mathcal{P}} E_P[\bar{F}^2(X) 1\{\bar{F}(X) > C\}] = 0$. By Theorems 2.8.4 and 2.8.2 in van der Vaart and Wellner (2000), this implies that $\mathcal{S}_P$ is Donsker and pre-Gaussian uniformly in $P \in \mathcal{P}$. This therefore ensures the conditions of Lemma E.12 (i). Note also that Assumption B.1-(i) ensures the conditions of Lemma E.12 (ii). Therefore, by Lemmas E.11-E.12 and Assumption 4.4, for any $\eta > 0$, there exists $C > 0$ such that $\inf_{P\in\mathcal{P}} P(E_n) \geqslant 1 - \eta$ for all $n$ sufficiently large.

Let $\theta, \theta' \in \Theta$. For each $j$, we have

$$\begin{aligned} &\left|\mathbb{G}_{n,j}^b(\theta) + \rho \hat{D}_{n,j}(\theta)\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta)) - \mathbb{G}_{n,j}^b(\theta') - \rho \hat{D}_{n,j}(\theta')\lambda - \varphi_j(\hat{\xi}_{n,j}(\theta'))\right| \\ &\qquad \leqslant |\mathbb{G}_{n,j}^b(\theta) - \mathbb{G}_{n,j}^b(\theta')| + \rho \|\hat{D}_{n,j}(\theta) - \hat{D}_{n,j}(\theta')\| \sup_{\lambda\in B^d} \|\lambda\| + |\varphi_j(\hat{\xi}_{n,j}(\theta)) - \varphi_j(\hat{\xi}_{n,j}(\theta'))|. \end{aligned} \tag{B.12}$$

Assume that the sample path $\{X_i\}_{i=1}^\infty$ is such that the event $E_n$ holds. Conditional on $\{X_i\}_{i=1}^\infty$ and using $\mathbb{G}_{n,j}^b(\theta) - \mathfrak{G}_{n,j}^b(\theta) = \mathfrak{G}_{n,j}^b(\theta)\eta_{n,j}(\theta)$,

$$\begin{aligned} |\mathbb{G}_{n,j}^b(\theta) - \mathbb{G}_{n,j}^b(\theta')| &\leqslant |\mathfrak{G}_{n,j}^b(\theta) - \mathfrak{G}_{n,j}^b(\theta')| + 2\sup_{\theta\in\Theta} |\mathfrak{G}_{n,j}^b(\theta)| \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \\ &\leqslant |\mathfrak{G}_{n,j}^b(\theta) - \mathfrak{G}_{n,j}^b(\theta')| + 2\sup_{\theta\in\Theta} |\mathfrak{G}_{n,j}^b(\theta)| \frac{C}{\sqrt{n}}. \end{aligned} \tag{B.13}$$

Define the event $F_n \in \mathcal{C}$ for the bootstrap weights by

$$F_n \equiv \left\{m_n \in Q : \sup_{\|\theta-\theta'\|\leqslant\tau_n} \|\mathfrak{G}_n^b(\theta) - \mathfrak{G}_n^b(\theta')\| \leqslant (\ln n)^2 \tau_n, \ \sup_{\theta\in\Theta} \|\mathfrak{G}_n^b(\theta)\| \leqslant C\right\}. \tag{B.14}$$

By Lemma E.11 (ii) and the asymptotic tightness of $\mathfrak{G}_n^b$, for any $\eta > 0$, there exists a $C$ such that $P_n^*(F_n) \geqslant 1 - \eta$ for all $n$ sufficiently large. Suppose that the multinomial bootstrap weight $M_n$ is such that $F_n$ holds. Then, the right hand side of (B.13) is bounded by $(\ln n)^2\tau_n + C/\sqrt{n}$ for some $C > 0$.

Next, by the triangle inequality and Assumption 4.4,

$$\begin{aligned} \|\hat{D}_{n,j}(\theta) - \hat{D}_{n,j}(\theta')\| &\leqslant \|\hat{D}_{n,j}(\theta) - D_{P,j}(\theta)\| + \|D_{P,j}(\theta) - D_{P,j}(\theta')\| + \|\hat{D}_{n,j}(\theta') - D_{P,j}(\theta')\| \\ &\leqslant C\beta_n + C\tau_n. \end{aligned} \tag{B.15}$$

Finally, note that by the Lipschitzness of $\varphi_j$, $|\varphi_j(\hat{\xi}_{n,j}(\theta)) - \varphi_j(\hat{\xi}_{n,j}(\theta'))| \leqslant C|\hat{\xi}_{n,j}(\theta) - \hat{\xi}_{n,j}(\theta')|$ and

$$\begin{aligned} &\hat{\xi}_{n,j}(\theta) - \hat{\xi}_{n,j}(\theta') \\ &\quad = \kappa_n^{-1}\Big[\sqrt{n}\Big(\frac{\bar{m}_{n,j}(\theta)}{\sigma_{P,j}(\theta)}(1 + \eta_{n,j}(\theta)) - \frac{E_P[m_j(X,\theta)]}{\sigma_{P,j}(\theta)}\Big) - \sqrt{n}\Big(\frac{\bar{m}_{n,j}(\theta')}{\sigma_{P,j}(\theta')}(1 + \eta_{n,j}(\theta')) - \frac{E_P[m_j(X,\theta')]}{\sigma_{P,j}(\theta')}\Big)\Big] \\ &\qquad + \kappa_n^{-1}\sqrt{n}\Big(\frac{E_P[m_j(X,\theta)]}{\sigma_{P,j}(\theta)} - \frac{E_P[m_j(X,\theta')]}{\sigma_{P,j}(\theta')}\Big). \end{aligned} \tag{B.16}$$

Hence,

$$|\hat{\xi}_{n,j}(\theta) - \hat{\xi}_{n,j}(\theta')| \leqslant \kappa_n^{-1}|\mathbb{G}_{n,j}(\theta) - \mathbb{G}_{n,j}(\theta')| + \kappa_n^{-1}\sqrt{n}\Big|\frac{\bar{m}_{n,j}(\theta)}{\sigma_{P,j}(\theta)}\eta_{n,j}(\theta) - \frac{\bar{m}_{n,j}(\theta')}{\sigma_{P,j}(\theta')}\eta_{n,j}(\theta')\Big| + \kappa_n^{-1}\sqrt{n}D_{P,j}(\bar{\theta})\|\theta - \theta'\|. \quad \text{(B.17)}$$

By Lemma E.11, the right hand side of (B.17) can be further bounded by

$$\begin{aligned}&\kappa_n^{-1}(\ln n)^2\tau_n + \kappa_n^{-1}\sqrt{n}\Big|\frac{\bar{m}_{n,j}(\theta)}{\sigma_{P,j}(\theta)} - \frac{\bar{m}_{n,j}(\theta')}{\sigma_{P,j}(\theta')}\Big||\eta_{n,j}(\theta)| \\ &\qquad + \kappa_n^{-1}\sqrt{n}\Big|\frac{\bar{m}_{n,j}(\theta')}{\sigma_{P,j}(\theta')}\Big||\eta_{n,j}(\theta) - \eta_{n,j}(\theta')| + C\kappa_n^{-1}\sqrt{n}\tau_n \\ &\qquad \leqslant \kappa_n^{-1}(\ln n)^2\tau_n + \kappa_n^{-1}\sqrt{n}\tau_n\frac{C}{\sqrt{n}} + C\kappa_n^{-1}\sqrt{n}\tau_n + C\kappa_n^{-1}\sqrt{n}\tau_n, \quad \text{(B.18)}\end{aligned}$$

where the last inequality follows from Condition (i) and Lemma E.12 (ii).

Combining (B.12), (B.13), (B.15), and (B.16)-(B.18), we obtain

$$\Big|\mathbb{G}_{n,j}^b(\theta) + \hat{D}_{n,j}(\theta)\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta)) - \mathbb{G}_{n,j}^b(\theta') - \hat{D}_{n,j}(\theta')\lambda - \varphi_j(\hat{\xi}_{n,j}(\theta'))\Big| \leqslant C\varepsilon_n. \quad \text{(B.19)}$$

In particular, if $\mathbf{1}\left(\Lambda_n^b(\theta,\rho,\hat{c}_n(\theta)) \cap \{p'\lambda = 0\} \neq \varnothing\right) = 1$, it also holds that $\mathbf{1}\big(\Lambda_n^b(\theta',\rho,\hat{c}_n(\theta) + C\varepsilon_n) \cap \{p'\lambda = 0\} \neq \varnothing\big) = 1$ because

$$\mathbb{G}_{n,j}^b(\theta') + \hat{D}_{n,j}(\theta')\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta')) \leqslant \mathbb{G}_{n,j}^b(\theta) + \hat{D}_{n,j}(\theta)\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta)) + C\varepsilon_n \leqslant \hat{c}_n(\theta) + C\varepsilon_n,$$

Recalling that $P_n^*(F_n) \geqslant 1 - \eta$ for all $n$ sufficiently large, we then have

$$\begin{aligned}&P_n^*\left(\left\{\Lambda_n^b(\theta',\rho,\hat{c}_n(\theta) + C\varepsilon_n) \cap \{p'\lambda = 0\} \neq \varnothing\right\}\right) \\ &\qquad \geqslant P_n^*\left(\left\{\Lambda_n^b(\theta',\rho,\hat{c}_n(\theta) + C\varepsilon_n) \cap \{p'\lambda = 0\} \neq \varnothing\right\} \cap F_n\right) \\ &\qquad \geqslant P_n^*\left(\left\{\Lambda_n^b(\theta,\rho,\hat{c}_n(\theta)) \cap \{p'\lambda = 0\} \neq \varnothing\right\} \cap F_n\right) \geqslant 1 - \alpha - \eta. \quad \text{(B.20)}\end{aligned}$$

Since $\eta$ is arbitrary, we have

$$\hat{c}_n(\theta') \leqslant \hat{c}_n(\theta) + C\varepsilon_n.$$

Reversing the roles of $\theta$ and $\theta'$ and noting that $\sup_{P\in\mathcal{P}} P(E_n) \to 0$ yields the first claim of the lemma.

**Part 2.** To obtain the result in equation (B.3), we use that for any $\theta, \theta' \in \Theta$ such that $\|\theta - \theta'\| \leqslant \tau_n$, $|\hat{c}_n(\theta) - \hat{c}_n(\theta')| \leqslant C\varepsilon_n$ with probability approaching 1 uniformly in $P \in \mathcal{P}$ by the result in Part 1. This implies

$$\begin{aligned}|\hat{c}_n(\theta) - \hat{c}_{n,\tau_n}(\theta)| &= \left|\int_{\mathbb{R}^d} \hat{c}_n(\theta - \nu)\phi_{\tau_n}(\nu)d\nu - \hat{c}_n(\theta)\right| \leqslant \int_{\mathbb{R}^d} |\hat{c}_n(\theta - \nu) - \hat{c}_n(\theta)|\,\phi_{\tau_n}(\nu)d\nu \\ &= \int_{\mathbb{B}_{\tau_n}} |\hat{c}_n(\theta - \nu) - \hat{c}_n(\theta)|\,\phi_{\tau_n}(\nu)d\nu \leqslant C\varepsilon_n \int_{\mathbb{B}_{\tau_n}} \phi_{\tau_n}(\nu)d\nu \leqslant C\varepsilon_n.\end{aligned}$$

**Part 3.** By the construction of the mollified version of the critical value, we have $\hat{c}_{n,\tau_n} \in \mathcal{C}^\infty(\Theta)$ (Adams and Fournier, 2003, Theorem 2.29). Therefore it has derivatives of all order. Using the multi-index notation, for any $s > 0$ and $|\alpha| \leqslant s$, the partial derivative $\nabla^\alpha \hat{c}_{n,\tau_n}$ is bounded by some constant $M > 0$ on the compact set $\Theta$, and

hence

$$\int_{\Theta} |\nabla^{\alpha} \hat{c}_{n,\tau_n}(\theta)|^2 d\upsilon(\theta) \leqslant M\upsilon(\Theta) < \infty,$$

where $\upsilon$ denote the Lebesgue measure on $\mathbb{R}^d$. This ensures $\nabla^{\alpha} \hat{c}_{n,\tau_n} \in L^2_{\upsilon}(\Theta)$ for all $|\alpha| \leqslant s$. Hence, $\hat{c}_{n,\tau_n}$ is in the Sobolev-Hilbert space $H^s(\Theta^o)$ for any $s > 0$. Note that when a Matérn kernel with $\nu < \infty$ is used and $\hat{c}_{n,\tau_n}$ is continuous, Lemma 3 in Bull (2011) implies that the RKHS-norm $\|\cdot\|_{\mathcal{H}_{\bar{\beta}}}$ (in $\mathcal{H}_{\bar{\beta}}(\Theta)$) and the Sobolev-Hilbert norm $\|\cdot\|_{H^{\nu+d/2}}$ are equivalent. Hence, there is $R > 0$ such that $\|\hat{c}_{n,\tau_n}\|_{\mathcal{H}_{\beta}} \leqslant C\|\hat{c}_{n,\tau_n}\|_{H^{\nu+d/2}} \leqslant R$.

**Part 4.** By Part 2 and the definition of $\hat{c}_{n,\rho,\tau}$ in (B.4), it follows that

$$\begin{aligned} \hat{c}_{n,\rho,\tau_n}(\theta_n) &\geqslant \hat{c}_{n,\rho}(\theta_n) - e_n \\ &\geqslant c^I_{n,\rho}(\theta_n) - e_n, \end{aligned} \tag{B.21}$$

for some $e_n = O_{\mathcal{P}}(\varepsilon_n)$, where the second inequality follows from the construction of $c^I_{n,\rho}$ in the proof of Lemma E.1. Note that Lemma E.3 and the fact that $\varepsilon_n = o_{\mathcal{P}}(1)$ by Part 1 imply $c^I_{n,\rho}(\theta_n) - e_n \xrightarrow{P_n} c^*_{\pi *}$. Replicate equation (E.22) with $\hat{c}_{n,\rho,\tau_n}$ replacing $\hat{c}_{n,\rho}$, and mimic the argument following (E.22) in the proof of Lemma E.1. Then, the conclusion of the lemma follows. □

## B.2 The kernel of the Gaussian Process and its Associated Function Space

Following Bull (2011), we consider two commonly used classes of kernels. The first one is the Gaussian kernel, which is given by

$$K_{\beta}(\theta - \theta') = \exp\big(-\sum_{k=1}^{d} |(\theta_k - \theta'_k)/\beta_k|^2\big), \ \beta_k \in [\underline{\beta}_k, \overline{\beta}_k], \ k = 1, \cdots, d, \tag{B.22}$$

where $0 < \underline{\beta}_k < \overline{\beta}_k < \infty$ for all $k$. The second one is the class of Matérn kernels defined by

$$K_{\beta}(\theta - \theta') = \frac{2^{1-\nu}}{D(\nu)} \Big(\sqrt{2}\nu \sum_{k=1}^{d} |(\theta_k - \theta'_k)/\beta_k|^2\Big)^{\nu} k_{\nu}\Big(\sqrt{2}\nu \sum_{k=1}^{d} |(\theta_k - \theta'_k)/\beta_k|^2\Big), \ \nu \in (0, \infty), \ \nu \notin \mathbb{N},$$

where $D$ is the gamma function, and $k_{\nu}$ is the modified Bessel function of the second kind.[44] The index $\nu$ controls the smoothness of $K_{\beta}$. In particular, the Fourier transform $\hat{K}_{\beta}(\zeta)$ of the Matérn kernel is bounded from above and below by the order of $\|\zeta\|^{-2\nu-d}$ as $\|\zeta\| \to \infty$, i.e. $\hat{K}_{\beta}(\zeta) = \Theta(\|\zeta\|^{-2\nu-d})$. Similarly, the Fourier transform of the Gaussian kernel satisfies $\hat{K}_{\beta}(\zeta) = O(\|\zeta\|^{-2\nu-d})$ for any $\nu > 0$. Below, we treat the Gaussian kernel as a kernel associated with $\nu = \infty$.

Each kernel is associated with a space of functions $\mathcal{H}_{\beta}(\mathbb{R}^d)$, called the reproducing kernel Hilbert space (RKHS). Below, we give some background on this space and refer to Steinwart and Christmann (2008); van der Vaart and van Zanten (2008) for further details. For $D \subseteq \mathbb{R}^d$, let $K : D \times D \to \mathbb{R}$ be a symmetric and positive definite function. $K$ is said to be a reproducing kernel of a Hilbert space $\mathcal{H}(D)$ if $K(\cdot, \theta') \in \mathcal{H}(D)$ for all $\theta' \in D$, and

$$f(\theta) = \langle f, K(\cdot, \theta) \rangle_{\mathcal{H}(D)}$$

holds for all $f \in \mathcal{H}(D)$ and $\theta \in D$. The space $\mathcal{H}(D)$ is called a reproducing kernel Hilbert space (RKHS) over $D$ if for all $\theta \in D$, the point evaluation functional $\delta_{\theta} : \mathcal{H}(D) \to \mathbb{R}$ defined by $\delta_{\theta}(f) = f(\theta)$ is continuous. When

[44] The requirement $\nu \notin \mathbb{N}$ is not essential for the convergence result. However, it simplifies some of the arguments as one can exploit the $2\nu$-Hölder continuity of $K_{\beta}$ at the origin without a log factor (Bull, 2011, Assumption 4).

$K(\theta,\theta') = K_\beta(\theta-\theta')$ is used as the correlation functional of the Gaussian process, we denote the associated RKHS by $\mathcal{H}_\beta(D)$. Using Fourier transforms, the norm on $\mathcal{H}_\beta(D)$ can be written as

$$\|f\|_{\mathcal{H}_\beta} \equiv \inf_{g|_D=f} \int \frac{\hat{g}(\zeta)}{\hat{K}_\beta(\zeta)} d\zeta, \tag{B.23}$$

where the infimum is taken over functions $g : \mathbb{R}^d \to \mathbb{R}$ whose restrictions to $D$ coincide with $f$, and we take $0/0 = 0$.

The RKHS has a connection to other well-known classes of functions. In particular, when $D$ is a Lipschitz domain, i.e. the boundary of $D$ is locally the graph of a Lipschitz function (Tartar, 2007) and the kernel is associated with $\nu \in (0,\infty)$, $\mathcal{H}_\beta(D)$ is equivalent to the Sobolev-Hilbert space $H^{\nu+d/2}(D^o)$, which is the space of functions on $D^o$ such that

$$\|f\|^2_{H^{\nu+d/2}} \equiv \inf_{g|_{D^o}=f} \int \frac{\hat{g}(\zeta)}{(1+\|\zeta\|^2)^{\nu+d/2}} d\zeta \tag{B.24}$$

is finite, where the infimum is taken over functions $g : \mathbb{R}^d \to \mathbb{R}$ whose restrictions to $D^o$ coincide with $f$. Further, if $\nu = \infty$, $\mathcal{H}_\beta(D)$ is continuously embedded in $H^s(D^o)$ for all $s > 0$ (Bull, 2011, Lemma 3).

Theorem 3.1 requires that $c$ has a finite RKHS norm. This is to ensure that the approximation error made by the best linear predictor $c_L$ of the Gaussian process regression is controlled uniformly (Narcowich, Ward, and Wendland, 2003). When a Matérn kernel is used, it suffices to bound the norm in the Sobolev-Hilbert space $H^{\nu+d/2}$ to bound $c$'s RKHS norm. We do so in Theorem B.1 by introducing a mollified version of $\hat{c}_n$.

## B.3 A Reformulation of the M-step as a Nonlinear Program

In (3.13), $\theta^{(L+1)}$ is defined as the maximizer of the following maximization problem

$$\max_{\theta\in\Theta} (p'\theta - p'\theta^*_L)_+ \Big(1 - \Phi\Big(\frac{\bar{g}(\theta) - c_L(\theta)}{\hat{\varsigma} s_L(\theta)}\Big)\Big), \tag{B.25}$$

where $\bar{g}(\theta) = \max_{j=1,\cdots,J} g_j(\theta)$. Since $\Phi$ is strictly increasing, one may rewrite the objective function as

$$(p'\theta - p'\theta^*_L)_+ \Big(1 - \max_{j=1,\cdots,J} \Phi\Big(\frac{g_j(\theta) - c_L(\theta)}{\hat{\varsigma} s_L(\theta)}\Big)\Big) = \min_{j=1,\cdots,J} (p'\theta - p'\theta^*_L)_+ \Big(1 - \Phi\Big(\frac{g_j(\theta) - c_L(\theta)}{\hat{\varsigma} s_L(\theta)}\Big)\Big).$$

Hence, $\theta^{(L+1)}$ is a solution to the maximin problem:

$$\max_{\theta\in\Theta} \min_{j=1,\cdots,J} (p'\theta - p'\theta^*_L)_+ \Big(1 - \Phi\Big(\frac{g_j(\theta) - c_L(\theta)}{\hat{\varsigma} s_L(\theta)}\Big)\Big),$$

which can be solved, for example, by Matlab's `fminimax` function. It can also be rewritten as a nonlinear program:

$$\max_{(\theta,v)\in\Theta\times\mathbb{R}} v$$
$$s.t. (p'\theta - p'\theta^*_L)_+ \Big(1 - \Phi\Big(\frac{g_j(\theta) - c_L(\theta)}{\hat{\varsigma} s_L(\theta)}\Big)\Big) \geqslant v, j = 1,\cdots,J,$$

which can be solved by nonlinear optimization solvers, e.g. Matlab's `fmincon` or `KNITRO`. We note that the objective function and constraints together with their gradients are available in closed form.

## B.4 Root-Finding Algorithm Used to Compute $\hat{c}_n(\theta)$

This section explains in detail how $\hat{c}_n(\theta)$ in equation (3.5) is computed. For a given $\theta \in \Theta$, $P^*(\Lambda^b_n(\theta,\rho,c) \cap \{p'\lambda = 0\} \neq \varnothing)$ increases in $c$ (with $\Lambda^b_n(\theta,\rho,c)$ defined in (3.1)), and so $\hat{c}_n(\theta)$ can be quickly computed via a root-

finding algorithm, such as the Brent-Dekker Method (BDM), see Brent (1971) and Dekker (1969). To do so, define $h_\alpha(c) = \frac{1}{B}\sum_{b=1}^{B}\psi_b(c) - (1-\alpha)$ where

$$\psi_b(c(\theta)) = \mathbf{1}(\Lambda_n^b(\theta,\rho,c) \cap \{p'\lambda = 0\} \neq \varnothing).$$

Let $\bar{c}(\theta)$ be an upper bound on $\hat{c}_n(\theta)$ (for example, the asymptotic Bonferroni bound $\bar{c}(\theta) \equiv \Phi^{-1}(1-\alpha/J)$). It remains to find $\hat{c}_n(\theta)$ so that $h_\alpha(\hat{c}_n(\theta)) = 0$ if $h_\alpha(0) \leqslant 0$. It is possible that $h_\alpha(0) > 0$ in which case we output $\hat{c}_n(\theta) = 0$. Otherwise, we use BDM to find the unique root to $h_\alpha(c)$ on $[0, \bar{c}(\theta)]$ where, by construction, $h_\alpha(\bar{c}_n(\theta)) \geqslant 0$. We propose the following algorithm:

**Step 0** (Initialize)

(i) Set *Tol* equal to a chosen tolerance value;

(ii) Set $c_L = 0$ and $c_U = \bar{c}(\theta)$ (values of $c$ that bracket the root $\hat{c}_n(\theta)$);

(iii) Set $c_{-1} = c_L$ and $c_{-2} =$ [] to be undefined for now (proposed values of $c$ from 1 and 2 iterations prior). Also set $c_0 = c_L$ and $c_1 = c_U$.

(iv) Compute $\varphi_j(\hat{\xi}_{n,j}(\theta))$ $j = 1, \cdots, J$;

(v) Compute $\hat{D}_{P,n}(\theta)$;

(vi) Compute $\mathbb{G}_{n,j}^b$ for $b = 1, \cdots, B$, $j = 1, \cdots, J$;

(vii) Compute $\psi_b(c_L)$ and $\psi_b(c_U)$ for $b = 1, \cdots, B$;

(viii) Compute $h_\alpha(c_L)$ and $h_\alpha(c_U)$.

**Step 1** (Method Selection)

Use the BDM rule to select the updated value of $c$, say $c_2$. The value is updated using one of three methods: Inverse Quadratic Interpolation, Secant, or Bisection. The selection rule is based on the values of $c_i$, $i = -2, -1, 0, 1$ and the corresponding function values.

**Step 2** (Update Value Function)

Update the value of $h_\alpha(c_2)$. We can exploit previous computation and monotonicity function $\psi_b(c_2)$ to reduce computational time:

1. If $\psi_b(c_L) = \psi_b(c_U) = 0$, then $\psi_b(c_2) = 0$;
2. If $\psi_b(c_L) = \psi_b(c_U) = 1$, then $\psi_b(c_2) = 1$.

**Step 3** (Update)

(i) If $h_\alpha(c_2) \geqslant 0$, then set $c_U = c_2$. Otherwise set $c_L = c_2$.

(ii) Set $c_{-2} = c_{-1}$, $c_{-1} = c_0$, $c_0 = c_L$, and $c_1 = c_U$.

(iii) Update corresponding function values $h_\alpha(\cdot)$.

**Step 4** (Convergence)

(i) If $h_\alpha(c_U) \leqslant Tol$ or if $|c_U - c_L| \leqslant Tol$, then output $\hat{c}_n(\theta) = c_U$ and exit. Note: $h_\alpha(c_U) \geqslant 0$, so this criterion ensures that we have *at least* $1-\alpha$ coverage.

(ii) Otherwise, return to **Step 1**.

The computationally difficult part of the algorithm is computing $\psi_b(\cdot)$ in **Step 2**. This is simplified for two reasons. First, evaluation of $\psi_b(c)$ entails determining whether a constraint set comprised of $J + 2d - 2$ linear inequalities in $d - 1$ variables is feasible. This can be accomplished efficiently employing commonly used software.[45] Second, we exploit monotonicity in $\psi_b(\cdot)$, reducing the number of linear programs needed to be solved.

# Appendix C Verification of Assumptions for the Canonical Moment (In)equalities Examples

In this section we verify: (i) Assumption B.1 which is the crucial condition in Theorem B.1, and (ii) Assumption 4.3-(II), for the canonical examples in the moment (in)equalities literature:

1. **Mean with interval data (of which missing data is a special case)**. Here we assume that $W_0, W_1$ are two observable random variables such that $P(W_0 \leqslant W_1) = 1$. The identified set is defined as

$$\Theta_I(P) = \{\theta \in \Theta \subset \mathbb{R} : E_P(W_0) - \theta \leqslant 0, \theta - E_P(W_1) \leqslant 0\}. \tag{C.1}$$

2. **Linear regression with interval outcome data and discrete regressors.** Here the modeling assumption is that $W = Z'\theta + u$, where $Z = [Z_1; \ldots; Z_d]$ is a $d \times 1$ random vector with $Z_1 = 1$. We assume that $Z$ has $k$ points of support denoted $z^1, \ldots, z^k \in \mathbb{R}^d$ with $\max_{r=1,\ldots,k} \|z^r\| < M < \infty$. The researcher observes $\{W_0, W_1, Z\}$ with $P(W_0 \leqslant W \leqslant W_1 | Z = z^r) = 1, r = 1, \ldots, k$. The identified set is

$$\Theta_I(P) = \{\theta \in \Theta \subset \mathbb{R}^d : E_P(W_0|Z = z^r) - z^{r\prime}\theta \leqslant 0, z^{r\prime}\theta - E_P(W_1|Z = z^r) \leqslant 0, r = 1, \ldots, k\}. \tag{C.2}$$

3. **Best linear prediction with interval outcome data and discrete regressors.** Here the variables are defined as for the linear regression case. Beresteanu and Molinari (2008) show that the identified set for the parameters of a best linear predictor of $W$ conditional on $Z$ is given by the set $\Theta_I(P) = E_P(ZZ')^{-1}E_P(Z\mathbf{W})$, where $\mathbf{W} = [W_0, W_1]$ is a random closed set and, with some abuse of notation, $E_P(Z\mathbf{W})$ denotes the Aumann expectation of $Z\mathbf{W}$.

   Here we go beyond the results in Beresteanu and Molinari (2008) and derive a moment inequality representation for $\Theta_I(P)$ when $Z$ has a discrete distribution. We denote by $u^r$ the vector $u^r = e^{r\prime}(M_P'M_P)^{-1}M_P'E_P(ZZ')$, $r = 1, \ldots, k$, where $e^r$ is the $r$-th basis vector in $\mathbb{R}^k$ and $M_P$ is a $d \times K$ matrix with $r$-th column equal to $P(Z = z^r)z^r$; we let $q^r = u^r E_P(ZZ')^{-1}$. Observe that for any selection $\tilde{W} \in \mathbf{W}$ *a.s.* one has $u^r E_P(ZZ')^{-1}E_P(Z\tilde{W}) = e^{r\prime}[E_P(\tilde{W}|Z = z^1); \ldots; E_P(\tilde{W}|Z = z^k)]$, so that the support function in direction $u^r$ is maximized/minimized by setting $E_P(\tilde{W}|Z = z^r)$ equal to $E_P(W_1|Z = z^r)$ and $E_P(W_0|Z = z^r)$, respectively. Hence, the identified set can be written in terms of moment inequalities as

$$\begin{aligned}\Theta_I(P) = \{\theta \in \Theta \subset \mathbb{R}^d : {} & q^r[E_P(Z(Z'\theta - W_0 - \mathbf{1}(q^r Z > 0)(W_1 - W_0)))] \leqslant 0 \\ & - q^r[E_P(Z(Z'\theta - W_0 - \mathbf{1}(q^r Z < 0)(W_1 - W_0)))] \leqslant 0, r = 1, \ldots, k\}.\end{aligned} \tag{C.3}$$

   The set is expressed through evaluation of its support function, given in Bontemps, Magnac, and Maurin (2012, Proposition 2), at directions $\pm u^r$; these are the directions orthogonal to the flat faces of $\Theta_I(P)$.

4. **Complete information entry games with pure strategy Nash equilibrium as solution concept.**

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

Here again we assume that the vector $Z$ has $k$ points of support with bounded norm, and the identified set is

$$\Theta_I(P) = \{\theta \in \Theta \subset \mathbb{R}^d : \text{ equations (5.1), (5.2), (5.3), (5.4) hold for all } Z = z^r, r = 1, \ldots, k\}. \tag{C.4}$$

In the first three examples we let $X \equiv (W_0, W_1, Z)'$. In the last example we let $X \equiv (Y_1, Y_2, Z)'$. Throughout, we propose to estimate $E_P(W_\ell | Z = z^r)$ and $E_P(Y_1 = s, Y_2 = t | Z = z^r)$, $\ell = 0, 1$, $(s, t) \in \{0, 1\} \times \{0, 1\}$ and $r = 1, \ldots, k$, using

$$\hat{E}_n(W_\ell | Z = z^r) = \frac{\sum_{i=1}^n W_{\ell,i} \mathbf{1}(Z_i = z^r)}{\sum_{i=1}^n \mathbf{1}(Z_i = z^r)}, \tag{C.5}$$

$$\hat{E}_n(Y_1 = s, Y_2 = t | Z = z^r) = \frac{\sum_{i=1}^n \mathbf{1}(Y_{1,i} = s, Y_{2,i} = t, Z_i = z^r)}{\sum_{i=1}^n \mathbf{1}(Z_i = z^r)}, \tag{C.6}$$

as it is done in, e.g., Ciliberto and Tamer (2009). We assume that for each of the four canonical examples under consideration, Assumption 4.1 as well as one of the assumptions below hold.

ASSUMPTION C.1: *The model $\mathcal{P}$ for $P$ satisfies* $\min_{\ell=0,1} \min_{r=1,\ldots,k} Var_P(W_\ell | Z = z^r) > \underline{\sigma} > 0$ *and* $\min_{r=1,\ldots,k} P(Z = z^r) > \varpi > 0$.

ASSUMPTION C.2: *The model $\mathcal{P}$ for $P$ satisfies: (1)* $\text{eig}(M_P' M_P) > \varsigma$*; (2)* $\text{eig}(E_P(ZZ')) > \varsigma$*; (3)* $\text{eig}(Corr_P([vech(ZZ'); W_0])) > \varsigma$ *and* $\text{eig}(Corr_P([vech(ZZ'); W_1])) > \varsigma$*; for some $\varsigma > 0$, where $vech(A)$ denotes the half-vectorization of the matrix $A$.*

ASSUMPTION C.3: *The model $\mathcal{P}$ for $P$ satisfies* $\min_{r=1,\ldots,k, (s,t) \in \{0,1\} \times \{0,1\}} P(Y_1 = s, Y_2 = t, Z = z^r) > \varpi > 0$.

These are simple to verify low level conditions. We note that Imbens and Manski (2004) and Stoye (2009) directly assume the unconditional version of C.1, while Beresteanu and Molinari (2008) assume C.1 itself.

## C.1 Verification of Assumption B.1 in Theorem B.1

We show that in each of the four examples $\frac{m_j(x,\theta)}{\sigma_{P,j}(\theta)}$, $j = 1, \ldots, J$ is Lipschitz continuous in $\theta \in \Theta$ for all $x \in \mathcal{X}$ and that $D_P$ can be estimated at rate $n^{-1/2}$.

1. **Mean with interval data**. Here $\sigma_{P,\ell}(\theta) = \sigma_{P,\ell}$, and under Assumption C.1 it is uniformly bounded from below. Then

$$\left| \frac{m_j(x,\theta)}{\sigma_{P,j}} - \frac{m_j(x,\theta')}{\sigma_{P,j}} \right| = \frac{\|(\theta' - \theta)\|}{\sigma_{P,j}(\theta)}, \quad \ell = 0, 1,$$

$$D_{P,\ell}(\theta) = \frac{(-1)^{(1-\ell)}}{\sigma_{P,\ell}}, \quad \ell = 0, 1.$$

   Assumption C.1 then guarantees that Assumption B.1 is satisfied.

2. **Linear regression with interval outcome data and discrete regressors.** Here again $\sigma_{P,\ell r}(\theta) = \sigma_{P,\ell r}$, and under Assumptions C.1-C.2 it is uniformly bounded from below. We first consider the rescaled function $\frac{(-1)^j (W_\ell \mathbf{1}(Z = z^r)/P(Z = z^r) - z^{r\prime}\theta)}{\sigma_{P,\ell r}}$:

$$\left| \frac{(-1)^j (W_\ell \mathbf{1}(Z = z^r)/P(Z = z^r) - z^{r\prime}\theta)}{\sigma_{P,\ell r}} - \frac{(-1)^j (W_\ell \mathbf{1}(Z = z^r)/P(Z = z^r) - z^{r\prime}\theta')}{\sigma_{P,\ell r}} \right| = \|z^r\| \frac{\|(\theta' - \theta)\|}{\sigma_{P,\ell r}(\theta)}, \quad \ell = 0, 1,$$

so that Assumption B.1 is satisfied for these rescaled functions by Assumptions C.1-C.2. Next, we observe that

$$D_{P,j} = \frac{(-1)^{(1-j)} z^{r\prime}}{\sigma_{P,\ell r}}, \quad \ell = 0, 1, r = 1, \ldots, k,$$

and it can be estimated at rate $n^{-1/2}$ by Lemma E.12. Theorem B.1 then holds observing that $|P(Z = z^r)/\sum_{i=1}^n \mathbf{1}(Z_i = z^r) - 1| = O_{\mathcal{P}}(n^{-1/2})$ and treating this random element similarly to how we treat $\eta_{n,j}(\cdot)$ in the proof of Theorem B.1.

3. **Best linear prediction with interval outcome data and discrete regressors.** Here

$$m_r(X_i, \theta) = q^r[Z_i(Z_i'\theta - (W_{0,i} + \mathbf{1}(q^r Z_i > 0)(W_{1,i} - W_{0,i})))] \tag{C.7}$$

hence is Lipschitz in $\theta$ with constant $Z_i Z_i'$. Under Assumptions C.1-C.2, $Var_P(m_r(X_i, \theta))$ is uniformly bounded from below, and Lipschitz in $\theta$ with a constant that depends on $Z_i^4$. Hence $\frac{m_r(X_i,\theta)}{\sigma_{P,r}(\theta)}$ is Lipschitz in $\theta$ with a constant that depends on powers of $Z$. Because $Z$ has bounded support, Assumption B.1 is satisfied. A simple argument yields that $D_P$ can be estimated at rate $n^{-1/2}$.

4. **Complete information entry games with pure strategy Nash equilibrium as solution concept.** Here again $\sigma_{P,str}(\theta) = \sigma_{P,str}$, and under Assumptions 4.1 and C.3 it is uniformly bounded from below. The result then follows from a similar argument as the one used in Example 2 (Linear regression with interval outcome data and discrete regressors), observing that the rescaled function of interest is now

$$\frac{\mathbf{1}(Y_1 = s, Y_2 = t|Z = z^r)/P(Z = z^r) - g_{str}(\theta)}{\sigma_{P,str}}, \quad (s,t) \in \{0,1\} \times \{0,1\}, r = 1, \ldots, k,$$

and the gradient is

$$\frac{1}{\sigma_{P,str}} \nabla_\theta g_{str}(\theta), \quad (s,t) \in \{0,1\} \times \{0,1\}, r = 1, \ldots, k,$$

where $g_{str}(\theta)$ are model-implied entry probabilities, and hence taking their values in $[0,1]$. The entry models typically posited assume that payoff shocks have smooth distributions (e.g., multivariate normal), yielding that $\nabla_\theta g_{str}(\theta)$ is well defined and bounded.

## C.2 Verification of Assumption 4.3-(II)

Here we verify Assumption 4.3-(II) for the canonical examples in the moment (in)equalities literature:

1. **Mean with interval data**. In the generalization of this example in Imbens and Manski (2004) and Stoye (2009), equations (4.1)-(4.2) are satisfied by construction, equation (4.3) is directly assumed.

2. **Linear regression with interval outcome data and discrete regressors.** Equation (4.1) is satisfied by construction. Given the estimator that we use for the population moment conditions, we verify equation (4.3) for the variances of the limit distribution of the vector $[\sqrt{n}(\hat{E}_n(W_\ell|Z = z^r) - E_P(W_\ell|Z = z^r))]_{\ell \in \{0,1\}, r=1,\ldots,k}$. We then have that equation (4.3) follows from Assumption C.1. Concerning equation (4.3), this needs to be verified for the correlation matrix of the limit distribution of a $r \times 1$ random vector that for each $r = 1, \ldots, k$ equals any choice in $\{\sqrt{n}(\hat{E}_n(W_0|Z = z^r) - E_P(W_0|Z = z^r)), \sqrt{n}(\hat{E}_n(W_1|Z = z^r) - E_P(W_1|Z = z^r))\}$, which suffices for our results to hold. We then have that (4.2) holds because the correlation matrix is diagonal.

3. **Best linear prediction with interval outcome data and discrete regressors.** Equation (4.1) is again satisfied by construction. Equation (4.2) holds under Assumptions C.1-C.2. Equation (4.3) is verified to hold under Assumption C.1 in Beresteanu and Molinari (2008, p. 808).

4. **Complete information entry games with pure strategy Nash equilibrium as solution concept.** In this case equations (5.3) and (5.4) are paired, but the corresponding moment functions differ by the model implied probability of the region of multiplicity, hence equation (4.1) is satisfied by construction. Given the estimator that we use for the population moment conditions, we verify equations (4.2) and (4.3) for the variances and for the correlation matrix of the limit distribution of the vector $\sqrt{n}(\hat{E}_n(Y_1 = s, Y_2 = t|Z = z^r) - E_P(Y_1 = s, Y_2 = t|Z = z^r)_{(s,t)\in\{0,1\}\times\{0,1\},r=1,\ldots,k})$, which suffices for our results to hold. Equation (4.2) holds provided that $|Corr(Y_{i1}(1 - Y_{i2}), Y_{i1}Y_{i2})| < 1 - \epsilon$ for some $\epsilon > 0$ and Assumption C.3 holds.[46] To see that equation (4.3) also holds, note that Assumption C.3 yields that $P(Y_{i1} = 1, Y_{i2} = 0, Z_i = z^r)$ is uniformly bounded away from 0 and 1, thereby implying that for each $(s,t) \in \{0,1\} \times \{0,1\}, r = 1,\ldots,k$, $(P(Y_1 = s, Y_2 = t|Z = z^r)(1 - P(Y_1 = s, Y_2 = t|Z = z^r)))/(P(Z = z^r)(1 - P(Z = z^r)))$ is uniformly bounded away from zero.

[46] In more general instances with more than two players, it follows if the multinomial distribution of outcomes of the game (reduced by one element) has a correlation matrix with eigenvalues uniformly bounded away from zero.

# Appendix D Proof of Theorems 4.1, 4.2, 4.3 and 4.4

## D.1 Notation and Structure of the Proof of Theorem 4.1

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 $\sup_{P\in\mathcal{P}} P(|X_n/a_n| > \epsilon) < \eta, \forall n \geqslant N$. 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 $\sup_{P\in\mathcal{P}} P(|X_n/a_n| > M) < \eta, \forall n \geqslant N$.

Table D.0: Important notation. Here $(P_n, \theta_n) \in \{(P,\theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ is a subsequence as defined in (D.3)-(D.4) below, $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$, $B^d = \{x \in \mathbb{R}^d : |x_i| \leqslant 1, i = 1,\ldots,d\}$, $B^d_{n,\rho} \equiv \frac{\sqrt{n}}{\rho}(\Theta - \theta_n) \cap B^d$, $\mathfrak{B}^d_\rho = \lim_{n\to\infty} B^d_{n,\rho}$, 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. $\theta$, 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 $\forall \theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$ by Lemma E.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. |
| $\varphi^*_j(\xi)$ | = | $\begin{cases} \varphi_j(\xi) & \pi_{1,j} = 0 \\ -\infty & \pi_{1,j} < 0 \\ 0 & j = J_1+1,\cdots,J. \end{cases}$ | Infeasible GMS that is less conservative than $\varphi_j$. |
| $u_{n,j,\theta_n}(\lambda)$ | = | $\{\mathbb{G}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\}(1 + \eta_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}))$ | Mean value expansion of nonlinear constraints with sample empirical process and "oracle" GMS, with $\bar{\theta}_n$ componentwise between $\theta_n$ and $\theta_n + \frac{\lambda\rho}{\sqrt{n}}$. |
| $U_n(\theta_n, c)$ | = | $\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap u_{n,j,\theta_n}(\lambda) \leqslant c,\ \forall j = 1,\ldots,J\}$ | Feasible set for nonlinear sample problem intersected with $p'\lambda = 0$. |
| $\mathfrak{w}_j(\lambda)$ | = | $\mathbb{Z}_j + \rho D_j\lambda + \pi^*_{1,j}$ | Linearized constraints with a Gaussian shift and "oracle" GMS. |
| $\mathfrak{W}(c)$ | = | $\{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c,\ \forall j = 1,\ldots,J\}$ | Feasible set for linearized limit problem intersected with $p'\lambda = 0$. |
| $c_{\pi*}$ | = | $\inf\{c \in \mathbb{R}_+ : \Pr(\mathfrak{W}(c) \neq \varnothing) \geqslant 1-\alpha\}$. | Limit problem critical level. |
| $v^b_{n,j,\theta_n'}(\lambda)$ | = | $\mathbb{G}^b_{n,j}(\theta_n') + \rho\hat{D}_{n,j}(\theta_n')\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta_n'))$ | Linearized constraints with bootstrap empirical process and sample GMS. |
| $V^b_n(\theta_n', c)$ | = | $\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap v^b_{n,j,\theta_n'}(\lambda) \leqslant c,\ \forall j = 1,\ldots,J\}$ | Feasible set for linearized bootstrap problem with sample GMS and $p'\lambda = 0$. |
| $v^I_{n,j,\theta_n'}(\lambda)$ | = | $\mathbb{G}^b_{n,j}(\theta_n') + \rho\hat{D}_{n,j}(\theta_n')\lambda + \varphi^*_j(\hat{\xi}_{n,j}(\theta_n'))$ | Linearized constraints with bootstrap empirical process and infeasible sample GMS. |
| $V^I_n(\theta_n', c)$ | = | $\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap v^I_{n,j,\theta_n'}(\lambda) \leqslant c,\ \forall j = 1,\ldots,J\}$ | Feasible set for linearized bootstrap problem with infeasible sample GMS and $p'\lambda = 0$. |
| $\hat{c}_n(\theta)$ | = | $\inf\{c \in \mathbb{R}_+ : P^*(V^b_n(\theta, c) \neq \varnothing) \geqslant 1-\alpha\}$ | Bootstrap critical level. |
| $\hat{c}_{n,\rho}(\theta)$ | = | $\inf_{\lambda\in B^d_{n,\rho}} \hat{c}_n(\theta + \frac{\lambda\rho}{\sqrt{n}})$ | Smallest value of the bootstrap critical level in a $B^d_{n,\rho}$ neighborhood of $\theta$. |
| $\hat{\sigma}^M_{n,j}(\theta)$ | = | $\hat{\mu}_{n,j}(\theta)\hat{\sigma}_{n,j}(\theta) + (1 - \hat{\mu}_{n,j}(\theta))\hat{\sigma}_{n,j+R_1}(\theta)$ | Weighted sum of the estimators of the standard deviations of paired inequalities |

Figure D.1: Structure of Lemmas used in the proof of Theorem 4.1.

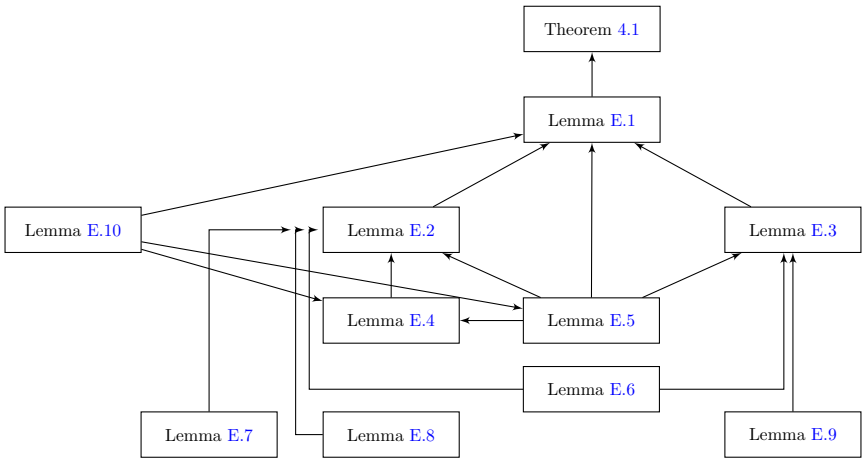


Table D.1: Heuristics for the role of each Lemma in the proof of Theorem 4.1. Notes: (i) Uniformity in Theorem 4.1 is enforced arguing along subsequences; (ii) When needed, random variables are realized on the same probability space as shown in Lemma E.1 and Lemma E.17 (see Appendix E.3 for details); (iii) Here $(P_n, \theta_n) \in \{(P, \theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ is a subsequence as defined in (D.3)-(D.4) below; (iv) All results hold for any $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$.

| | |
|---|---|
| Theorem 4.1 | $P_n(p'\theta_n \in CI) \geqslant P_n\left(U_n(\theta_n, \hat{c}_{n,\rho}(\theta_n)) \neq \varnothing\right)$.<br>Coverage is conservatively estimated by the probability that $U_n$ is nonempty. |
| Lemma E.1 | $\liminf P_n\left(U_n(\theta_n, \hat{c}_{n,\rho}(\theta_n)) \neq \varnothing\right) \geqslant 1 - \alpha$. |
| Lemma E.2 | $P_n(U(\theta_n, c_n^I(\theta_n)) \neq \varnothing, \mathfrak{W}(c_{\pi*}) = \varnothing) + P_n(U(\theta_n, c_n^I(\theta_n)) = \varnothing, \mathfrak{W}(c_{\pi*}) \neq \varnothing) = o_{\mathcal{P}}(1)$.<br>Argued by comparing $U_n$ and its limit $\mathfrak{W}$ (after coupling). |
| Lemma E.3 | $P_n^*(V_n^I(\theta_n', c) \neq \varnothing) - \Pr(\mathfrak{W}(c) \neq \varnothing) \to 0$ and $c_n^I(\theta_n') \xrightarrow{P_n} c_{\pi*}$ if $c_{\pi*} > 0$.<br>The bootstrap critical value that uses the less conservative GMS yileds a convergent critical value. |
| Lemma E.4 | $\sup_{\lambda \in B^d} \lvert \max_j(u_{n,j,\theta_n}(\lambda) - c_n^I(\theta_n)) - \max_j(\mathfrak{w}_j(\lambda) - c_{\pi*})\rvert = o_{\mathcal{P}}(1)$, and similarly for $\mathfrak{w}_j$ and $v_{n,j,\theta_n'}^I$.<br>The criterion functions entering $U_n$ and $\mathfrak{W}$ converge to each other. |
| Lemma E.5 | Local-to-binding constraints are selected by GMS uniformly over the $\rho$-box (intuition: $\rho n^{-1/2} = o_{\mathcal{P}}(\kappa_n^{-1})$),<br>and $\lVert\hat{\xi}_n(\theta_n') - \kappa_n^{-1}\sqrt{n}\sigma_{P_n,j}^{-1}(\theta_n')E_{P_n}[m_j(X_i, \theta_n')]\rVert = o_{\mathcal{P}}(1)$. |
| Lemma E.6 | $\forall \eta > 0\ \exists \delta > 0, : \Pr(\{\mathfrak{W}(c) \neq \varnothing\} \cap \{\mathfrak{W}^{-\delta}(c) = \varnothing\}) < \eta$, and similarly for $V_n^I$.<br>It is unlikely that these sets are nonempty but become empty upon slightly tightening stochastic constraints. |
| Lemma E.7 | Intersections of constraints whose gradients are almost linearly dependent are unlikely to realize inside $\mathfrak{W}$.<br>Hence, we can ignore irregularities that occur as linear dependence is approached. |
| Lemma E.8 | If there are weakly more equality constraints than parameters, then $c$ is uniformly bounded away from zero.<br>This simplifies some arguments. |
| Lemma E.9 | If two paired inequalities are local to binding, then they are also asymptotically identical up to sign.<br>This justifies "merging" them. |
| Lemma E.10 | $\eta_{n,j}(\cdot)$ converges to zero uniformly in $P$ and $\theta$. |

## D.2 Proof of Theorems 4.1 and 4.2

### D.2.1 Main Proofs

**Proof of Theorem 4.1**

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

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

$\gamma_{2,P} = (s(p, \Theta_I(P)), vech(\Omega_P(\theta)), vec(D_P(\theta)))$, 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

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

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

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

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

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

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

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

The projection of $\theta_n$ is covered when

$$\begin{aligned}
&-s(-p, \mathcal{C}_n(\hat{c}_n)) \leqslant p'\theta_n \leqslant 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)} \leqslant \hat{c}_n(\vartheta), \forall j \end{array} \right\} \leqslant p'\theta_n \leqslant \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)} \leqslant \hat{c}_n(\vartheta), \forall j \end{array} \right\} \\
&\Leftrightarrow \left\{ \begin{array}{c} \inf_\lambda p'\lambda \\ s.t. \lambda \in \frac{\sqrt{n}}{\rho}(\Theta - \theta_n), \quad \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} \leqslant \hat{c}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \forall j \end{array} \right\} \leqslant 0 \\
&\qquad \leqslant \left\{ \begin{array}{c} \sup_\lambda p'\lambda \\ s.t. \lambda \in \frac{\sqrt{n}}{\rho}(\Theta - \theta_n), \quad \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} \leqslant \hat{c}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \forall j \end{array} \right\} \qquad \text{(D.6)} \\
&\Leftrightarrow \left\{ \begin{array}{c} \inf_\lambda p'\lambda \\ s.t. \lambda \in \frac{\sqrt{n}}{\rho}(\Theta - \theta_n), \\ \{\mathbb{G}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})\}(1 + \eta_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})) \leqslant \hat{c}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \forall j \end{array} \right\} \leqslant 0 \\
&\qquad \leqslant \left\{ \begin{array}{c} \sup_\lambda p'\lambda \\ s.t. \lambda \in \frac{\sqrt{n}}{\rho}(\Theta - \theta_n), \\ \{\mathbb{G}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1 + \eta_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})) \leqslant \hat{c}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \forall j \end{array} \right\}, \qquad \text{(D.7)}
\end{aligned}$$

with $\eta_{n,j}(\cdot) \equiv \sigma_{P,j}(\cdot)/\hat{\sigma}_{n,j}(\cdot) - 1$ and where we localized $\vartheta$ in a $\sqrt{n}/\rho$-neighborhood of $\Theta - \theta_n$ and we took a mean

value expansion yielding $\forall j$

$$\frac{\sqrt{n}\bar{m}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}=\{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})). \tag{D.8}$$

Denote $B^d_{n,\rho}\equiv\frac{\sqrt{n}}{\rho}(\Theta-\theta_n)\cap B^d$, with $B^d=\{x\in\mathbb{R}^d:|x_i|\leqslant 1,i=1,\ldots,d\}$. The event in (D.7) is implied by

$$\begin{aligned}\Leftarrow&\left\{\begin{array}{c}\inf_\lambda p'\lambda\\ s.t.\lambda\in B^d_{n,\rho},\\ \{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}),\forall j\end{array}\right\}\leqslant 0\\ \leqslant&\left\{\begin{array}{c}\sup_\lambda p'\lambda\\ s.t.\lambda\in B^d_{n,\rho},\\ \{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}),\forall j\end{array}\right\},\end{aligned} \tag{D.9}$$

**Step 2.** This step is used only when Assumption 4.3-(II) is invoked. When this assumption is invoked, recall that in equation (2.5) we use the estimator specified in Lemma E.10 equation (E.188) for $\sigma_{P,j},j=1,\ldots,2R_1$ (with $R_1\leqslant J_1/2$ defined in the statement of the assumption). In equation (3.1) we use the sample analog estimators of $\sigma_{P,j}$ for all $j=1,\ldots,J$. To keep notation manageable, we explicitly denote the estimator used in (2.5) by $\hat{\sigma}^M_j$ only in this step but in almost all other parts of this Appendix we use the generic notation $\hat{\sigma}_j$.

For each $j=1,\ldots,R_1$ such that

$$\pi^*_{1,j}=\pi^*_{1,j+R_1}=0, \tag{D.10}$$

where $\pi^*_1$ is defined in (D.5), let

$$\tilde{\mu}_j=\left\{\begin{array}{cc}1 & \text{if } \gamma_{1,P_n,j}(\theta_n)=0=\gamma_{1,P_n,j+R_1}(\theta_n),\\ \frac{\gamma_{1,P_n,j+R_1}(\theta_n)(1+\eta_{n,j+R_1}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))}{\gamma_{1,P_n,j+R_1}(\theta_n)(1+\eta_{n,j+R_1}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))+\gamma_{1,P_n,j}(\theta_n)(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))} & \text{otherwise,}\end{array}\right. \tag{D.11}$$

$$\tilde{\mu}_{j+R_1}=\left\{\begin{array}{cc}0 & \text{if } \gamma_{1,P_n,j}(\theta_n)=0=\gamma_{1,P_n,j+R_1}(\theta_n),\\ \frac{\gamma_{1,P_n,j}(\theta_n)(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))}{\gamma_{1,P_n,j+R_1}(\theta_n)(1+\eta_{n,j+R_1}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))+\gamma_{1,P_n,j}(\theta_n)(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))} & \text{otherwise,}\end{array}\right. \tag{D.12}$$

For each $j=1,\ldots,R_1$, replace the constraint indexed by $j$, that is

$$\frac{\sqrt{n}\bar{m}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}^M_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}\leqslant\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}), \tag{D.13}$$

with the following weighted sum of the paired inequalities

$$\tilde{\mu}_j\frac{\sqrt{n}\bar{m}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}^M_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}-\tilde{\mu}_{j+R_1}\frac{\sqrt{n}\bar{m}_{j+R_1,n}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}^M_{n,j+R_1}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}\leqslant\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}), \tag{D.14}$$

and for each $j=1,\ldots,R_1$, replace the constraint indexed by $j+R_1$, that is

$$\frac{\sqrt{n}\bar{m}_{j+R_1,n}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}^M_{n,j+R_1}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})}\leqslant\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}), \tag{D.15}$$

with

$$-\tilde{\mu}_j \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}^M_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} + \tilde{\mu}_{j+R_1} \frac{\sqrt{n}\bar{m}_{j+R_1,n}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}^M_{n,j+R_1}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} \leqslant \hat{c}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \tag{D.16}$$

It then follows from Assumption 4.3-(II) that these replacements are conservative because

$$\frac{\bar{m}_{j+R_1,n}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}^M_{n,j+R_1}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} \leqslant -\frac{\bar{m}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}^M_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})},$$

and therefore (D.14) implies (D.13) and (D.16) implies (D.15).

**Step 3.** Next, we make the following comparisons:

$$\pi^*_{1,j} = 0 \Rightarrow \pi^*_{1,j} \geqslant \sqrt{n}\gamma_{1,P_n,j}(\theta_n), \tag{D.17}$$

$$\pi^*_{1,j} = -\infty \Rightarrow \sqrt{n}\gamma_{1,P_n,j}(\theta_n) \to -\infty. \tag{D.18}$$

For any constraint $j$ for which $\pi^*_{1,j} = 0$, (D.17) yields that replacing $\sqrt{n}\gamma_{1,P_n,j}(\theta_n)$ in (D.9) with $\pi^*_{1,j}$ introduces a conservative distortion. Under Assumption 4.3-(II), for any $j$ such that (D.10) holds, the substitutions in (D.14) and (D.16) yield $\tilde{\mu}_j\sqrt{n}\gamma_{1,P_n,j}(\theta_n)(1 + \eta_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})) - \tilde{\mu}_{j+R_1}\sqrt{n}\gamma_{1,P_n,j+R_1}(\theta_n)(1 + \eta_{n,j+R_1}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})) = 0$, and therefore replacing this term with $\pi^*_{1,j} = 0 = \pi^*_{1,j+R_1}$ is inconsequential.

For any $j$ for which $\pi^*_{1,j} = -\infty$, (D.18) 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 4.4 (i)-(ii), and $\lambda \in B^d_{n,\rho}$, it follows that

$$\rho D_{P_n,j}(\bar{\theta}_n)\lambda \leqslant \rho\sqrt{d}(\|D_{P_n,j}(\bar{\theta}_n) - D_{P_n,j}(\theta_n)\| + \|D_{P_n,j}(\theta_n)\|) \leqslant \rho\sqrt{d}(\rho M/\sqrt{n} + \bar{M}), \tag{D.19}$$

where $\bar{M}$ and $M$ are as defined in Assumption 4.4-(i) and (ii) respectively, and we used that $\bar{\theta}_n$ lies component-wise between $\theta_n$ and $\theta_n + \frac{\lambda\rho}{\sqrt{n}}$. Using that $\mathbb{G}_{n,j}$ is asymptotically tight by Assumption 4.5, we have that for any $\tau > 0$, there exists a $T > 0$ and $N_1 \in \mathbb{N}$ such that for all $n \geqslant N_1$,

$$P_n\left(\max_{j:\pi^*_{1,j}=-\infty}\{\mathbb{G}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1 + \eta_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})) \leqslant 0,\ \forall\lambda \in B^d_{n,\rho}\right) > 1 - \tau/2. \tag{D.20}$$

To see this, note that $\pi^*_{ij} = -\infty$ if and only if $\lim_{n\to\infty}\frac{\sqrt{n}}{\kappa_n}\gamma_{1P_nj}(\theta_n) = \pi_{1j} \in [-\infty, 0)$. Suppose first that $\pi_{1j} > -\infty$. Then for all $\epsilon > 0$ there exists $N_2 \in \mathbb{N}$ such that $\left|\frac{\sqrt{n}}{\kappa_n}\gamma_{1P_nj}(\theta_n) - \pi_{1j}\right| \leqslant \epsilon$, for all $n \geqslant N_2$. Choose $\epsilon > 0$ such that

$\pi_{1j}+\epsilon<0$. Let $N=\max\{N_1,N_2\}$. Then we have

$$
\begin{aligned}
&P_n\left(\max_{j:\pi^*_{1,j}=-\infty}\{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant 0,\ \forall\lambda\in B^d_{n,\rho}\right)\\
\geqslant\ &P_n\left(\max_{j:\pi^*_{1,j}=-\infty}\{T+\rho(\bar{M}+\rho M/\sqrt{n})+\sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant 0\cap\max_{j:\pi^*_{1,j}=-\infty}\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})\leqslant T\right)\\
\geqslant\ &P_n\left(\max_{j:\pi^*_{1,j}=-\infty}\{T+\rho(\bar{M}+\rho M/\sqrt{n})+\kappa_n(\pi_{1j}+\epsilon)\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant 0\cap\max_{j:\pi^*_{1,j}=-\infty}\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})\leqslant T\right)\\
=\ &P_n\left(\max_{j:\pi^*_{1,j}=-\infty}\left\{\frac{T}{\kappa_n}+\frac{\rho}{\kappa_n}(\bar{M}+\rho M/\sqrt{n})+(\pi_{1j}+\epsilon)\right\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant 0\cap\max_{j:\pi^*_{1,j}=-\infty}\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})\leqslant T\right)\\
=&P_n\left(\max_{j:\pi^*_{1,j}=-\infty}\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})\leqslant T\right)>1-\tau/2,\ \forall n\geqslant N.
\end{aligned}
$$

If $\pi_{1j}=-\infty$ the same argument applies a fortiori. We therefore have that for $n\geqslant N$,

$$
\begin{aligned}
&P_n\left(\left\{\begin{array}{c}\inf_\lambda p'\lambda\\ s.t.\lambda\in B^d_{n,\rho},\\ \{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}),\forall j\end{array}\right\}\leqslant 0\right.\\
&\qquad\left.\leqslant\left\{\begin{array}{c}\sup_\lambda p'\lambda\\ s.t.\lambda\in B^d_{n,\rho},\\ \{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}),\forall j\end{array}\right\}\right)
\end{aligned}
\tag{D.21}
$$

$$
\begin{aligned}
\geqslant&P_n\left(\left\{\begin{array}{c}\inf_\lambda p'\lambda\\ s.t.\lambda\in B^d_{n,\rho},\\ \{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\pi^*_{1,j}\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}),\forall j\end{array}\right\}\leqslant 0\right.\\
&\qquad\left.\leqslant\left\{\begin{array}{c}\sup_\lambda p'\lambda\\ s.t.\lambda\in B^d_{n,\rho},\\ \{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\pi^*_{1,j}\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}),\forall j\end{array}\right\}\right)-\tau/2.
\end{aligned}
\tag{D.22}
$$

Since the choice of $\tau$ is arbitrary, the limit of the term in (D.21) is not smaller than the limit of the first term in (D.22). Hence, we continue arguing for the event whose probability is evaluated in (D.22).

Finally, by definition $\hat{c}_n(\cdot)\geqslant 0$ and therefore $\inf_{\lambda\in B^d_{n,\rho}}\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}})$ exists. Therefore, the event whose probability is evaluated in (D.22) is implied by the event

$$
\begin{aligned}
&\left\{\begin{array}{c}\inf_\lambda p'\lambda\\ s.t.\lambda\in B^d_{n,\rho},\\ \{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\pi^*_{1,j}\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant\inf_{\lambda\in B^d_{n,\rho}}\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}),\forall j\end{array}\right\}\leqslant 0\\
&\qquad\leqslant\left\{\begin{array}{c}\sup_\lambda p'\lambda\\ s.t.\lambda\in B^d_{n,\rho},\\ \{\mathbb{G}_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}})+\rho D_{P_n,j}(\bar{\theta}_n)\lambda+\pi^*_{1,j}\}(1+\eta_{n,j}(\theta_n+\frac{\lambda\rho}{\sqrt{n}}))\leqslant\inf_{\lambda\in B^d_{n,\rho}}\hat{c}_n(\theta_n+\frac{\lambda\rho}{\sqrt{n}}),\forall j\end{array}\right\}
\end{aligned}
\tag{D.23}
$$

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

$$u_{n,j,\theta_n}(\lambda) \equiv \{\mathbb{G}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\}(1 + \eta_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})), \tag{D.24}$$

where under Assumption 4.3-(II) when $\pi^*_{1,j} = 0$ and $\pi^*_{1,j+R_1} = 0$ the substitutions of equation (D.13) with equation (D.14) and of equation (D.15) with equation (D.16) have been performed. Let

$$U_n(\theta_n, c) \equiv \left\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap u_{n,j,\theta_n}(\lambda) \leqslant c,\ \forall j = 1, \ldots, J\right\}, \tag{D.25}$$

and define

$$\hat{c}_{n,\rho} \equiv \inf_{\lambda \in B^d_{n,\rho}} \hat{c}_n(\theta + \frac{\lambda\rho}{\sqrt{n}}). \tag{D.26}$$

Then by (D.23) and the definition of $U_n$, we obtain

$$P_n(p'\theta_n \in CI_n) \geqslant P_n\left(U_n(\theta_n, \hat{c}_{n,\rho}) \neq \varnothing\right). \tag{D.27}$$

By passing to a further subsequence, we may assume that

$$D_{P_n}(\theta_n) \to D, \tag{D.28}$$

for some $J \times d$ matrix $D$ such that $\|D\| \leqslant M$ and $\Omega_{P_n} \xrightarrow{u} \Omega$ for some correlation matrix $\Omega$. By Lemma 2 in Andrews and Guggenberger (2009) and Assumption 4.5 (i), uniformly in $\lambda \in B^d$, $\mathbb{G}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) \xrightarrow{d} \mathbb{Z}$ for a normal random vector with the correlation matrix $\Omega$. By Lemma E.1,

$$\liminf_{n\to\infty} P_n\left(U_n(\theta_n, \hat{c}_{n,\rho}) \neq \varnothing\right) \geqslant 1 - \alpha. \tag{D.29}$$

The conclusion of the theorem then follows from (D.2), (D.3), (D.27), and (D.29). □

**Proof of Theorem 4.2**

The argument of proof is the same as for Theorem 4.1, with the following modification. Take $(P_n, \theta_n)$ as defined following equation (D.4). Then $f(\theta_n)$ is covered when

$$\begin{Bmatrix} \inf f(\vartheta) \\ s.t.\ \vartheta \in \Theta, \quad \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leqslant \hat{c}^f_n(\vartheta), \forall j \end{Bmatrix} \leqslant f(\theta_n) \leqslant \begin{Bmatrix} \sup f(\vartheta) \\ s.t.\ \vartheta \in \Theta, \quad \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leqslant \hat{c}^f_n(\vartheta), \forall j \end{Bmatrix}$$

$$\Leftrightarrow \begin{Bmatrix} \inf_\lambda \nabla f(\tilde{\theta}_n)\lambda \\ s.t. \lambda \in \frac{\sqrt{n}}{\rho}(\Theta - \theta_n), \quad \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} \leqslant \hat{c}^f_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \forall j \end{Bmatrix} \leqslant 0$$

$$\leqslant \begin{Bmatrix} \sup_\lambda \nabla f(\tilde{\theta}_n)\lambda \\ s.t. \lambda \in \frac{\sqrt{n}}{\rho}(\Theta - \theta_n), \quad \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} \leqslant \hat{c}^f_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \forall j \end{Bmatrix},$$

where we took a mean value expansion yielding

$$f(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) = f(\theta_n) + \frac{\rho}{\sqrt{n}}\nabla f(\tilde{\theta}_n)\lambda, \tag{D.30}$$

for $\tilde{\theta}_n$ a mean value that lies componentwise between $\theta_n$ and $\theta_n + \frac{\lambda\rho}{\sqrt{n}}$, and we used that the sign of the last term in (D.30) is the same as the sign of $\nabla f(\tilde{\theta}_n)\lambda$. With the objective function in (D.30) so redefined, all expression in the proof of Theorem 4.1 up to (D.24) continue to be valid. We can then redefine the set $U_n(\theta_n, c)$ in (D.25) as

$$U_n(\theta_n, c) \equiv \left\{\lambda \in B^d_{n,\rho} : \|\nabla f(\tilde{\theta}_n)\|^{-1}\nabla f(\tilde{\theta}_n)\lambda = 0 \cap u_{n,j,\theta_n}(\lambda) \leqslant c,\ \forall j = 1,\dots,J\right\}.$$

Replace $p'$ with $\|\nabla f(\tilde{\theta}_n)\|^{-1}\nabla f(\tilde{\theta}_n)$ in all expressions involving the set $U_n(\theta_n, \hat{c}^f_{n,\rho}(\theta_n))$, and replace $p'$ with $\|\nabla f(\theta_n)'\|^{-1}\nabla f(\theta'_n)$ in all expressions for the sets $V^I_n(\theta'_n, \hat{c}^f_n(\theta'_n))$, and in all the almost sure representation counterparts of these sets. Observe that we can select a convergent subsequence from $\{\|\nabla f(\theta_n)'\|^{-1}\nabla f(\theta'_n)\}$ that converges to some $p$ in the unit sphere, so that the form of $\mathfrak{W}(c_{\pi*})$ in (E.17) is unchanged. This yields the result, noting that by the assumption $\|\nabla f(\tilde{\theta}_n) - \nabla f(\theta'_n)\| = O_{\mathcal{P}}(\rho/\sqrt{n})$ □

### D.2.2 A High Level Condition Replacing Assumption 4.3 and the $\rho$-Box Constraints

Next, we consider an assumption which is composed of two parts. The first part aims at informally mimicking Assumption A.2 in Bugni, Canay, and Shi (2017) and replaces Assumption 4.3. The second part replaces the use of the $\rho$-box constraints. Below, for a given set $A \subset \mathbb{R}^d$, let $\|A\|_H = \sup_{a\in A}\|a\|$ denote its Hausdorff norm.

ASSUMPTION D.1: *Consider any sequence* $\{P_n, \theta_n\} \in \{(P,\theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ *such that*

$$\begin{aligned}\kappa_n^{-1}\sqrt{n}\gamma_{1,P_n,j}(\theta_n) &\to \pi_{1j} \in \mathbb{R}_{[-\infty]},\ j = 1,\dots,J,\\ \Omega_{P_n} &\xrightarrow{u} \Omega,\\ D_{P_n}(\theta_n) &\to D.\end{aligned}$$

*Let* $\pi^*_{1j} = 0$ *if* $\pi_{1j} = 0$ *and* $\pi^*_{1j} = -\infty$ *if* $\pi_{1j} < 0$. *Let* $\mathbb{Z}$ *be a Gaussian process with covariance kernel* $\Omega$. *Let*

$$\mathfrak{w}_j(\lambda) \equiv \mathbb{Z}_j + \rho D_j\lambda + \pi^*_{1,j}. \tag{D.31}$$

*(I) Let*

$$\mathfrak{W}(c) \equiv \left\{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c,\ \forall j = 1,\dots,J\right\}, \tag{D.32}$$

$$c_{\pi*} \equiv \inf\{c \in \mathbb{R}_+ : \Pr(\mathfrak{W}(c) \neq \varnothing) \geqslant 1-\alpha\}. \tag{D.33}$$

*Then:*

*(a) If* $c_{\pi*} > 0$, $\Pr(\mathfrak{W}(c) \neq \varnothing)$ *is continuous and strictly increasing at* $c = c_{\pi*}$.

*(b) If* $c_{\pi*} = 0$, $\liminf_{n\to\infty} P_n(U_n(\theta_n, 0) \neq \varnothing) \geqslant 1-\alpha$, *where* $U_n(\theta_n, c)$, $c \geqslant 0$ *is as in* (D.25).

*(II) Let*

$$\bar{\mathfrak{W}}(c) \equiv \left\{\lambda \in \mathbb{R}^d : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c,\ \forall j = 1,\dots,J\right\},$$

*which differs from* (D.32) *by not constraining* $\lambda$ *to* $\mathfrak{B}^d_\rho$, *and let* $\bar{c} \equiv \Phi^{-1}(1-\alpha/J)$ *denote the asymptotic Bonferroni critical value. Then for every* $\eta > 0$ *there exists* $M_\eta < \infty$ *s.t.* $\Pr(\|\bar{\mathfrak{W}}(\bar{c})\|_H > M_\eta) \leqslant \eta$.

### D.2.3 Proof of Theorem 4.1 with High Level Assumption D.1-(I) Replacing Assumption 4.3, and Dropping the $\rho$-Box Constraints Under Assumption D.1-(II)

LEMMA D.1: *Suppose that Assumption 4.1, 4.2, 4.4 and 4.5 hold.*

*(I) Let also Assumption D.1-(I) hold. Let* $0 < \alpha < 1/2$. *Then,*

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

*(II) Let also Assumption D.1-(II) and either Assumption 4.3 or D.1-(I) hold. Let* $\hat{c}_n = \inf\{c \in \mathbb{R}_+ : P^*(\{\Lambda_n^b(\theta, +\infty, c) \cap \{p'\lambda = 0\}\} \neq \varnothing) \geqslant 1-\alpha\}$, *where* $\Lambda_n^b$ *is defined in equation (3.1) and* $CI_n \equiv [-s(-p, \mathcal{C}_n(\hat{c}_n)), s(p, \mathcal{C}_n(\hat{c}_n))]$ *with* $s(q, \mathcal{C}_n(\hat{c}_n)), q \in \{p, -p\}$ *defined in equation (2.5). Then*

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

*Proof.* We establish each part of the Lemma separately.

**Part (I).** This part of the lemma replaces Assumptions 4.3 with Assumption D.1-(I). Hence we establish the result by showing that all claims that were made under Assumption 4.3 remain valid under Assumption D.1-(I). We proceed in steps.

Step 1. Revisiting the proof of Lemma E.6, equation (E.133).

Let $\mathcal{J}^*$ be as defined in (E.29). If $\mathcal{J}^* = \varnothing$ we immediately have that Lemma E.6 continues to hold. Hence we assume that $\mathcal{J}^* \neq \varnothing$. To keep the notation simple, below we argue as if all $j = 1, \ldots, J$ belong to $\mathcal{J}^*$.

Consider the case that $c_{\pi *} > 0$. For some $c_{\pi *} > \delta > 0$, let

$$\mathfrak{W}(c-\delta) \equiv \left\{\lambda \in \mathfrak{B}_\rho^d : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c-\delta,\ \forall j = 1, \ldots, J\right\}, \tag{D.34}$$

where we emphasize that the set $\mathfrak{W}(c-\delta)$ is obtained by a $\delta$-contraction of all constraints, including those indexed by $j = J_1 + 1, \ldots, J$. By Assumption D.1-(I), for any $\eta > 0$ there exists a $\delta$ such that

$$\eta \geqslant \left|\Pr\left(\mathfrak{W}(c_{\pi *}) \neq \varnothing\right) - \Pr\left(\mathfrak{W}(c_{\pi *} - \delta) \neq \varnothing\right)\right| = \Pr\left(\{\mathfrak{W}(c_{\pi *}) \neq \varnothing\} \cap \{\mathfrak{W}(c_{\pi *} - \delta) = \varnothing\}\right),$$
$$\eta \geqslant \left|\Pr\left(\mathfrak{W}(c_{\pi *} + \delta) \neq \varnothing\right) - \Pr\left(\mathfrak{W}(c_{\pi *}) \neq \varnothing\right)\right| = \Pr\left(\{\mathfrak{W}(c_{\pi *} + \delta) \neq \varnothing\} \cap \{\mathfrak{W}(c_{\pi *}) = \varnothing\}\right).$$

The result follows.

Step 2. Revisiting the proof of Lemma E.2.

Case 1 of Lemma E.2 is unaltered. Case 2 of Lemma E.2 follows from the same argument as used in Case 1 of Lemma E.2, because under Assumption D.1-(I) as shown in step 1 of this proof all inequalities are tightened. In Case 3 of Lemma E.2 the result in (D.29) holds automatically by Assumption D.1-(I)-(ii). (As a remark, Lemmas E.7-E.8 are no longer needed to establish Lemma E.2.)

Step 3. Revisiting the proof of Lemma E.3. Under Assumption D.1 we do not need to merge paired inequalities. Hence, part (iii) of Lemma E.3 holds automatically because $\varphi_j^*(\xi) \leqslant \varphi_j(\xi)$ for any $j$ and $\xi$. We are left to establish parts (i) and (ii) of Lemma E.3. These follow immediately, because Lemma E.6 remains valid as shown in step 1 and by Assumption D.1-(I), $\Pr(\mathfrak{W}(c) \neq \varnothing)$ is strictly increasing at $c = c_{\pi *}$ if $c_{\pi *} > 0$. (As a remark, Lemma E.9 is no longer needed to establish Lemma E.3.)

In summary, the desired result follows by applying Lemma E.1 in the proof of Theorem 4.1 as Lemmas E.2, E.3 and E.6 remain valid, Lemmas E.4, E.5, E.10 and the Lemmas in Appendix E.3 are unaffected, and Lemmas E.7, E.8, E.9 are no longer needed.

**Part (II).** This is established by adapting the proof of Theorem 4.1 as follows:

In the main proof, we pass to an a.s. representation early on, so that $\mathfrak{W}$ realizes jointly with other random variables (we denote almost sure representations adding a superscript "*" on the original variable). At the same

time, we entirely drop $\rho$. This means that algebraic expressions, e.g. in the main proof, simplify as if $\rho = 1$, but it also removes any constraints along the lines of $\lambda \in B^d_{n,\rho}$ in equation (D.9). Indeed, (D.9) is replaced by:

$$\cdots \Leftarrow \left\{ \begin{array}{c} \inf_\lambda p'\lambda \\ s.t. \lambda \in \bar{\mathfrak{W}}^*(\bar{c}), \\ \{\mathbb{G}^*_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1 + \eta_{n,j}(\theta_n + \lambda/\sqrt{n})) \leqslant \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\} \leqslant 0$$

$$\leqslant \left\{ \begin{array}{c} \sup_\lambda p'\lambda \\ s.t. \lambda \in \bar{\mathfrak{W}}^*(\bar{c}), \\ \{\mathbb{G}^*_{n,j}(\theta_n + \lambda/\sqrt{n}) + D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\}(1 + \eta_{n,j}(\theta_n + \lambda/\sqrt{n})) \leqslant \hat{c}_n(\theta_n + \lambda/\sqrt{n}), \forall j \end{array} \right\},$$

yielding a new definition of the set $U^*_n$ as

$$U^*_n(\theta_n, c) \equiv \left\{\lambda \in \bar{\mathfrak{W}}^*(\bar{c}) : p'\lambda = 0 \cap u^*_{n,j,\theta_n}(\lambda) \leqslant c,\ \forall j = 1, \ldots, J\right\}.$$

Subsequent uses of $\rho$ in the main proof use that $\|\lambda\| \leqslant \sqrt{d}\rho = O_{\mathcal{P}}(1)$. For example, consider the argument following equation (E.30) or the argument just preceding equation (D.29), and so on. All these continue to go through because $\bar{\mathfrak{W}}^*(\bar{c}) = O(1)$ by assumption.

Similar uses occur in Lemma E.1. The next major adaptation is that in (E.27) and (E.28): we again drop $\rho$ but nominally introduce the constraint that $\lambda \in \bar{\mathfrak{W}}^*(\bar{c})$. However, for $c \leqslant \bar{c}$, this condition cannot constrain $\mathfrak{W}^*(c)$, and so we can as well drop it: The modified $\mathfrak{W}^*(c)$ equals $\bar{\mathfrak{W}}^*(c)$.

Next we argue that Lemma E.7 continues to hold, now claimed for $\bar{\mathfrak{W}}^*$. To verify that this is the case, replace $B^d$ with $\bar{\mathfrak{W}}(\bar{c})$ throughout in Lemma E.7. This requires straightforward adaptation of algebra as $\bar{\mathfrak{W}}(\bar{c})$ is only stochastically and not deterministically bounded.

Finally, in Lemma E.3 we remove the $\rho$-constraint from $V^b_n$ and $V^I_n$ without replacement, and note that the lemma is now claimed for $\theta'_n \in \theta + \|\bar{\mathfrak{W}}(\bar{c})\|_H/\sqrt{n}B^d$. Recall that in the lemma the a.s. representation of a set $A$ is denoted by $\tilde{A}$, and with some abuse of notation let the a.s. representation of $\bar{\mathfrak{W}}$ be denoted $\widetilde{\bar{\mathfrak{W}}}$. Now we compare $\tilde{V}^b_n$ and $\tilde{V}^I_n$ with $\widetilde{\bar{\mathfrak{W}}}$. To ensure that $\lambda$ is uniformly stochastically bounded in expressions like (E.95), we verify that the modified $\tilde{V}^b_n$ and $\tilde{V}^I_n$ inherit the property in Assumption D.1-(II). To see this, fix any unit vector $t \perp p$ and notice that any $t = \lambda/\|\lambda\|$ for $\lambda \in \widetilde{\bar{\mathfrak{W}}}(c)$ or for $\lambda \in \tilde{V}^b_n(\theta'_n, c)$ or for $\lambda \in \tilde{V}^I_n(\theta'_n, c)$, $0 < c \leqslant \bar{c}$, satisfies this condition. By Assumption D.1-(II) and the Cauchy-Schwarz inequality, $\max_{\lambda \in \widetilde{\bar{\mathfrak{W}}}(c)} t'\lambda = O(1)$ for any $c \leqslant \bar{c}$. Since the value of this program is necessarily attained by a basic solution whose associated gradients span $t$, it must be the case that such solution is itself $O(1)$. Formally, let $C$ be the index set characterizing the solution, $\mathbb{Z}^C_i$ be the vector of realizations $\mathbb{Z}^j_i$ corresponding to $j \in C$, and $K^C(\theta'_n)$ the matrix that stacks the corresponding gradients; then $(K^C(\theta'_n))^{-1}(\bar{c}\mathbf{1} - \mathbb{Z}^C_i) = O(1)$. By Lemma E.7 and the fact that $\hat{D}_n(\theta'_n) \xrightarrow{P} D$ by Assumption 4.4, we then also have that $(\hat{K}^C(\theta'_n))^{-1}(\bar{c}\mathbf{1} - \mathbb{G}^b_{n,j}) = O_{\mathcal{P}}(1)$, and so for $c \leqslant \bar{c}$, $V^b$ is bounded in this same direction. It follows that, by similar reasoning to the preceding paragraph, the comparison between $V^I_n(\theta'_n, c)$ and $\bar{\mathfrak{W}}(c)$ in Lemma E.3 goes through. □

## D.3 Proof of Theorems 4.3 and 4.4

### D.3.1 Assumptions in Pakes, Porter, Ho, and Ishii (2011), Chernozhukov, Hong, and Tamer (2007), and Bugni, Canay, and Shi (2017) That Allow for Simplifications of the Method

We analyze calibrated projection under assumptions that are more stringent than for Theorem 4.1. The reward is considerable computational simplification and, in some cases, removal of a tuning parameter. The additional assumptions have been used in the related literature. Their logical relation to each other and to explicit constraint qualifications is further analyzed in Kaido, Molinari, and Stoye (2017). For our purposes in this paper, we just state without proof that, given Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5, all assumptions below, including the minorant assumptions attributed to other papers, are implied by assumptions in Pakes, Porter, Ho, and Ishii (2011); hence, all results reported below apply under the Pakes, Porter, Ho, and Ishii (2011) assumptions.[47]

For $\theta \in \partial\Theta_I(P)$, denote by $\mathcal{J}(P,\theta)$ the set of inequalities $j$ s.t. $E_P\left(m_j(X_i,\theta)\right) = 0$. Denote by $\mathcal{N}(P,\theta)$ the positive span of $(D_{P,j})_{j\in\mathcal{J}(P,\theta)}$ and by $\mathcal{T}(P,\theta) = \{t : D'_{P,j}t \leqslant 0, j \in \mathcal{J}(P,\theta)\}$ the corresponding dual cone. (These are the normal and tangent cones of $\Theta_I(P)$ at $\theta$.) For a given $p \in \mathbb{R}^d : \|p\| = 1$, let $s(p,\Theta_I(P)) = \max_{\theta\in\Theta_I(P)} p'\theta$ and $H(p,\Theta_I(P)) \equiv \arg\max_{\theta\in\Theta_I(P)} p'\theta$.

Assumption D.2 (A weakening of Assumption 4(a) in Pakes, Porter, Ho, and Ishii (2011)): *There is a class of DGPs $\mathcal{Q} \subset \mathcal{P}$ such that any $P \in \mathcal{Q}$ satisfies the following conditions:*

1. *There exists a (universal) $\varepsilon_D > 0$ s.t.*

$$\min_{\theta\in H(p,\Theta_I(P))} \min_{\|t\|=1} \max_{\substack{j\in\{1,\ldots,J\}:\\ E_P(m_j(X_i,\theta)/\sigma_j(\theta))>-\varepsilon_D}} t'D_{P,j}(\theta) < -\varepsilon_D.$$

2. *There exists a (universal) $\varepsilon_D > 0$ s.t.*

$$\max_{\theta\in H(p,\Theta_I(P))} \min_{\|t\|=1} \max_{\substack{j\in\{1,\ldots,J\}:\\ E_P(m_j(X_i,\theta)/\sigma_j(\theta))>-\varepsilon_D}} t'D_{P,j}(\theta) < -\varepsilon_D.$$

There are two layers to these assumptions. First, they say that from some support point (part (1)) or all support points (part (2)), there are directions that point uniformly inside $\Theta_I(P)$ in the sense of all moment inequalities decreasing in value. The obvious counterexample would be an extremely pointy corner (a "spike").

In addition, the assumptions apply to "tightened" tangent cones that use all inequalities which are almost binding, where "almost" is operationalized with the small but positive constant $\varepsilon_D$. Together with smoothness of moment conditions, this implies that, by moving a small (but boundedly nonzero) distance in the direction of steepest descent from the support point, one can find a point $\theta$ at which $\max_j E_P\left(m_j(X_i,\theta)/\sigma_j(\theta)\right)$ is boundedly negative. This implies that the sample analog of $\Theta_I(P)$ is nonempty with probability approaching 1 (the proof in Appendix D.3.2 includes a formal version of this argument). In particular, it implies that a vestige of the "degeneracy" assumption in Chernozhukov, Hong, and Tamer (2007) is imposed. Some invocations of the assumption strictly speaking only use one of the two features (again, see Kaido, Molinari, and Stoye (2017) for details), but we do not disentangle them here. Note, however, that the second implication renders the assumption implausible whenever the sample analog of $\Theta_I(P)$ is empty, an empirically frequent occurrence.

[47] Our own assumptions meaningfully exceed those of Pakes, Porter, Ho, and Ishii (2011) only through Assumption 4.3. The absence of such an assumption in Pakes, Porter, Ho, and Ishii (2011) is actually an oversight, and ours or a similar assumption must be added for their Theorem 2 to hold.

Next, consider:

ASSUMPTION D.3 (Linear Minorant – Chernozhukov, Hong, and Tamer (2007) display (4.5)): *There exist universal constants $C, \delta > 0$ and a class of DGPs $\mathcal{Q} \subset \mathcal{P}$ such that for each $P \in \mathcal{Q}$,*

$$\max_{j=1,\ldots,J} E_P\left(m_j(X_i,\theta)/\sigma_j(\theta)\right) \geqslant C \min\left\{\delta, d\left(\theta, \Theta_I(P)\right)\right\}.$$

ASSUMPTION D.4 (Linear Minorant Along Support Plane – Bugni, Canay, and Shi (2017) Assumption A3(a)): *There exist universal constants $C, \delta > 0$ and a class of DGPs $\mathcal{Q} \subset \mathcal{P}$ such that for each $P \in \mathcal{Q}$ and for each $q \in \{p, -p\}$,*

$$\max_{j=1,\ldots,J} E_P\left(m_j(X_i,\theta)/\sigma_j(\theta)\right) \geqslant C \min\left\{\delta, d\left(\theta, H(q, \Theta_I(P))\right)\right\}$$

*for all $\theta$ with $q'\theta = s(q, \Theta_I(P))$.*

These assumptions are lifted from the cited papers. In the original papers, they are polynomial minorant conditions: The minima are raised to some power $\chi$. However, for our setting and criterion function, the special case $\chi = 1$ applies. Note also that Assumption D.4 is closely analogous to Assumption D.3 but imposes the minorant condition on the "null restricted model" in which the parameter space is restricted to the true supporting hyperplane of $\Theta_I(P)$. It is easy to see that the assumptions are logically independent.

A further strengthening of assumptions is:

ASSUMPTION D.5 (A Weakening of Assumption 3 in Pakes, Porter, Ho, and Ishii (2011)): *There exists a universal constant $\bar{\delta} > 0$ and a class of DGPs $\mathcal{Q} \subset \mathcal{P}$ such that for any $P \in \mathcal{Q}$ and for each $q \in \{p, -p\}$ and any $\theta \in H(q, \Theta_I(P))$, $\mathcal{T}(\theta) \subseteq \{t : q't/\|t\| \leqslant -\bar{\delta}\}$.*

Note the implication that $\mathcal{T}(P, \theta)$ is uniformly pointy. The assumption is weaker than in Pakes, Porter, Ho, and Ishii (2011) because they also assume $\Theta_I(P) \subseteq \mathcal{T}(\theta)$ and separately (although it is also an implication) that $H(p, \Theta_I(P))$ is a singleton.

Our final assumption gives a further strengthening by requiring the support set in direction of projection to be a singleton:

ASSUMPTION D.6 (Assumption 1 in Pakes, Porter, Ho, and Ishii (2011)): *There is a class of DGPs $\mathcal{Q} \subset \mathcal{P}$ such that for any $P \in \mathcal{Q}$ and $q \in \{p, -p\}$, $H(q, \Theta_I(P))$ is a singleton. (Its sole element will be denoted $\theta_q^*$ below.)*

### D.3.2 Proof of Theorem 4.3: Simplifications for Calibrated Projection

**Part I**

Let $\theta_p^*$ attain the outer minimum in Assumption D.2-1, let $t^*$ attain the inner minimum given $\theta_p^*$, and consider any $\eta \leqslant \varepsilon_D/2M$, where $\varepsilon_D$ is from Assumption D.2-1 and $M$ is from Assumption 4.4(ii). Then a Mean Value Theorem yields

$$\begin{aligned} \frac{E_P(m_j(X_i, \theta_p^* + \eta t^*))}{\sigma_{P,j}(\theta_p^* + \eta t^*)} &= \frac{E_P(m_j(X_i, \theta_p^*))}{\sigma_{P,j}(\theta_p^*)} + \eta D_{P_j}(\bar{\theta}) t^* \\ &\leqslant 0 + \eta(\eta M - \varepsilon_D) \\ \implies \max_j \frac{E_P(m_j(X_i, \theta_p^* + \eta t^*))}{\sigma_{P,j}(\theta_p^* + \eta t^*)} &\leqslant -\eta\varepsilon_D/2. \end{aligned} \tag{D.35}$$

This will be used later but also implies

$$\begin{aligned} & \max_j \frac{E_P(m_j(X_i, \theta_p^* + t^*\varepsilon_D/2M))}{\sigma_{P,j}(\theta_p^* + t^*\varepsilon_D/2M)} \leqslant -\varepsilon_D^2/4M < 0 \qquad \text{(D.36)} \\ \Longrightarrow \quad & P\left(\max_j \frac{\bar{m}_j(X_i, \theta_p^* + t^*\varepsilon_D/2M)}{\hat{\sigma}_j(\theta_p^* + t^*\varepsilon_D/2M))} < 0\right) \to 1 \\ \Longrightarrow \quad & P(\theta_p^* + t^*\varepsilon_D/2M \in CI_n) \to 1 \end{aligned}$$

uniformly in $\mathcal{Q}$. Hence, noncoverage risk for any $\gamma \in \left[-s(-p, \Theta_I(P)), p'(\theta_p^* + t^*\varepsilon_D/2M)\right]$ is entirely driven by the possibility that $CI_n$ is too high, and conversely for $\gamma \in \left[p'(\theta_p^* + t^*\varepsilon_D/2M, s(p, \Theta_I(P)))\right]$. As these noncoverage risks are monotonic in $\gamma$, the simplification is justified. □

**Part II**

Note first that, as an immediate implication of D.36, the event that $\min_{\theta\in\Theta} \max_j |\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)|_+ = 0$, hence this value is attained on $\hat{\Theta}_I$, occurs w.p.a. 1 uniformly in $\mathcal{Q}$.

Next, we show that $\sqrt{n}(s(p, \hat{\Theta}_I) - s(p, \Theta_I(P))) = O_{\mathcal{Q}}(1)$. Define

$$\mathcal{C}(-\varepsilon) = \left\{\theta \in \Theta : \max_{j=1,\ldots,J} E_P\left(m_j(X_i, \theta)/\sigma_{P,j}(\theta)\right) \leqslant -\varepsilon\right\}.$$

Note that in this notation, $\Theta_I(P) = \mathcal{C}(0)$. By (D.36) and because $\mathcal{C}(-\varepsilon)$ is closed by assumptions on $m_j$, we have that $H(p, \mathcal{C}(-\varepsilon))$ is nonempty for $\varepsilon \in [0, \varepsilon_D^2/4M]$. Next, consider any $\eta \leqslant \varepsilon_D/2M$, then $p'(\theta_p^* + \eta t^*) \geqslant p'\theta_p^* - \eta$, which together with (D.35) implies

$$s(p, \mathcal{C}(-\eta\varepsilon_D/2)) - s(p, \Theta_I) \geqslant -\eta.$$

Set $\varepsilon = \eta\varepsilon_D/2$, then equivalently we find that for $\varepsilon \leqslant \varepsilon_D^2/4M$, $s(p, \mathcal{C}(-\varepsilon) - s(p, \Theta_I)) \geqslant -2\varepsilon/\varepsilon_D$. Next, we have that uniformly over $\theta \in \cup_{\varepsilon\in[0,\varepsilon_D^2/4M]} H(p, \mathcal{C}(-\varepsilon))$,

$$\begin{aligned} \sqrt{n}\max_j |\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)|_+ &= \max_j \left\{(1 - \eta_{n,j}(\theta))\left|\mathbb{G}_{n,j}(\theta) + \sqrt{n}E_P\left(m_j(X_i, \theta)/\sigma_{P,j}(\theta)\right)\right|_+\right\} \\ &\leqslant \sum_j (1 - \eta_{n,j}(\theta))\left|\mathbb{G}_{n,j}(\theta) + \sqrt{n}E_P\left(m_j(X_i, \theta)/\sigma_{P,j}(\theta)\right)\right|_+ \\ &\leqslant J(1 + o_{\mathcal{Q}}(1))\left|O_{\mathcal{Q}}(1) - \sqrt{n}\varepsilon\right|_+, \end{aligned}$$

so in analogy to CHT (Theorem 4.2, step 1 of proof) we find $\sqrt{n}|s(p, \hat{\Theta}_I) - s(p, \Theta_I)|_- = O_{\mathcal{Q}}(1)$. On the other hand, from Assumption D.3 we have that uniformly over $\theta \in \Theta$,

$$\begin{aligned} \sqrt{n}\max_j |\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)|_+ &= \max_j \left\{(1 - \eta_{n,j}(\theta))\left|\mathbb{G}_{n,j}(\theta) + \sqrt{n}E_P\left(m_j(X_i, \theta)/\sigma_{P,j}(\theta)\right)\right|_+\right\} \\ &\geqslant \frac{1}{J}\sum_{j=1}^J (1 - \eta_{n,j}(\theta))\left|\mathbb{G}_{n,j}(\theta) + \sqrt{n}E_P\left(m_j(X_i, \theta)/\sigma_{P,j}(\theta)\right)\right|_+ \\ &\geqslant \frac{1}{J}\sum_{j=1}^J (1 + o_{\mathcal{Q}}(1))\left|O_{\mathcal{Q}}(1) + \sqrt{n}C\min\left\{\delta, d(\theta, \Theta_I(P))\right\}\right|_+, \end{aligned}$$

hence $\sqrt{n}|s(p, \hat{\Theta}_I) - s(p, \Theta_I(P))|_+ = O_{\mathcal{Q}}(1)$.

We next argue that $d(\hat{\theta}_p, H(p, \Theta_I(P))) = O_{\mathcal{Q}}(n^{-1/2})$ (the proof for $d(\hat{\theta}_{-p}, H(-p, \Theta_I(P)))$ is identical). To do so, let $\hat{k} \equiv s(p, \Theta_I(P)) - s(p, \hat{\Theta}_I)$ and define $\tilde{\theta} = \hat{\theta}_p + \hat{k}p$, noting that $p'\tilde{\theta} = s(p, \Theta_I(P))$ by construction and so

Assumption D.4 applies to $\tilde{\theta}$. Let $\bar{\theta} \in H(p, \Theta_I(P))$ be such that $d(\tilde{\theta}, H(p, \Theta_I(P))) = \|\bar{\theta} - \tilde{\theta}\|$, then

$$d\left(\hat{\theta}_p, H(p, \Theta_I(P))\right) \leqslant d\left(\hat{\theta}_p, \tilde{\theta}\right) + d\left(\tilde{\theta}, H(p, \Theta_I(P))\right) \leqslant \left\|\hat{\theta}_p - \tilde{\theta}\right\| + \left\|\tilde{\theta} - \bar{\theta}\right\| = \left\|\tilde{\theta} - \bar{\theta}\right\| + \hat{k}.$$

We already have $\sqrt{n}\hat{k} = O_{\mathcal{Q}}(1)$, so it suffices to show $\sqrt{n}|\tilde{\theta} - \bar{\theta}| = O_{\mathcal{Q}}(1)$. Using Assumption D.4, we have

$$\begin{aligned} C \min\left\{\delta, \|\tilde{\theta} - \bar{\theta}\|\right\} &\leqslant \max_{j=1,\ldots,J} \left\{ \frac{E_P(m_j(X_i, \tilde{\theta}))}{\sigma_{P,j}(\tilde{\theta})} \right\} \\ &= \max_{j=1,\ldots,J} \left\{ \frac{E_P(m_j(X_i, \hat{\theta}_p))}{\sigma_{P,j}(\hat{\theta}_p)} + \hat{k} D_{P_j}(\check{\theta}_j) p \right\} \\ &\leqslant \max_{j=1,\ldots,J} \left\{ \frac{E_P(m_j(X_i, \hat{\theta}_p))}{\sigma_{P,j}(\hat{\theta}_p)} \right\} + \max_{j=1,\ldots,J} \left\{ \hat{k} D_{P_j}(\check{\theta}_j) p \right\}. \end{aligned}$$

Here, the equality step uses that $\tilde{\theta} = \hat{\theta}_p + \hat{k}p$ and introduces $\check{\theta}_j$, which lies componentwise between $\tilde{\theta}$ and $\hat{\theta}_p$. In the last line, the first term equals 0 w.p.a. 1 because $\hat{\theta}_p \in \hat{\Theta}_I$, and the second term is bounded by $\hat{k}\bar{M}$, hence the result. To justify Simplification 2, combine the above algebra with the following observations:

(i) For a sequence $P_n \in \mathcal{Q}$, coverage of $p'\theta$ for some $\theta \in H(p, \Theta_I(P_n))$ implies coverage of $s(p, \Theta_I(P_n))$. In the proof of Theorem 4.1, starting with display D.7, it therefore suffices to show the claim for some, possibly data dependent, sequence $\theta_n \in H(p, \Theta_I(P_n))$, and then again (in case of two-sided testing) for a sequence $\theta_n \in H(-p, \Theta_I(P_n))$.

(ii) All proofs go through if coverage is evaluated at $\theta_n$ but $D_{P,j}$ and $\mathbb{G}_{n,j}$ are estimated at some $\hat{\theta}_{n,p} = \theta_n + O_{\mathcal{Q}}(n^{-1/2})$. To give one example, Assumption 4.4 implies that $\|\hat{D}_{n,j}(\hat{\theta}_{n,p}) - D_{P_n,j}(\theta_n)\| = o_{\mathcal{Q}}(1)$.

**Part III**

This is established by showing that Assumption D.1-(II) is implied. Thus, let $\mathfrak{W}$ be as in (D.32). Because the marginals of $\mathbb{Z}$ are standard normal, for any $\eta > 0$ we have the Bonferroni bounds

$$\Pr\left(\mathfrak{W}(\bar{c}) \subseteq L_\eta\right) \geqslant 1 - \eta,$$

where

$$\begin{aligned} L_\eta &= \left\{ \lambda \in \mathbb{R}^d : p'\lambda = 0 \cap \max_j \left\{ \Phi^{-1}(\eta/J) + D_{P_j}\lambda \right\} \leqslant \bar{c} \right\} \\ &= \left\{ \lambda \in \mathbb{R}^d : p'\lambda = 0 \cap \max_j D_{P_j}\lambda \leqslant \bar{c} + \Phi^{-1}(1 - \eta/J) \right\}. \end{aligned}$$

It remains to bound $\|L_\eta\|_H = \max\{\|\lambda\| : \lambda \in L_\eta\}$. To do so, we show below that

$$p'\lambda = 0 \Rightarrow \max_j \left\{ D_{P_j}\lambda / \|\lambda\| \right\} \geqslant \underbrace{\frac{\sqrt{1 + \bar{\delta}^2} - 1}{\sqrt{1 + \bar{\delta}^2} + 1}}_{=:a} \varepsilon_D, \tag{D.37}$$

where $\bar{\delta}$ is from Assumption D.5 and $\varepsilon_D$ is from Assumption D.2. Solving (D.37) for $\|\lambda\|$ and inspecting the definition of $L_\eta$ yields

$$\max\{\|\lambda\| : \lambda \in L_\eta\} \leqslant \frac{\bar{c} + \Phi^{-1}(1 - \eta/J)}{a\varepsilon_D}$$

and therefore an $O(1)$ upper bound on $\|\mathfrak{W}(\bar{c})\|$. It remains to show (D.37). Suppose by contradiction that $\max_j \{D_j\lambda / \|\lambda\|\} < a\varepsilon_D$. Let the unit vector $t^*$ achieve the minimum from Assumption D.2-2, then $\max_j \{D_j(\lambda/\|\lambda\| + dt^*)\} <$

0 and therefore $t \equiv \lambda/\|\lambda\| + dt^* \in \mathcal{T}$. We compute

$$\frac{\lambda' t}{\|\lambda\|\,\|t\|} = \frac{\lambda'\left(\frac{\lambda}{\|\lambda\|} + at^*\right)}{\|\lambda\|\left\|\frac{\lambda}{\|\lambda\|} + at^*\right\|} = \frac{1 + a\frac{\lambda' t^*}{\|\lambda\|}}{\left\|\frac{\lambda}{\|\lambda\|} + at^*\right\|} > \frac{1-a}{1+a} = 1/\sqrt{1+\bar{\delta}^2},$$

where the inequality is strict because $\lambda \neq t^*$. We conclude that $\max_{t\in\mathcal{T}} \frac{\lambda' t}{\|\lambda\|\|t\|} > 1/\sqrt{1+\bar{\delta}^2}$. In particular, if $\hat{\lambda}$ is the projection of $\lambda$ onto $\mathcal{T}$, then $\frac{\lambda'\hat{\lambda}}{\|\lambda\|\|\hat{\lambda}\|} > 1/\sqrt{1+\bar{\delta}^2}$.[48]

However, we also have $p'\hat{\lambda}/\|\hat{\lambda}\| \leqslant -\bar{\delta}$ by Assumption D.5. It follows that $p'(\lambda - \hat{\lambda})/\|\hat{\lambda}\| \geqslant \bar{\delta}$, hence $\|\lambda - \hat{\lambda}\|^2 \geqslant \bar{\delta}^2\|\hat{\lambda}\|^2$ by Cauchy-Schwarz (recall $p$ is a unit vector). But also $\|\lambda - \hat{\lambda}\|^2 + \|\hat{\lambda}\|^2 = \|\lambda\|^2$. Simple algebra then yields $\|\hat{\lambda}\|/\|\lambda\| \leqslant 1/\sqrt{1+\bar{\delta}^2}$. But $\|\hat{\lambda}\|/\|\lambda\|$ is also the cosine of the angle formed by $\lambda$ and $\hat{\lambda}$. Thus, $\frac{\lambda'\hat{\lambda}}{\|\lambda\|\|\hat{\lambda}\|} \leqslant 1/\sqrt{1+\bar{\delta}^2}$, a contradiction.[49]

### D.3.3 Proof of Theorem 4.4: Asymptotic Equivalence with BCS-Profiling in Well-Behaved Cases

Recall that under this Theorem's assumptions, $H(p, \Theta_I)$ is a singleton $\{\theta_p^*\}$ whose element is $\sqrt{n}$-consistently estimated by a sample analog $\hat{\theta}_p$. We restrict attention to $s \geqslant p'(\theta_p^* + t^*\varepsilon_D/2M)$, where terms are as in the proof of Theorem 4.3-(I). The proof for $s < p'(\theta_p^* + t^*\varepsilon_D/2M)$ is analogous. Similarly to earlier proofs, consider a sequence $(P_n, s_n)$ that asymptotically minimizes the probability from the Theorem. If $\sqrt{n}(s_n - s(p, \Theta_I(P_n))) \to \infty$, then $\min_{p'\theta = s_n} T_n(\theta) \to \infty$ by arguments in the proof of Theorem 4.3-(II), and the conclusion obtains because both indicator functions vanish. Similarly, if $\sqrt{n}(s_n - s(p, \Theta_I(P_n))) \to -\infty$, then both indicator functions equal 1 with probability approaching 1 (indeed, recall the sample support function is $\sqrt{n}$-consistent). It remains to analyze the case where $\sqrt{n}(s_n - s(p, \Theta_I(P_n))) = O_{\mathcal{Q}}(1)$.

Recalling that no $\rho$-box is used, $\hat{c}_n(\hat{\theta}_p)$ is the $(1-\alpha)$ quantile of

$$\begin{aligned}
T_n^b &= \min_{p'\lambda=0} \max_j \left\{ \mathbb{G}_{n,j}^b(\hat{\theta}_p) + \kappa_n^{-1}\sqrt{n}\left|\bar{m}_{n,j}(\hat{\theta}_p)/\hat{\sigma}_{n,j}(\hat{\theta}_p)\right|_- + \hat{D}_{n,j}(\hat{\theta}_p)\lambda \right\} \\
&\overset{(1)}{=} \min_{p'\lambda=0} \max_j \left\{ \mathbb{G}_{n,j}^b(\hat{\theta}_p) + \kappa_n^{-1}\sqrt{n}E_P\left|m_j(X_i, \hat{\theta}_p)/\sigma_{P,j}(\hat{\theta}_p)\right|_- + \hat{D}_{n,j}(\hat{\theta}_p)\lambda \right\} + o_{\mathcal{Q}}(1) \\
&\overset{(2)}{=} \min_{p'\lambda=0} \max_j \left\{ \mathbb{G}_{n,j}^b(\theta_p^*) + \kappa_n^{-1}\sqrt{n}E_P\left|m_j(X_i, \theta_p^*)/\sigma_{P,j}(\theta_p^*)\right|_- + D_{n,j}(\theta_p^*)\lambda \right\} + o_{\mathcal{Q}}(1),
\end{aligned}$$

Here, (1) uses Lemma E.5-(iii). Step (2) uses that by Theorem 4.3-(III), the values of $\lambda$ solving the optimization problems are $O_{\mathcal{Q}}(1)$; by 4.3-(II), $\sqrt{n}(\hat{\theta}_p - \theta_p^*) = O_{\mathcal{Q}}(1)$; and smoothness conditions as well as consistent estimation of gradients. These jointly imply that $|\hat{D}_{n,j}(\hat{\theta}_p)\lambda - D_{n,j}(\theta_p^*)\lambda| = o_{\mathcal{Q}}(1)$ uniformly over the relevant range of $\lambda$.

To compare BCS-profiling, let $\hat{\theta}_{p,s_n}$ be the selection from $\arg\min_{p'\theta=s_n} |\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)|_+$ that solves the problem in the definition of $T_n^{DR}(s_n)$ below. Arguments very similar to Theorem 4.3-(II) imply that $\sqrt{n}(\hat{\theta}_{p,s_n} - \theta_p^*) =$

[48] Verbally, if $\lambda$ is near tangential to all constraints, it is near tangential to $\mathcal{T}$. The counterexample to this would be a "spike," which is excluded by Assumption D.2-2.

[49] Verbally, if $p'\lambda = 0$, then $\lambda$ cannot be near tangential to $\mathcal{T}$ because of the "pointy cone" assumption D.5, yielding a contradiction.

$O_{\mathcal{Q}}(1)$. We can use this, again Lemma E.5-(iii), and smoothness conditions to write

$$\begin{aligned}
T_n^{DR}(s_n) &= \min_{\theta} \max_j \left\{ \mathbb{G}_{n,j}^b(\theta) + \kappa_n^{-1}\sqrt{n}\,|\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)|_- \right\} \quad s.t. \;\; \theta \in \arg\min_{p'\theta = s_n} \max_j |\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)|_+ \\
&= \max_j \left\{ \mathbb{G}_{n,j}^b(\hat{\theta}_{p,s_n}) + \kappa_n^{-1}\sqrt{n}\big|\bar{m}_{n,j}(\hat{\theta}_{p,s_n})/\hat{\sigma}_{n,j}(\hat{\theta}_{p,s_n})\big|_- \right\} \\
&= \max_j \left\{ \mathbb{G}_{n,j}^b(\hat{\theta}_{p,s_n}) + \kappa_n^{-1}\sqrt{n}E_P\big|m_j(X_i,\hat{\theta}_{p,s_n})/\sigma_{P,j}(\hat{\theta}_{p,s_n})\big|_- \right\} + o_{\mathcal{Q}}(1) \\
&= \max_j \left\{ \mathbb{G}_{n,j}^b(\theta_p^*) + \kappa_n^{-1}\sqrt{n}E_P\left|m_j(X_i,\theta_p^*)/\sigma_{P,j}(\theta_p^*)\right|_- \right\} + o_{\mathcal{Q}}(1).
\end{aligned}$$

Next,

$$\begin{aligned}
T_n^{PR}(s_n) &= \min_{\theta\in\Theta: p'\theta = s_n} \max_j \left\{ \mathbb{G}_{n,j}^b(\theta) + \kappa_n^{-1}\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta) \right\} \\
&\overset{(1)}{=} \min_{\theta\in\Theta: p'\theta = s_n} \max_j \left\{ \mathbb{G}_{n,j}^b(\theta) + \kappa_n^{-1}\sqrt{n}E_P\left(m_j(X_i,\theta)/\sigma_{P,j}(\theta)\right)\right\} + o_{\mathcal{Q}}(1) \\
&\overset{(2)}{=} \min_{\theta\in\Theta: p'\theta = s(p,\Theta_I)} \max_j \left\{ \mathbb{G}_{n,j}^b(\theta) + \kappa_n^{-1}\sqrt{n}E_P\left(m_j(X_i,\theta)/\sigma_{P,j}(\theta)\right)\right\} + o_{\mathcal{Q}}(1) \\
&\overset{(3)}{=} \min_{p'\lambda = 0} \max_j \left\{ \mathbb{G}_{n,j}^b(\theta_p^* + \lambda\kappa_n n^{-1/2}) + \kappa_n^{-1}\sqrt{n}E_P\big(m_j(X_i,\theta_p^* + \lambda\kappa_n n^{-1/2})/\sigma_{P,j}(\theta_p^* + \lambda\kappa_n n^{-1/2})\big)\right\} + o_{\mathcal{Q}}(1) \\
&\overset{(4)}{=} \min_{p'\lambda = 0} \max_j \left\{ \mathbb{G}_{n,j}^b(\theta_p^*) + \kappa_n^{-1}\sqrt{n}E_P\left(m_j(X_i,\theta_p^*)/\sigma_{P,j}(\theta_p^*)\right) + D_{P,j}(\theta_p^*)\lambda\right\} + o_{\mathcal{Q}}(1) \\
&\overset{(5)}{=} \min_{p'\lambda = 0} \max_j \left\{ \mathbb{G}_{n,j}^b(\theta_p^*) + \kappa_n^{-1}\sqrt{n}E_P\left|m_j(X_i,\theta_p^*)/\sigma_{P,j}(\theta_p^*)\right|_- + D_{P,j}(\theta_p^*)\lambda\right\} + o_{\mathcal{Q}}(1)
\end{aligned}$$

Here, (1) uses Lemma E.5-(iii). The first crucial step is (2), which uses that the distance between the hyperplanes $\{p'\theta = s_n\}$ and $\{p'\theta = s(\Theta_I, p)\}$ is of order $O_{\mathcal{Q}}(n^{-1/2})$, together with smoothness conditions. Step (3) reparameterizes $\theta = \theta_p^* + \lambda\kappa_n n^{-1/2}$. Crucially, BCS prove that the $\lambda$ solving the problem is $O_{\mathcal{Q}}(1)$. This means the problem can be uniformly linearized, justifying step (4). Step (4) also observes cancellation of factors multiplying $D_{P,j}(\theta_p^*)\lambda$. Step (5) uses that $\theta_p^* \in \Theta_I$. Finally, Assumption 4.3 ensures that the true distribution of $T_n$, as well as the above approximations, are of order $O_{\mathcal{Q}}(1)$. We conclude that $T_n^{PR}(s_n)$ asymptotically agrees with, and $T_n^{DR}(s_n)$ asymptotically dominates, $T_n^b$. □

# Appendix E Auxiliary Lemmas

## E.1 Lemmas Used to Prove Theorems 4.1 and 4.2

Throughout this Appendix, we let $(P_n, \theta_n) \in \{(P,\theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ be a subsequence as defined in the proof of Theorem 4.1. That is, along $(P_n, \theta_n)$, one has

$$\kappa_n^{-1}\sqrt{n}\gamma_{1,P_n,j}(\theta_n) \to \pi_{1j} \in \mathbb{R}_{[-\infty]}, \;\; j = 1,\ldots,J, \tag{E.1}$$

$$\Omega_{P_n} \xrightarrow{u} \Omega, \tag{E.2}$$

$$D_{P_n}(\theta_n) \to D. \tag{E.3}$$

Fix $c \geqslant 0$. For each $\lambda \in \mathbb{R}^d$ and $\theta \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$, let

$$\mathfrak{w}_j(\lambda) \equiv \mathbb{Z}_j + \rho D_j\lambda + \pi_{1,j}^*, \tag{E.4}$$

where $\pi^*_{1,j}$ is defined in (D.5) and we used Lemma E.5. Under Assumption 4.3-(II) if

$$\pi^*_{1,j} = 0 = \pi^*_{1,j+R_1}, \tag{E.5}$$

we replace the constraints

$$\mathbb{Z}_j + \rho D_j \lambda \leqslant c, \tag{E.6}$$
$$\mathbb{Z}_{j+R_1} + \rho D_{j+R_1} \lambda \leqslant c, \tag{E.7}$$

with

$$\mu_j(\theta)\{\mathbb{Z}_j + \rho D_j \lambda\} - \mu_{j+R_1}(\theta)\{\mathbb{Z}_{j+R_1} + \rho D_{j+R_1} \lambda\} \leqslant c, \tag{E.8}$$
$$-\mu_j(\theta)\{\mathbb{Z}_j + \rho D_j \lambda\} + \mu_{j+R_1}(\theta)\{\mathbb{Z}_{j+R_1} + \rho D_{j+R_1} \lambda\} \leqslant c, \tag{E.9}$$

where

$$\mu_j(\theta) = \begin{cases} 1 & \text{if } \gamma_{1,P_n,j}(\theta) = 0 = \gamma_{1,P_n,j+R_1}(\theta), \\ \frac{\gamma_{1,P_n,j+R_1}(\theta)}{\gamma_{1,P_n,j+R_1}(\theta)+\gamma_{1,P_n,j}(\theta)} & \text{otherwise,} \end{cases} \tag{E.10}$$

$$\mu_{j+R_1}(\theta) = \begin{cases} 0 & \text{if } \gamma_{1,P_n,j}(\theta) = 0 = \gamma_{1,P_n,j+R_1}(\theta), \\ \frac{\gamma_{1,P_n,j}(\theta)}{\gamma_{1,P_n,j+R_1}(\theta)+\gamma_{1,P_n,j}(\theta)} & \text{otherwise,} \end{cases} \tag{E.11}$$

When Assumption 4.3-(II) is invoked with hard-threshold GMS, replace constraints $j$ and $j+R_1$ in the definition of $\Lambda^b_n(\theta'_n, \rho, c), \theta'_n \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$ in equation (3.1) as described on p.14 of the paper; when it is invoked with a GMS function $\varphi$ that is smooth in its argument, replace them, respectively, with

$$\hat{\mu}_{n,j}(\theta'_n)\{\mathbb{G}^b_{n,j}(\theta'_n) + \hat{D}_{n,j}(\theta'_n)\lambda\} - \hat{\mu}_{n,j+R_1}(\theta'_n)\{\mathbb{G}^b_{n,j+R_1}(\theta'_n) + \hat{D}_{n,j+R_1}(\theta'_n)\lambda\} + \varphi_j(\hat{\xi}_{n,j}(\theta'_n)) \leqslant c, \tag{E.12}$$
$$-\hat{\mu}_{n,j}(\theta'_n)\{\mathbb{G}^b_{n,j}(\theta'_n) + \hat{D}_{n,j}(\theta'_n)\lambda\} + \hat{\mu}_{n,j+R_1}(\theta'_n)\{\mathbb{G}^b_{n,j+R_1}(\theta'_n) + \hat{D}_{n,j+R_1}(\theta'_n)\lambda\} + \varphi_{j+R_1}(\hat{\xi}_{n,j+R_1}(\theta'_n)) \leqslant c, \tag{E.13}$$

where

$$\hat{\mu}_{n,j+R_1}(\theta'_n) = \min\left\{\max\left(0, \frac{\frac{\bar{m}_{n,j}(\theta'_n)}{\hat{\sigma}_{n,j}(\theta'_n)}}{\frac{\bar{m}_{n,j+R_1}(\theta'_n)}{\hat{\sigma}_{n,j+R_1}(\theta'_n)} + \frac{\bar{m}_{n,j}(\theta'_n)}{\hat{\sigma}_{n,j}(\theta'_n)}}\right), 1\right\}, \tag{E.14}$$
$$\hat{\mu}_{n,j}(\theta'_n) = 1 - \hat{\mu}_{n,j+R_1}(\theta'_n). \tag{E.15}$$

Let $\mathfrak{B}^d_\rho = \lim_{n\to\infty} B^d_{n,\rho}$. Let the intersection of $\{\lambda \in \mathfrak{B}^d_\rho : \ p'\lambda = 0\}$ with the level set associated with the so defined function $\mathfrak{w}_j(\lambda)$ be

$$\mathfrak{W}(c) \equiv \big\{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c, \ \forall j = 1, \ldots, J\big\}. \tag{E.16}$$

Due to the substitutions in equations (E.6)-(E.9), the paired inequalities (i.e., inequalities for which (E.5) holds under Assumption 4.3-(II)) 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

$$\mathfrak{W}^\delta(c) \equiv \big\{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c + \delta, \ \forall j = 1, \ldots, J_1, \\ \cap \mathfrak{w}_j(\lambda) \leqslant c, \ \forall j = J_1 + 1, \ldots, J\big\}. \tag{E.17}$$

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

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

Given a square matrix $A$, we let $\text{eig}(A)$ denote its smallest eigenvalue. In all Lemmas below, we assume $\alpha < 1/2$.

LEMMA E.1: *Let Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. Let $\{P_n, \theta_n\}$ be a sequence such that $P_n \in \mathcal{P}$ and $\theta_n \in \Theta_I(P_n)$ for all $n$ and $\kappa_n^{-1}\sqrt{n}\gamma_{1,P_n,j}(\theta_n) \to \pi_{1j} \in \mathbb{R}_{[-\infty]}$, $j = 1, \ldots, J$, $\Omega_{P_n} \xrightarrow{u} \Omega$, and $D_{P_n}(\theta_n) \to D$. Then,*

$$\liminf_{n\to\infty} P_n\left(U_n(\theta_n, \hat{c}_{n,\rho}) \neq \varnothing\right) \geqslant 1 - \alpha. \tag{E.19}$$

*Proof.* We consider a subsequence along which $\liminf_{n\to\infty} P_n(U_n(\theta_n, \hat{c}_{n,\rho} \neq \varnothing)$ is achieved as a limit. For notational simplicity, we use $\{n\}$ for this subsequence below.

Below, we construct a sequence of critical values such that

$$\hat{c}_n(\theta_n') \geqslant c_n^I(\theta_n') + o_{P_n}(1), \tag{E.20}$$

and $c_n^I(\theta_n') \xrightarrow{P_n} c_{\pi*}$ for any $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$. The construction is as follows. When $c_{\pi*} = 0$, let $c_n^I(\theta_n') = 0$ for all $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$, and hence $c_n^I(\theta_n') \xrightarrow{P_n} c_{\pi*}$. If $c_{\pi*} > 0$, let $c_n^I(\theta_n) \equiv \inf\{c \in \mathbb{R}_+ : P_n^*(V_n^I(\theta_n, c)) \geqslant 1 - \alpha\}$, where $V_n^I$ is defined as in Lemma E.3. By Lemma E.3 (iii), this critical value sequence satisfies (E.20) with probability approaching 1. Further, by Lemma E.3 (ii), $c_n^I(\theta_n') \xrightarrow{P_n} c_{\pi*}$ for any $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$.

For each $\theta \in \Theta$, let

$$c_{n,\rho}^I(\theta) \equiv \inf_{\lambda \in B_{n,\rho}^d} c_n^I(\theta + \frac{\lambda\rho}{\sqrt{n}}). \tag{E.21}$$

Since the $o_{P_n}(1)$ term in (E.20) does not affect the argument below, we redefine $c_{n,\rho}^I(\theta_n)$ as $c_{n,\rho}^I(\theta_n) + o_{P_n}(1)$. By (E.20) and simple addition and subtraction,

$$\begin{aligned} P_n\Big(U_n(\theta_n, \hat{c}_{n,\rho}(\theta_n)) \neq \varnothing\Big) &\geqslant P_n\Big(U_n(\theta_n, c_{n,\rho}^I(\theta_n)) \neq \varnothing\Big) \\ &= \Pr(\mathfrak{W}(c_{\pi*}) \neq \varnothing) + \Big[P_n\Big(U_n(\theta_n, c_{n,\rho}^I(\theta_n)) \neq \varnothing\Big) - \Pr\Big(\mathfrak{W}(c_{\pi*}) \neq \varnothing\Big)\Big]. \end{aligned} \tag{E.22}$$

As previously argued, $\mathbb{G}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) \xrightarrow{d} \mathbb{Z}$. Moreover, by Lemma E.10, $\sup_{\theta\in\Theta} \|\eta_n(\theta)\| \xrightarrow{p} 0$ uniformly in $\mathcal{P}$, and by Lemma E.3, $c_{n,\rho}^I(\theta_n) \xrightarrow{p} c_{\pi*}$. Therefore, uniformly in $\lambda \in B^d$, the sequence $\{(\mathbb{G}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \eta_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), c_{n,\rho}^I(\theta_n))\}$ satisfies

$$(\mathbb{G}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \eta_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), c_{n,\rho}^I(\theta_n)) \xrightarrow{d} (\mathbb{Z}, 0, c_{\pi*}). \tag{E.23}$$

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

For each $\lambda \in \mathbb{R}^d$, we define analogs to the quantities in (D.24) and (E.4) as

$$u^*_{n,j,\theta_n}(\lambda) \equiv \{\mathbb{G}^*_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\}(1 + \eta^*_{n,j}), \tag{E.24}$$

$$\mathfrak{w}^*_j(\lambda) \equiv \mathbb{Z}^*_j + \rho D_j \lambda + \pi^*_{1,j}. \tag{E.25}$$

where we used that by Lemma E.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 $\theta'_n \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$ and therefore $\pi^*_{1,j}$ is constant over this neighborhood, and we applied a similar replacement as described in equations (E.6)-(E.9) for the case that $\pi^*_{1,j} = 0 = \pi^*_{1,j+R_1}$. Similarly, we define analogs to the sets in (D.25) and (E.16) as

$$U^*_n(\theta_n, c^*_n) \equiv \big\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap u^*_{n,j,\theta_n}(\lambda) \leqslant c^*_n,\ \forall j = 1, \ldots, J\big\}, \tag{E.26}$$

$$\mathfrak{W}^*(c_{\pi *}) \equiv \big\{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}^*_j(\lambda) \leqslant c_{\pi *},\ \forall j = 1, \ldots, J\big\}. \tag{E.27}$$

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

$$P_n\Big(U_n(\theta_n, \hat{c}_{n,\rho}(\theta_n)) \neq \varnothing\Big) \geqslant \mathbf{P}(\mathfrak{W}^*(c_{\pi *}) \neq \varnothing) + \Big[\mathbf{P}\Big(U^*_n(\theta_n, c^*_n) \neq \varnothing\Big) - \mathbf{P}\Big(\mathfrak{W}^*(c_{\pi *}) \neq \varnothing\Big)\Big]. \tag{E.28}$$

By the definition of $c_{\pi *}$, we have $\mathbf{P}(\mathfrak{W}^*(c_{\pi *}) \neq \varnothing) \geqslant 1 - \alpha$. Therefore, we are left to show that the second term on the right hand side of (E.28) tends to 0 as $n \to \infty$.

Define

$$\mathcal{J}^* \equiv \{j = 1, \cdots, J : \pi^*_{1,j} = 0\}. \tag{E.29}$$

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

$$U^*_n(\theta_n, c^*_n) = \{\lambda \in B^d_{n,\rho} : p'\lambda = 0\}, \quad \mathfrak{W}^*(c_{\pi *}) = \{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0\}, \tag{E.30}$$

with probability 1, and hence

$$\mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) \neq \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi *}) \neq \varnothing\}\Big) = 1. \tag{E.31}$$

This in turn implies that

$$\Big|\mathbf{P}\Big(U^*_n(\theta_n, c^*_n) \neq \varnothing\}\Big) - \mathbf{P}\Big(\mathfrak{W}^*(c_{\pi *}) \neq \varnothing\}\Big)\Big| = 0, \tag{E.32}$$

where we used $|\mathbf{P}(A) - \mathbf{P}(B)| \leqslant \mathbf{P}(A \Delta B) \leqslant 1 - \mathbf{P}(A \cap B)$ for any pair of events $A$ and $B$. Hence, the term in the square brackets in (E.28) is 0.

**Case 2.** Now consider the case that $\mathcal{J}^* \neq \varnothing$. We show that the term in the square brackets in (E.28) converges to 0. To that end, note that for any events $A, B$,

$$\Big|\mathbf{P}(A \neq \varnothing) - \mathbf{P}(B \neq \varnothing)\Big| \leqslant \Big|\mathbf{P}(\{A = \varnothing\} \cap \{B \neq \varnothing\}) + \mathbf{P}(\{A \neq \varnothing\} \cap \{B = \varnothing\})\Big| \tag{E.33}$$

Hence, we aim to establish that for $A = U^*_n(\theta_n, c^*_n)$, $B = \mathfrak{W}^*(c_{\pi *})$, the right hand side of equation (E.33) converges to zero. But this is guaranteed by Lemma E.2. Therefore, the conclusion of the lemma follows. □

Lemma E.2: *Let Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. Let $(P_n, \theta_n)$ have the almost sure representations given in Lemma E.1, and let $\mathcal{J}^*$ be defined as in (E.29). Assume that $\mathcal{J}^* \neq \varnothing$. Then for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that*

$$\mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) \neq \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi *}) = \varnothing\}\Big) \leqslant \eta/2, \tag{E.34}$$

$$\mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) = \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi *}) \neq \varnothing\}\Big) \leqslant \eta/2, \tag{E.35}$$

*for all $n \geqslant N$, where the sets in the above expressions are defined in equations (E.26) and (E.27).*

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

$$\left(\mathbb{G}^*_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda + \pi^*_{1,j}\right)(1 + \eta^*_{n,j}) \leqslant c^*_n,$$
$$\mathbb{Z}^*_j + \rho D_j \lambda + \pi^*_{1,j} \leqslant c_{\pi *}$$

are satisfied with probability approaching one by similar arguments as in (D.20). Hence, we can redefine the sets of interest as

$$U^*_n(\theta_n, c^*_n) \equiv \left\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap u^*_{n,j,\theta_n}(\lambda) \leqslant c^*_n,\ \forall j \in \mathcal{J}^*\right\}, \tag{E.36}$$
$$\mathfrak{W}^*(c_{\pi *}) \equiv \left\{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}^*_j(\lambda) \leqslant c_{\pi *},\ \forall j \in \mathcal{J}^*\right\}. \tag{E.37}$$

We first show (E.34). For this, we start by defining the events

$$A_n \equiv \left\{\sup_{\lambda \in B^d} \max_{j \in \mathcal{J}^*} \left|(u^*_{n,j,\theta_n}(\lambda) - c^*_n) - (\mathfrak{w}^*_j(\lambda) - c_{\pi *})\right| \geqslant \delta\right\}. \tag{E.38}$$

By Lemma E.4, using the assumption that $\mathcal{J}^* \neq \varnothing$, for any $\eta > 0$ there exists $N \in \mathbb{N}$ such that

$$\mathbf{P}(A_n) < \eta/2,\ \forall n \geqslant N. \tag{E.39}$$

Define the sets of $\lambda$s, $U^{*,+\delta}_n$ and $\mathfrak{W}^{*,+\delta}$ by relaxing the constraints shaping $U^*_n$ and $\mathfrak{W}^*$ by $\delta$:

$$U^{*,+\delta}_n(\theta_n, c) \equiv \{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap u^*_{n,j,\theta_n}(\lambda) \leqslant c + \delta,\ j \in \mathcal{J}^*\}, \tag{E.40}$$
$$\mathfrak{W}^{*,+\delta}(c) \equiv \{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}^*_j(\lambda) \leqslant c + \delta,\ j \in \mathcal{J}^*\}. \tag{E.41}$$

Compared to the set in equation (E.17), here we replace $u^*_{n,j,\theta_n}(\lambda)$ for $u_{n,j,\theta_n}(\lambda)$ and $\mathfrak{w}^*_j(\lambda)$ for $\mathfrak{w}_j(\lambda)$, 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 $L_n \equiv \{U^*_n(\theta_n, c^*_n) \subset \mathfrak{W}^{*,+\delta}(c_{\pi *})\}$ and note that $A^c_n \subseteq L_n$.

We may then bound the left hand side of (E.34) as

$$\mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) \neq \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi *}) = \varnothing\}\Big) \leqslant \mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) \neq \varnothing\} \cap \{\mathfrak{W}^{*,+\delta}(c_{\pi *}) = \varnothing\}\Big)$$
$$+ \mathbf{P}\Big(\{\mathfrak{W}^{*,+\delta}(c_{\pi *}) \neq \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi *}) = \varnothing\}\Big), \tag{E.42}$$

where we used $P(A \cap B) \leqslant P(A \cap C) + P(B \cap C^c)$ for any events $A, B$, and $C$. The first term on the right hand side of (E.42) can further be bounded as

$$\mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) \neq \varnothing\} \cap \{\mathfrak{W}^{*,+\delta}(c_{\pi *}) = \varnothing\}\Big) \leqslant \mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) \nsubseteq \mathfrak{W}^{*,+\delta}(c_{\pi *})\}\Big)$$
$$= \mathbf{P}(L^c_n) \leqslant \mathbf{P}(A_n) < \eta/2,\ \forall n \geqslant N\ , \tag{E.43}$$

where the penultimate inequality follows from $A^c_n \subseteq L_n$ as argued above, and the last inequality follows from (E.39). For the second term on the left hand side of (E.42), by Lemma E.6, there exists $N' \in \mathbb{N}$ such that

$$\mathbf{P}\Big(\{\mathfrak{W}^{*,+\delta}(c_{\pi *}) \neq \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi *}) = \varnothing\}\Big) \leqslant \eta/2,\ \forall n \geqslant N'. \tag{E.44}$$

Hence, (E.34) follows from (E.42), (E.43), and (E.44).

To establish (E.35), we distinguish three cases.

**Case 1.** Suppose first that $J_2 = 0$ (recalling that under Assumption 4.3-(II) this means that there is no $j = 1, \ldots, R_1$ such that $\pi^*_{1,j} = 0 = \pi^*_{1,j+R_1}$), and hence one has only moment inequalities. In this case, by (E.36) and (E.37), one may write

$$U^*_n(\theta_n, c) \equiv \left\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap u^*_{n,j,\theta_n}(\lambda) \leqslant c,\ j \in \mathcal{J}^*\right\}, \tag{E.45}$$

$$\mathfrak{W}^{*,-\delta}(c) \equiv \left\{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}^*_j(\lambda) \leqslant c - \delta,\ j \in \mathcal{J}^*\right\}, \tag{E.46}$$

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

$$R_{2n} \equiv \{\mathfrak{W}^{*,-\delta}(c_{\pi *}) \subset U^*_n(\theta_n, c^*_n)\}, \tag{E.47}$$

and note that $A^c_n \subseteq R_{2n}$. The result in equation (E.35) then follows by Lemma E.6 using again similar steps to (E.42)-(E.44).

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

$$\mathfrak{W}^{*,-\delta}(c_{\pi *}) \equiv \{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}^*_j(\lambda) \leqslant c - \delta,\ j \in \mathcal{J}^*\}, \tag{E.48}$$

that is, the same set as in (E.133) with $\mathfrak{w}^*_j(\lambda)$ replacing $\mathfrak{w}_j(\lambda)$ and defining the set using only inequalities in $\mathcal{J}^*$. Note that, by Lemma E.8, there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$ $c^I_n(\theta)$ 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 $c_{\pi *}$ 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 $\delta < \underline{c}$, $\mathfrak{W}^{*,-\delta}(c_{\pi *})$ is nonempty with probability approaching 1 by Lemma E.6, and hence its superset $\mathfrak{W}^*(c_{\pi *})$ is also non-empty with probability approaching 1. However, note that $A^c_n \subseteq R_{2n}$, where $R_{2n}$ is in (E.47) now defined using the tightened constraint set $\mathfrak{W}^{*,-\delta}(c_{\pi *})$ being defined as in (E.48), and therefore the same argument as in the previous case applies.

**Case 3.** Finally, suppose that $1 \leqslant J_2 < d$. Recall that, with probability 1 (under $\mathbf{P}$),

$$c_{\pi *} = \lim_{n \to \infty} c^*_n, \tag{E.49}$$

and note that by construction $c_{\pi *} \geqslant 0$. Consider first the case that $c_{\pi *} > 0$. Then, by taking $\delta < c_{\pi *}$, the argument in Case 2 applies.

Next consider the case that $c_{\pi *} = 0$. Observe that

$$\begin{aligned}\mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) = \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi *}) \neq \varnothing\}\Big) \leqslant\ & \mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) = \varnothing\} \cap \{\mathfrak{W}^{*,-\delta}(0) \neq \varnothing\}\Big) \\ & + \mathbf{P}\Big(\{\mathfrak{W}^{*,-\delta}(0) = \varnothing\} \cap \{\mathfrak{W}^*(0) \neq \varnothing\}\Big),\end{aligned} \tag{E.50}$$

with $\mathfrak{W}^{*,-\delta}(0)$ defined as in (E.17) with $c = 0$ and with $\mathfrak{w}^*_j(\lambda)$ replacing $\mathfrak{w}_j(\lambda)$. By Lemma E.6, for any $\eta > 0$ there exists $\delta > 0$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\Big(\{\mathfrak{W}^{*,-\delta}(0) = \varnothing\} \cap \{\mathfrak{W}^*(0) \neq \varnothing\}\Big) < \eta/3\ \forall n \geqslant N. \tag{E.51}$$

Therefore, the second term on the right hand side of (E.50) can be made arbitrarily small.

We now consider the first term on the right hand side of (E.50). Let $g$ be a $J+2d+2$ vector with

$$g_j = \begin{cases} -\mathbb{Z}_j, & j \in \mathcal{J}^*, \\ 0, & j \in \{1, \cdots, J\} \backslash \mathcal{J}^*, \\ 1, & j = J+1, \ldots, J+2d, \\ 0, & j = J+2d+1, J+2d+2, \end{cases} \tag{E.52}$$

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 (E.36)-(E.37).

For a given set $C \subset \{1, \ldots, J+2d+2\}$, let the vector $g^C$ collect the entries of $g^C$ corresponding to indices in $C$. Let

$$K \equiv \begin{bmatrix} [\rho D_j]_{j=1}^{J_1+J_2} \\ [-\rho D_{j-J_2}]_{j=J_1+J_2+1}^{J} \\ I_d \\ -I_d \\ p' \\ -p' \end{bmatrix}. \tag{E.53}$$

Let the matrix $K^C$ collect the rows of $K$ corresponding to indices in $C$.

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 $\mathcal{C}$ equal the first element of $\widetilde{\mathcal{C}}$ s.t. $\det K^C \neq 0$ and $\lambda^C = (K^C)^{-1} g^C \in \mathfrak{W}^{*,-\delta}(0)$ if such an element exists; else, let $\mathcal{C} = \{J+1, ..., J+d\}$ and $\lambda^C = \mathbf{1}_d$, where $\mathbf{1}_d$ denotes a $d$ vector with each entry equal to 1. Recall that $\mathfrak{W}^{*,-\delta}(0)$ 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 $\mathfrak{W}^{*,-\delta}(0) \neq \varnothing$, then $\mathfrak{W}^{*,-\delta}(0)$ has extreme points, each of which is characterized as the intersection of $d$ (not necessarily unique) linearly independent constraints interpreted as equalities. Therefore, $\mathfrak{W}^{*,-\delta}(0) \neq \varnothing$ implies that $\lambda^{\mathcal{C}} \in \mathfrak{W}^{*,-\delta}(0)$ and therefore also that $\mathcal{C} \subset \mathcal{J}^* \cup \{J+1, \ldots, J+2d+2\}$. Note that the associated random vector $\lambda^{\mathcal{C}}$ is a measurable selection of a random closed set that equals $\mathfrak{W}^{*,-\delta}(0)$ if $\mathfrak{W}^{*,-\delta}(0) \neq \varnothing$ and equals $\mathfrak{B}^d_\rho$ otherwise, see, e.g., Molchanov (2005, Definition 1.2.2).

Lemma E.7 establishes that for any $\eta > 0$, there exist $\varepsilon_\eta > 0$ and $N$ s.t. $n \geqslant N$ implies

$$\mathbf{P}\left(\mathfrak{W}^{*,-\delta}(0) \neq \varnothing, \left|\det K^{\mathcal{C}}\right| \leqslant \varepsilon_\eta\right) \leqslant \eta, \tag{E.54}$$

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

$$\mathbf{P}\left(\left|\det\left(K^{\mathcal{C}}\right)^{-1}\right| \leqslant M\right) \geqslant 1-\eta, \ \forall n \geqslant N. \tag{E.55}$$

Let $g_n$ be a $J+2d+2$ vector with

$$g_{n,j}(\theta + \lambda/\sqrt{n}) \equiv \begin{cases} c^*_n/(1+\eta^*_{n,j}) - \mathbb{G}^*_{n,j}(\theta + \frac{\lambda\rho}{\sqrt{n}}) & \text{if } j \in \mathcal{J}^*, \\ 0, & \text{if } j \in \{1, \cdots, J\} \backslash \mathcal{J}^*, \\ 1, & \text{if } j = J+1, \ldots, J+2d, \\ 0, & \text{if } j = J+2d+1, J+2d+2, \end{cases} \tag{E.56}$$

using again that $\pi^*_{1,j} = 0$ for $j \in \mathcal{J}^*$. For each $P \in \mathcal{P}$, let

$$K_P(\theta, \rho) \equiv \begin{bmatrix} [\rho D_{P,j}(\theta)]_{j=1}^{J_1+J_2} \\ [-\rho D_{P,j-J_2}(\theta)]_{j=J_1+J_2+1}^{J} \\ I_d \\ -I_d \\ p' \\ -p' \end{bmatrix}. \tag{E.57}$$

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

$$\phi_n(\lambda) \equiv \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\right)^{-1} g^{\mathcal{C}}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \tag{E.58}$$

where the notation $\bar{\theta}(\theta_n, \lambda)$ emphasizes that $\bar{\theta}$ depends on $\theta_n$ and $\lambda$ because it lies component-wise between $\theta_n$ and $\theta_n + \frac{\lambda\rho}{\sqrt{n}}$. We show that $\phi_n$ is a contraction mapping and hence has a fixed point.

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

$$\begin{aligned} \|\phi_n(\lambda) - \phi_n(\lambda')\| &= \left\| \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\right)^{-1} g^{\mathcal{C}}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) - \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda'), \rho)\right)^{-1} g^{\mathcal{C}}_n(\theta_n + \frac{\lambda'\rho}{\sqrt{n}}) \right\| \\ &\leqslant \left\| \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\right)^{-1} \right\|_2 \left\| g^{\mathcal{C}}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) - g^{\mathcal{C}}_n(\theta_n + \frac{\lambda'\rho}{\sqrt{n}}) \right\| \\ &\quad + \left\| \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\right)^{-1} - \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda'), \rho)\right)^{-1} \right\|_2 \left\| g^{\mathcal{C}}_n(\theta_n + \frac{\lambda'\rho}{\sqrt{n}}) \right\|, \end{aligned} \tag{E.59}$$

where $\|\cdot\|_2$ denotes the spectral norm (induced by the Euclidean norm).

By Assumption 4.5 (ii), for any $\eta > 0$, $k > 0$, there is $N \in \mathbb{N}$ such that

$$\begin{aligned} &\mathbf{P}\left( \left\| g^{\mathcal{C}}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) - g^{\mathcal{C}}_n(\theta_n + \frac{\lambda'\rho}{\sqrt{n}}) \right\| \leqslant k\|\lambda - \lambda'\| \right) \\ &\qquad = \mathbf{P}\left( \|\mathbb{G}^{*,\mathcal{C}}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) - \mathbb{G}^{*,\mathcal{C}}_n(\theta_n + \frac{\lambda'\rho}{\sqrt{n}})\| \leqslant k\|\lambda - \lambda'\| \right) \geqslant 1 - \eta, \ \forall n \geqslant N. \end{aligned} \tag{E.60}$$

Moreover, by arguing as in equation (D.20), for any $\eta$ there exist $0 < L < \infty$ and $N \in \mathbb{N}$ such that $\forall n \geqslant N$

$$\mathbf{P}\left( \sup_{\lambda' \in B^d} \left\| g^{\mathcal{C}}_n(\theta_n + \frac{\lambda'\rho}{\sqrt{n}}) \right\| \leqslant L \right) \geqslant 1 - \eta. \tag{E.61}$$

For any invertible matrix $K$, $\|K^{-1}\|_2 = (\min\{\sqrt{\alpha} : \ \alpha \ \text{ is an eigenvalue of } KK'\})^{-1}$. Hence, by the proof of Lemma E.7 and the definition of $\mathcal{C}$, for any $\eta > 0$, there exist $0 < L < \infty$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\left( \left\| \left(K^{\mathcal{C}}\right)^{-1} \right\|_2 \leqslant L \right) \geqslant 1 - \eta, \ \forall n \geqslant N, \tag{E.62}$$

By Horn and Johnson (1985, ch. 5.8), for any invertible matrices $K, \tilde{K}$ such that $\|\tilde{K}^{-1}(K - \tilde{K})\|_2 < 1$,

$$\|K^{-1} - \tilde{K}^{-1}\|_2 \leqslant \frac{\|\tilde{K}^{-1}(K - \tilde{K})\|_2}{1 - \|\tilde{K}^{-1}(K - \tilde{K})\|_2} \|\tilde{K}^{-1}\|_2. \tag{E.63}$$

By the assumption that $D_{P_n}(\theta_n) \to D$ and Assumption 4.4, for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that

$$\sup_{\lambda \in B^d} \|K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho) - K^{\mathcal{C}}\|_2 \leqslant \eta, \ \forall n \geqslant N. \tag{E.64}$$

By (E.63), the definition of the spectral norm, and the triangle inequality, for any $\eta > 0$, there exist $0 < L_1, L_2 < \infty$ and $N \in \mathbb{N}$ such that

$$\begin{aligned}
&\mathbf{P}\big( \sup_{\lambda \in B^d} \big\| \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\big)^{-1} \big\|_2 \leqslant 2L_1 \big) \\
&\quad \geqslant \mathbf{P}\big( \big\| \big(K^{\mathcal{C}}\big)^{-1} \big\|_2 + \sup_{\lambda \in B^d} \| K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)^{-1} - (K^{\mathcal{C}})^{-1} \|_2 \leqslant 2L_1 \big) \\
&\quad \geqslant \mathbf{P}\left( \big\| \big(K^{\mathcal{C}}\big)^{-1} \big\|_2 \leqslant L_1, \frac{\| \big(K^{\mathcal{C}}\big)^{-1} \|_2^2}{1 - \| \big(K^{\mathcal{C}}\big)^{-1} (K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho) - K^{\mathcal{C}}) \|_2} \leqslant L_2, \ \sup_{\lambda \in B^d} \| K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho) - K^{\mathcal{C}} \|_2 \leqslant \frac{L_1}{L_2} \right) \\
&\quad \geqslant 1 - 2\eta, \ \forall n \geqslant N,
\end{aligned} \tag{E.65}$$

Again by applying (E.63), for any $k > 0$, there exists $N \in \mathbb{N}$ such that

$$\begin{aligned}
&\mathbf{P}\big( \big\| \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda))\big)^{-1} - \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda'))\big)^{-1} \big\|_2 \leqslant k \|\lambda - \lambda'\| \big) \\
&\qquad \geqslant \mathbf{P}\Big( \sup_{\lambda \in B^d} \big\| \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda))\big)^{-1} \big\|_2^2 M\rho \|\bar{\theta}(\theta_n, \lambda) - \bar{\theta}(\theta_n, \lambda')\| \leqslant k \|\lambda - \lambda'\| \Big) \geqslant 1 - \eta, \ \forall n \geqslant N,
\end{aligned} \tag{E.66}$$

where the first inequality follows from $\|K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda)) - K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda'))\|_2 \leqslant M\rho \|\bar{\theta}(\theta_n, \lambda) - \bar{\theta}(\theta_n, \lambda')\| \leqslant M\rho^2/\sqrt{n} \|\lambda - \lambda'\|$ by Assumption 4.4 (ii), and the last inequality follows from (E.65).

By (E.59)-(E.61) and (E.65)-(E.66), it then follows that there exists $\beta \in [0, 1)$ such that for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that

$$\mathbf{P}\left( |\phi_n(\lambda) - \phi_n(\lambda')| \leqslant \beta \|\lambda - \lambda'\|, \ \ \forall \lambda, \lambda' \in B^d \right) \geqslant 1 - \eta, \ \forall n \geqslant N. \tag{E.67}$$

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

$$\mathbf{P}\left( \exists \lambda_n^f : \lambda_n^f = \phi_n(\lambda_n^f) \right) \geqslant 1 - \eta, \ \forall n \geqslant N. \tag{E.68}$$

Next, define the mapping

$$\psi_n(\lambda) \equiv \big(K^{\mathcal{C}}\big)^{-1} g^{\mathcal{C}}. \tag{E.69}$$

This map is constant in $\lambda$ and hence is uniformly continuous and a contraction with Lipschitz constant equal to zero. It therefore has $\lambda_n^{\mathcal{C}}$ as its fixed point. Moreover, by (E.58) and (E.69) arguing as in (E.59), it follows that for any $\lambda \in B^d$,

$$\begin{aligned}
\|\psi_n(\lambda) - \phi_n(\lambda)\| \leqslant &\Big\| \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\big)^{-1} \Big\|_2 \Big\| g^{\mathcal{C}} - g_n^{\mathcal{C}}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) \Big\| \\
&+ \Big\| \big(K^{\mathcal{C}}\big)^{-1} - \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\big)^{-1} \Big\|_2 \|g^{\mathcal{C}}\|.
\end{aligned} \tag{E.70}$$

By (E.52) and (E.56)

$$\begin{aligned}
\Big\| g^{\mathcal{C}} - g_n^{\mathcal{C}}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) \Big\| &\leqslant \max_{j \in \mathcal{J}^*} | - \mathbb{Z}_j^* - c_n^*/(1 + \eta_{n,j}^*) + \mathbb{G}_{n,j}^*(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) | \\
&\leqslant \max_{j \in \mathcal{J}^*} |\mathbb{Z}_j^* - \mathbb{G}_{n,j}^*(\theta_n + \frac{\lambda\rho}{\sqrt{n}})| + \max_{j \in \mathcal{J}^*} |c_n^*/(1 + \eta_{n,j}^*)|.
\end{aligned} \tag{E.71}$$

We note that when Assumption 4.3-(II) is used, for each $j = 1, \ldots, R_1$ such that $\pi_{1,j}^* = 0 = \pi_{1,j+R_1}^*$ 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 $\mu_j$ were defined in (D.11)-(D.12) and (E.10)-(E.11)

respectively. Moreover, $\mathbb{G}^*_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) \xrightarrow{a.s.} \mathbb{Z}^*$ and (E.49) implies $c^*_n \to 0$ so that we have

$$\sup_{\lambda \in B^d} \|g^{\mathcal{C}} - g^{\mathcal{C}}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}})\| \xrightarrow{a.s.} 0. \tag{E.72}$$

Further, by (E.63), $D_{P_n} \to D$ and, Assumption 4.4-(ii), for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that

$$\sup_{\lambda \in B^d} \left\| \left(K^{\mathcal{C}}\right)^{-1} - \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\right)^{-1} \right\|_2 \leqslant \eta, \ \forall n \geqslant N. \tag{E.73}$$

In sum, by (E.61), (E.65), and (E.71)-(E.73), for any $\eta, \nu > 0$, there exists $N \geqslant \mathbb{N}$ such that

$$\mathbf{P}\left( \sup_{\lambda \in B^d} \|\psi_n(\lambda) - \phi_n(\lambda)\| < \nu \right) \geqslant 1 - \eta, \ \forall n \geqslant \mathbb{N}. \tag{E.74}$$

Hence, for a specific choice of $\nu = \kappa(1-\beta)$, where $\beta$ is defined in equation (E.67), we have that $\sup_{\lambda \in B^d} \|\psi_n(\lambda) - \phi_n(\lambda)\| < \kappa(1-\beta)$ implies

$$\begin{aligned} \|\lambda^{\mathcal{C}}_n - \lambda^f_n\| &= \|\psi_n(\lambda^{\mathcal{C}}_n) - \phi_n(\lambda^f_n)\| \\ &\leqslant \|\psi_n(\lambda^{\mathcal{C}}_n) - \phi_n(\lambda^{\mathcal{C}}_n)\| + \|\phi_n(\lambda^{\mathcal{C}}_n) - \phi_n(\lambda^f_n)\| \\ &\leqslant \kappa(1-\beta) + \beta\|\lambda^{\mathcal{C}}_n - \lambda^f_n\| \end{aligned} \tag{E.75}$$

Rearranging terms, we obtain $\|\lambda^{\mathcal{C}}_n - \lambda^f_n\| \leqslant \kappa$. Note that by Assumptions 4.4 (i) and 4.5 (i), for any $\delta > 0$, there exists $\kappa_\delta > 0$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\Big( \sup_{\|\lambda - \lambda'\| \leqslant \kappa_\delta} |u^*_{n,j,\theta_n}(\lambda) - u^*_{n,j,\theta_n}(\lambda')| < \delta \Big) \geqslant 1 - \eta, \ \forall n \geqslant \mathbb{N}. \tag{E.76}$$

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

$$\mathfrak{w}^*_j(\lambda^{\mathcal{C}}_n) + \delta \leqslant 0, \ j \in \{1, \cdots, J_1\} \cap \mathcal{J}^*. \tag{E.77}$$

Hence, by (E.39), (E.49), and (E.76)-(E.77), $\|\lambda^{\mathcal{C}}_n - \lambda^f_n\| \leqslant \kappa_{\delta/4}$, for each $j \in \{1, \cdots, J_1\} \cap \mathcal{J}^*$ we have

$$u^*_{n,j,\theta_n}(\lambda^f_n) - c^*_n(\theta_n) \leqslant u^*_{n,j,\theta_n}(\lambda^{\mathcal{C}}_n) - c^*_n(\theta_n) + \delta/4 \leqslant \mathfrak{w}^*_j(\lambda^{\mathcal{C}}_n) + \delta/2 \leqslant 0. \tag{E.78}$$

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 $n \geqslant N$ we have

$$\begin{aligned} \mathbf{P}\Big( \{U^*_n(\theta_n, c^*_n) = \varnothing\} \cap \{\mathfrak{W}^{*,-\delta}(0) \neq \varnothing\} \Big) &\leqslant \mathbf{P}\Big( \nexists \lambda^f_n \in U^*_n(\theta_n, c^*_n), \exists \lambda^{\mathcal{C}}_n \in \mathfrak{W}^{*,-\delta}(0) \Big) \\ &\leqslant \mathbf{P}\left( \left\{ \sup_{\lambda \in B^d} \|\psi_n(\lambda) - \phi_n(\lambda)\| < \kappa_\delta(1-\beta) \cap A_n \right\}^c \right) \leqslant \eta/3, \end{aligned} \tag{E.79}$$

where $A^c$ denotes the complement of the set $A$, and the last inequality follows from (E.39) and (E.74). □

LEMMA E.3: *Suppose Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. Let $\{P_n, \theta_n\} \in \{(P, \theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ be a sequence satisfying (E.1)-(E.3). For each $j$, let*

$$v^I_{n,j,\theta_n}(\lambda) \equiv \mathbb{G}^b_{n,j}(\theta_n) + \rho\hat{D}_{n,j}(\theta_n)\lambda + \varphi^*_j(\hat{\xi}_{n,j}(\theta_n)), \tag{E.80}$$

$$\mathfrak{w}_j(\lambda) \equiv \mathbb{Z}_j + \rho D_j \lambda + \pi^*_{1,j}, \tag{E.81}$$

*where*

$$\varphi_j^*(\xi) = \begin{cases} \varphi_j(\xi) & \pi_{1,j} = 0 \\ -\infty & \pi_{1,j} < 0 \\ 0 & j = J_1 + 1, \cdots, J. \end{cases} \tag{E.82}$$

*For each $c \geqslant 0$, define*

$$V_n^I(\theta_n, c) \equiv \{\lambda \in B_{n,\rho}^d : p'\lambda = 0 \cap v_{n,j,\theta_n}^I(\lambda) \leqslant c, j = 1, \cdots, J\}, \tag{E.83}$$

$$\mathfrak{W}(c) \equiv \left\{\lambda \in \mathfrak{B}_\rho^d : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c, \ \forall j = 1, \ldots, J\right\}. \tag{E.84}$$

*We then let $c_n^I(\theta_n) \equiv \inf\{c \in \mathbb{R}_+ : P_n^*(V_n^I(\theta_n, c) \neq \varnothing) \geqslant 1 - \alpha\}$ and $c_{\pi*} \equiv \inf\{c \in \mathbb{R}_+ : \Pr(\mathfrak{W}(c) \neq \varnothing) \geqslant 1 - \alpha\}$.*
*Then, (i) for any $c > 0$ and $\{\theta_n'\} \subset \Theta$ such that $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$ for all $n$,*

$$P_n^*(V_n^I(\theta_n', c) \neq \varnothing) - \Pr(\mathfrak{W}(c) \neq \varnothing) \to 0, \tag{E.85}$$

*with probability approaching 1;*
*(ii) If $c_{\pi*} > 0$, $c_n^I(\theta_n') \xrightarrow{P_n} c_{\pi*}$;*
*(iii) For any $\{\theta_n'\} \subset \Theta$ such that $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$ for all $n$,*

$$\hat{c}_n(\theta_n') \geqslant c_n^I(\theta_n') + o_{P_n}(1). \tag{E.86}$$

*Proof.* Throughout, let $c > 0$ and let $\{\theta_n'\} \subset \Theta$ be a sequence such that $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$ for all $n$. By Lemma E.15, in $l^\infty(\Theta)$ uniformly in $\mathcal{P}$ conditional on $\{X_i\}_{i=1}^\infty$, and by Assumption 4.4 $\|\hat{D}_n(\theta_n') - D_{P_n}(\theta_n)\| \xrightarrow{p} 0$. Further, by Lemma E.5, $\hat{\xi}_{n,j}(\theta_n') \xrightarrow{P_n} \pi_{1,j}$. Therefore,

$$(\mathbb{G}_n^b(\theta_n'), \hat{D}_n(\theta_n'), \hat{\xi}_n(\theta_n'))|\{X_i\}_{i=1}^\infty \xrightarrow{d} (\mathbb{Z}, D, \pi_1). \tag{E.87}$$

for almost all sample paths $\{X_i\}_{i=1}^\infty$. By Lemma E.17, conditional on the sample path, there exists an almost sure representation $(\tilde{\mathbb{G}}_n^b(\theta_n'), \tilde{D}_n, \tilde{\xi}_n)$ of $(\mathbb{G}_n^b(\theta_n'), \hat{D}_n(\theta_n'), \hat{\xi}_n(\theta_n'))$ defined on another probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbf{P}})$ such that $(\tilde{\mathbb{G}}_n^b(\theta_n'), \tilde{D}_n, \tilde{\xi}_n) \overset{d}{=} (\mathbb{G}_n^b(\theta_n'), \hat{D}_n(\theta_n'), \hat{\xi}_n(\theta_n'))$ conditional on the sample path. In particular, conditional on the sample, $(\hat{D}_n(\theta_n'), \hat{\xi}_n(\theta_n'))$ are non-stochastic. Therefore, we set $(\tilde{D}_n, \tilde{\xi}_n) = (\hat{D}_n(\theta_n'), \hat{\xi}_n(\theta_n'))$, $\tilde{\mathbf{P}} - a.s.$ The almost sure representation satisfies $(\tilde{\mathbb{G}}_n^b(\theta_n'), \tilde{D}_n, \tilde{\xi}_{n,j}) \xrightarrow{a.s.} (\tilde{\mathbb{Z}}, D, \pi_1)$ for almost all sample paths, where $\tilde{\mathbb{Z}} \overset{d}{=} \mathbb{Z}$. The almost sure representation $(\tilde{\mathbb{G}}_n^b, \tilde{D}_n, \tilde{\xi}_n)$ is defined for each sample path $x^\infty = \{x_i\}_{i=1}^\infty$, but we suppress its dependence on $x^\infty$ for notational simplicity (see Appendix E.3 for details). Using this representation, define

$$\tilde{v}_{n,j,\theta_n'}^I(\lambda) \equiv \tilde{\mathbb{G}}_{n,j}^b(\theta_n') + \rho\tilde{D}_n\lambda + \varphi_j^*(\tilde{\xi}_{n,j}), \tag{E.88}$$

and

$$\tilde{\mathfrak{w}}_j(\lambda) \equiv \tilde{\mathbb{Z}}_j + \rho D_j\lambda + \pi_{1,j}^*, \tag{E.89}$$

where $\tilde{\mathbb{Z}} \overset{d}{=} \mathbb{Z}$, and $\tilde{\mathbb{G}}_n^b(\theta_n') \to \tilde{\mathbb{Z}}, \tilde{\mathbf{P}} - a.s.$ conditional on $\{X_i\}_{i=1}^\infty$. With this construction, one may write

$$\begin{aligned} |P_n^*(V_n^I(\theta_n', c) \neq \varnothing) - \Pr(\mathfrak{W}(c) \neq \varnothing)| &= |\tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n', c) \neq \varnothing) - \tilde{\mathbf{P}}(\tilde{\mathfrak{W}}(c) \neq \varnothing)| \\ &\leqslant |\tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n', c) = \varnothing \cap \tilde{\mathfrak{W}}(c) \neq \varnothing) + \tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n', c) \neq \varnothing \cap \tilde{\mathfrak{W}}(c) = \varnothing)|, \end{aligned} \tag{E.90}$$

where the inequality is due to (E.33). First, we bound the first term on the right hand side of (E.90). Note that

$$\tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n',c)=\varnothing\cap\tilde{\mathfrak{W}}(c)\neq\varnothing)\leqslant\tilde{\mathbf{P}}(\tilde{V}_n^{I,+\delta}(\theta_n',c)=\varnothing\cap\tilde{\mathfrak{W}}(c)\neq\varnothing)+\tilde{\mathbf{P}}(\tilde{V}_n^{I,+\delta}(\theta_n',c)\neq\varnothing\cap\tilde{V}_n^I(\theta_n',c)=\varnothing), \tag{E.91}$$

where $\tilde{V}_n^{I,+\delta}$ is defined as

$$\tilde{V}_n^{I,+\delta}\equiv\Big\{\lambda\in B_{n,\rho}^d:p'\lambda=0\cap\tilde{v}_{n,j,\theta_n'}^I(\lambda)\leqslant c+\delta,j\in\mathcal{J}^*\Big\}. \tag{E.92}$$

Let

$$A_n\equiv\Big\{\tilde{\omega}\in\tilde{\Omega}:\sup_{\lambda\in B^d}\max_{j\in\mathcal{J}^*}|\tilde{v}_{n,j,\theta_n'}^I(\lambda)-\tilde{\mathfrak{w}}_j(\lambda)|\geqslant\delta\Big\}. \tag{E.93}$$

Let

$$E\equiv\{\{x_i\}_{i=1}^\infty:\|\hat{D}_n(\theta_n')-D\|<\eta,\max_{j\in\mathcal{J}^*}|\varphi_j^*(\hat{\xi}_{n,j}(\theta_n'))-\pi_{1,j}^*|<\eta\}. \tag{E.94}$$

Note that, $P_n(E)\geqslant 1-\eta$ for all $n$ sufficiently large by Assumption 4.4 and Lemma E.5. On $E$, we therefore have $\|\tilde{D}_n-D\|<\eta$ and $\max_{j\in\mathcal{J}^*}|\tilde{\xi}_{n,j}-\pi_{1,j}^*|<\eta$, $\tilde{\mathbf{P}}-a.s.$ Below, we condition on $\{X_i\}_{i=1}^\infty\in E$. For any $j\in\mathcal{J}^*$,

$$|\tilde{v}_{n,j,\theta_n'}^I(\lambda)-\tilde{\mathfrak{w}}_j(\lambda)|\leqslant|\tilde{\mathbb{G}}_{n,j}^b(\theta_n')-\tilde{\mathbb{Z}}_j|+\rho\|\tilde{D}_{j,n}-D_j\|\|\lambda\|+|\varphi_j^*(\tilde{\xi}_{n,j})-\pi_{1,j}^*|\leqslant(2+\rho)\eta, \tag{E.95}$$

uniformly in $\lambda\in B^d$, where we used $\tilde{\mathbb{G}}_n^b\to\tilde{\mathbb{Z}},\tilde{\mathbf{P}}-a.s.$ Since $\eta$ can be chosen arbitrarily small, this in turn implies

$$\tilde{\mathbf{P}}\big(A_n\big)<\eta/2,$$

for all $n$ sufficiently large. Note also that $\sup_{\lambda\in B^d}\max_{j\in\mathcal{J}^*}|\tilde{v}_{n,j,\theta_n'}^I(\lambda)-\tilde{\mathfrak{w}}_j(\lambda)|<\delta$ implies $\tilde{\mathfrak{W}}(c)\subseteq\tilde{V}_n^{I,+\delta}(\theta_n',c)$, and hence $A_n^c$ is a subset of

$$L_n\equiv\Big\{\tilde{\omega}\in\tilde{\Omega}:\tilde{\mathfrak{W}}(c)\subseteq\tilde{V}_n^{I,+\delta}(\theta_n',c)\Big\}. \tag{E.96}$$

Using this,

$$\tilde{\mathbf{P}}(\tilde{V}_n^{I,+\delta}(\theta_n',c)=\varnothing\cap\tilde{\mathfrak{W}}(c)\neq\varnothing)\leqslant\tilde{\mathbf{P}}(\tilde{\mathfrak{W}}(c)\nsubseteq\tilde{V}_n^{I,+\delta}(\theta_n',c))=\tilde{\mathbf{P}}(L_n^c)\leqslant\tilde{\mathbf{P}}(A_n)<\eta/2, \tag{E.97}$$

for all $n$ sufficiently large. Also, by Lemma E.6,

$$\tilde{\mathbf{P}}(\tilde{V}_n^{I,+\delta}(\theta_n',c)\neq\varnothing\cap\tilde{V}_n^I(\theta_n',c)=\varnothing)<\eta/2, \tag{E.98}$$

for all $n$ sufficiently large.

Combining (E.91), (E.93), (E.97), (E.98), and using $P_n(E)\geqslant 1-\eta$ for all $n$, we have

$$\int_E\tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n',c)=\varnothing\cap\tilde{\mathfrak{W}}(c)\neq\varnothing)dP_n+\int_{E^c}\tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n',c)=\varnothing\cap\tilde{\mathfrak{W}}(c)\neq\varnothing)dP_n\leqslant\eta(1-\eta)+\eta\leqslant 2\eta. \tag{E.99}$$

The second term of the right hand side of (E.90) can be bounded similarly. Therefore, $|P^*(V_n^I(\theta_n',c)\neq\varnothing)-\Pr(\mathfrak{W}(c)\neq\varnothing)|\to 0$ with probability (under $P_n$) approaching 1. This establishes the first claim.

(ii) By Part (i), for $c>0$, we have

$$P_n^*(V_n^I(\theta_n',c)\neq\varnothing)-\Pr(\mathfrak{W}(c)\neq\varnothing)\to 0. \tag{E.100}$$

Fix $c > 0$, and set

$$g_j = \begin{cases} c - \mathbb{Z}_j, & j = 1, \ldots, J, \\ 1, & j = J+1, \ldots, J+2d, \\ 0, & j = J+2d+1, J+2d+2. \end{cases} \tag{E.101}$$

Mimic the argument following (E.137). Then, this yields

$$|\Pr(\mathfrak{W}(c) \neq \varnothing) - \Pr(\mathfrak{W}(c-\delta) \neq \varnothing)| = \Pr(\{\mathfrak{W}(c) \neq \varnothing\} \cap \{\mathfrak{W}(c-\delta) = \varnothing\}) \leqslant \eta, \tag{E.102}$$

$$|\Pr(\mathfrak{W}(c+\delta) \neq \varnothing) - \Pr(\mathfrak{W}(c) \neq \varnothing)| = \Pr(\{\mathfrak{W}(c+\delta) \neq \varnothing\} \cap \{\mathfrak{W}(c) = \varnothing\}) \leqslant \eta, \tag{E.103}$$

which therefore ensures that $c \mapsto \Pr(\mathfrak{W}(c) \neq \varnothing)$ is continuous at $c > 0$.

Next, we show $c \mapsto \Pr(\mathfrak{W}(c) \neq \varnothing)$ is strictly increasing at any $c > 0$. For this, consider $c > 0$ and $c - \delta > 0$ for $\delta > 0$. Define the $J$ vector $e$ to have elements $e_j = c - \mathbb{Z}_j,\ j = 1, \ldots, J$. 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

$$K = \begin{bmatrix} D^{[1:J^*,:]} \\ I_d \\ -I_d \\ p' \\ -p' \end{bmatrix},\ g = \begin{bmatrix} e^{[1:J^*]} \\ \rho \cdot \mathbf{1}_d \\ \rho \cdot \mathbf{1}_d \\ 0 \\ 0 \end{bmatrix},\ \tau = \begin{bmatrix} \mathbf{1}_{J^*} \\ \mathbf{0}_d \\ \mathbf{0}_d \\ 0 \\ 0 \end{bmatrix}. \tag{E.104}$$

By Farkas' lemma (Rockafellar, 1970, Theorem 22.1) and arguing as in (E.142),

$$\Pr(\{\mathfrak{W}(c) \neq \varnothing\} \cap \{\mathfrak{W}(c-\delta) = \varnothing\}) = \Pr\left(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu'(g - \delta\tau) < 0, \exists \mu \in \mathcal{M}\}\right), \tag{E.105}$$

where $\mathcal{M} = \{\mu \in \mathbb{R}_+^{J^*+2d+2} : \mu' K = 0\}$. By Minkowski-Weyl's theorem (Rockafellar and Wets, 2005, Theorem 3.52), there exists $\{\nu^t \in \mathcal{M}, t = 1, \cdots, T\}$, for which one may write

$$\mathcal{M} = \{\mu : \mu = b \sum_{t=1}^{T} a_t \nu^t, b > 0, a_t \geqslant 0, \sum_{t=1}^{T} a_t = 1\}. \tag{E.106}$$

This implies

$$\mu' g \geqslant 0,\ \forall \mu \in \mathcal{M} \ \Leftrightarrow\ \nu^{t\prime} g \geqslant 0,\ \forall t \in \{1, \cdots, T\} \tag{E.107}$$

$$\mu'(g - \delta\tau) < 0,\ \exists \mu \in \mathcal{M} \ \Leftrightarrow\ \nu^{t\prime} g < \delta \nu^{t\prime} \tau,\ \exists t \in \{1, \cdots, T\}. \tag{E.108}$$

Hence,

$$\Pr\left(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu'(g - \delta\tau) < 0, \exists \mu \in \mathcal{M}\}\right) = \Pr\left(0 \leqslant \nu^{s\prime} g,\ 0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau,\ \forall s, \exists t\right) \tag{E.109}$$

Note that by (E.104), for each $s \in \{1, \cdots, T\}$,

$$\nu^{s\prime} g = \nu^{s,[1:J^*]\prime}(c 1_{\mathcal{J}^*} - \mathbb{Z}_{\mathcal{J}^*}) + \rho \sum_{j=J^*+1}^{J^*+2d} \nu^{s,[j]}, \tag{E.110}$$

$$\nu^{s\prime} \tau = \sum_{j=1}^{J^*} \nu^{s,[j]}. \tag{E.111}$$

For each $s \in \{1, \cdots, T\}$, let

$$h_s^U \equiv c \sum_{j=1}^{J^*} \nu^{s,[j]} + \rho \sum_{j=J^*+1}^{J^*+2d} \nu^{s,[j]} \tag{E.112}$$

$$h_s^L \equiv (c-\delta) \sum_{j=1}^{J^*} \nu^{s,[j]}, \tag{E.113}$$

where $0 \leqslant h_s^L < h_s^U$ for all $s \in \{1, \cdots, T\}$ due to $0 < c - \delta < c$ and $\nu^s \in \mathbb{R}_+^{J^*+2d+2}$. One may therefore rewrite the probability on the right hand side of (E.109) as

$$\Pr\left(0 \leqslant \nu^{s\prime} g,\ 0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau,\ \forall s, \exists t\right) = \Pr\left(\nu^{s,[1:J^*]\prime} \mathbb{Z}_{\mathcal{J}^*} \leqslant h_s^U, h_t^L < \nu^{t,[1:J^*]\prime} \mathbb{Z}_{\mathcal{J}^*} \leqslant h_t^U\ \forall s, \exists t\right) > 0, \tag{E.114}$$

where the last inequality follows because $\mathbb{Z}_{\mathcal{J}^*}$'s correlation matrix $\Omega$ has an eigenvalue bounded away from 0 by Assumption 4.3. By (E.105), (E.109), and (E.114), $c \mapsto \Pr(\mathfrak{W}(c) \neq \varnothing)$ is strictly increasing at any $c > 0$.

Suppose that $c_{\pi *} > 0$, then arguing as in Lemma 5.(i) of Andrews and Guggenberger (2010), we obtain $c_n^I(\theta_n') \xrightarrow{P_n} c_{\pi *}$.

(iii) Begin with observing that one can equivalently express $\hat{c}_n$ (originally defined in (3.5)) as $\hat{c}_n(\theta) = \inf\{c \in \mathbb{R}_+ : P_n^*(V_n^b(\theta, c) \neq \varnothing) \geqslant 1 - \alpha\}$.

Suppose first that Assumption 4.3-(I) holds. In this case, there are no paired inequalities, and $V_n^I$ differs from $V_n^b$ only in terms of the function $\varphi_j^*$ in (E.82) used in place of the GMS function $\varphi_j$. In particular, $\varphi_j^*(\xi) \leqslant \varphi_j(\xi)$ for any $j$ and $\xi$, and therefore $\hat{c}_n(\theta_n) \geqslant c_n^I(\theta_n)$ by construction.

Next, suppose Assumption 4.3-(II) holds and $V_n^I(\theta_n', c)$ is defined with hard threshold GMS as in equation (3.3), i.e. with GMS function $\varphi^1$ in AS. The only case that might create concern is one in which

$$\pi_{1,j} \in [-1, 0) \text{ and } \pi_{1,j+R_1} = 0. \tag{E.115}$$

In this case, only the $j + R_1$-th inequality binds in the limit, but with probability approaching 1, GMS selects both of the pair. Therefore, we have

$$\pi_{1,j}^* = -\infty, \text{ and } \pi_{1,j+R_1}^* = 0, \tag{E.116}$$

$$\varphi_j(\hat{\xi}_{n,j}(\theta_n')) = 0, \text{ and } \varphi_{j+R_1}(\hat{\xi}_{n,j+R_1}(\theta_n')) = 0, \tag{E.117}$$

so that in $V_n^I(\theta_n', c)$, inequality $j + R_1$, which is

$$\mathbb{G}_{n,j+R_1}^b(\theta_n') + \rho \hat{D}_{n,j+R_1}(\theta_n')\lambda \leqslant c, \tag{E.118}$$

is replaced with inequality

$$-\mathbb{G}_{n,j}^b(\theta_n') - \rho \hat{D}_{n,j}(\theta_n')\lambda \leqslant c, \tag{E.119}$$

as explained in Section 4.1. In this case, $\hat{c}_n(\theta_n) \geqslant c_n^I(\theta_n)$ is not guaranteed in finite sample. However, let $v_n^{IP}$ be as in (E.80) but replacing $j + R_1$-th component $\mathbb{G}_{n,j+R_1}^b(\theta_n) + \hat{D}_{n,j+R_1}(\theta_n)\lambda + \varphi_{j+R_1}^*(\hat{\xi}_{n,j+R_1}(\theta_n))$ with $-\mathbb{G}_{n,j}^b(\theta_n) - \hat{D}_{n,j}(\theta_n)\lambda - \varphi_j^*(\hat{\xi}_{n,j}(\theta_n))$. Define $V_n^{IP}$ as in (E.83) but replacing $v_n^I$ with $v_n^{IP}$. Define $c_n^{IP}(\theta_n) \equiv \inf\{c \in \mathbb{R}_+ : P^*(V_n^{IP}(\theta_n, c)) \geqslant 1 - \alpha\}$. By construction, $\hat{c}_n(\theta_n') \geqslant c_n^{IP}(\theta_n')$ for any $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$. Therefore, it suffices to show that $c_n^{IP}(\theta_n') - c_n^I(\theta_n') \xrightarrow{P_n} 0$. For this, note that Lemma E.9-(3) establishes

$$\sup_{\lambda \in B_{n,\rho}^d} \|\mathbb{G}_{n,j+R_1}^b(\theta_n') + \rho \hat{D}_{n,j+R_1}(\theta_n')\lambda + \mathbb{G}_{n,j}^b(\theta_n') + \rho \hat{D}_{n,j}(\theta_n')\lambda\| = o_{P^*}(1), \tag{E.120}$$

for almost all sample paths $\{X_i\}_{i=1}^{\infty}$. Therefore, replacing the $j+R_1$-th inequality with the $j$-th inequality in $V_n^{IP}$ is asymptotically negligible. Mimicking the arguments in Parts (i) and (ii) then yields

$$c_n^{IP}(\theta_n') \xrightarrow{P_n} c_{\pi *}. \tag{E.121}$$

This therefore ensures $c_n^{IP}(\theta_n') - c_n^{I}(\theta_n') \xrightarrow{P_n} 0$.

If the set $V_n^I(\theta_n', c)$ is defined with a GMS function satisfying Assumption 4.2 and continuous in its argument, we can mimic the above argument using the replacements in (E.12)-(E.13) with $\hat{\mu}_{n,j+R_1}$ as defined in (E.14) and $\hat{\mu}_{n,j}(\theta_n')$ as in (E.15). Then when both $\pi_j \in (-\infty, 0]$ and $\pi_{j+R_1} \in (-\infty, 0]$ we have:

$$\begin{aligned}\Delta(\mu, \hat{\mu}) \equiv \Big\| &\hat{\mu}_{n,j}(\theta_n')\{\mathbb{G}_{n,j}^b(\theta_n') + \rho\hat{D}_{n,j}(\theta_n')\lambda\} - \hat{\mu}_{n,j+R_1}(\theta_n')\{\mathbb{G}_{n,j+R_1}^b(\theta_n') + \rho\hat{D}_{n,j+R_1}(\theta_n')\lambda\} \\ &-\mu_j(\theta_n')\{\mathbb{G}_{n,j}^b(\theta_n') + \rho\hat{D}_{n,j}(\theta_n')\lambda\} + \mu_{j+R_1}(\theta_n')\{\mathbb{G}_{n,j+R_1}^b(\theta_n') + \rho\hat{D}_{n,j+R_1}(\theta_n')\lambda\}\Big\| = o_{\mathcal{P}}(1),\end{aligned}$$

where $\mu_j, \mu_{j+R_1}$ are defined in equations (E.10)-(E.11) for $\theta \in \theta_n + (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$. Replacing $\hat{\mu}_{n,j} = 1 - \hat{\mu}_{n,j+R_1}$ and $\mu_j = 1 - \mu_{j+R_1}$ in the definition of $\Delta(\mu, \hat{\mu})$, we have

$$\Delta(\mu, \hat{\mu}) \leqslant \big|\hat{\mu}_{n,j+R_1}(\theta_n') - \mu_{j+R_1}(\theta_n')\big|\big\|\{\mathbb{G}_{n,j+R_1}^b(\theta_n') + \rho\hat{D}_{n,j+R_1}(\theta_n')\lambda\} + \{\mathbb{G}_{n,j}^b(\theta_n') + \rho\hat{D}_{n,j}(\theta_n')\lambda\}\big\|. \tag{E.122}$$

If both $\pi_j \in (-\infty, 0], \pi_{j+R_1} \in (-\infty, 0]$, the result follows by the fact that $\lambda \in B_{n,\rho}^d$ and $\hat{\mu}_{n,j}, \hat{\mu}_{n,j+R_1}, \mu_j, \mu_{j+R_1}$ are bounded in $[0, 1]$, by Lemma E.9-(3)-(4), and by Assumption 4.4-(i). The rest of the argument follows similarly as for the case of hard-threshold GMS. □

LEMMA E.4: *Let Assumptions 4.1, 4.2, 4.4, and 4.5 hold. Let $(P_n, \theta_n)$ be the sequence satisfying (E.1)-(E.3), let $\mathcal{J}^*$ be defined as in (E.29), and assume that $\mathcal{J}^* \neq \varnothing$. Then, for any $\varepsilon, \eta > 0$ and $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$, there exists $N' \in \mathbb{N}$ and $N'' \in \mathbb{N}$ such that for all $n \geqslant \max\{N', N''\}$,*

$$\mathbf{P}\left(\sup_{\lambda \in B^d}\left|\max_{j=1,\cdots,J}(u_{n,j,\theta_n}^*(\lambda) - c_n^*) - \max_{j=1,\cdots,J}(\mathfrak{w}_j^*(\lambda) - c_{\pi *})\right| \geqslant \varepsilon\right) < \eta, \tag{E.123}$$

$$\tilde{\mathbf{P}}\left(\sup_{\lambda \in B^d}\left|\max_{j=1,\cdots,J}\tilde{\mathfrak{w}}_j(\lambda) - \max_{j=1,\cdots,J}\tilde{v}_{n,j,\theta_n'(\lambda)}^I\right| \geqslant \varepsilon\right) < \eta, \ w.p.1, \tag{E.124}$$

*where the functions $u_n^*, \mathfrak{w}^*, \tilde{v}_n, \tilde{\mathfrak{w}}$ are defined in equations (E.24),(E.25), (E.88), and (E.89).*

*Proof.* We first establish (E.123). By definition, $\pi_{1,j}^* = -\infty$ for all $j \notin \mathcal{J}^*$ and therefore

$$\begin{aligned}&\mathbf{P}\Big(\sup_{\lambda \in B^d}|\max_{j=1,\cdots,J}(u_{n,j,\theta_n}^*(\lambda) - c_n^*) - \max_{j=1,\cdots,J}(\mathfrak{w}_j^*(\lambda) - c_{\pi *})| \geqslant \varepsilon\Big) \\ &\qquad = \mathbf{P}\Big(\sup_{\lambda \in B^d}|\max_{j \in \mathcal{J}^*}(u_{n,j,\theta_n}^*(\lambda) - c_n^*) - \max_{j \in \mathcal{J}^*}(\mathfrak{w}_j^*(\lambda) - c_{\pi *})| \geqslant \varepsilon\Big).\end{aligned} \tag{E.125}$$

Hence, for the conclusion of the lemma, it suffices to show, for any $\varepsilon > 0$,

$$\lim_{n \to \infty}\mathbf{P}\Big(\sup_{\lambda \in B^d}|\max_{j \in \mathcal{J}^*}(u_{n,j,\theta_n}^*(\lambda) - c_n^*) - \max_{j \in \mathcal{J}^*}(\mathfrak{w}_j^*(\lambda) - c_{\pi *})| \geqslant \varepsilon\Big) = 0.$$

For each $\lambda \in \mathbb{R}^d$, define $r_{n,j,\theta_n}(\lambda) \equiv (u_{n,j,\theta_n}^*(\lambda) - c_n^*) - (\mathfrak{w}_j^*(\lambda) - c_n)$. Using the fact that $\pi_{1,j}^* = 0$ for $j \in \mathcal{J}^*$,

and the triangle and Cauchy-Schwarz inequalities, for any $\lambda \in B^d \cap \frac{\sqrt{n}}{\rho}(\Theta - \theta_n)$ and $j \in \mathcal{J}^*$, we have

$$\begin{aligned}
|r_{n,j,\theta_n}(\lambda)| &\leqslant |\mathbb{G}^*_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) - \mathbb{Z}^*_j| + \rho\|D_{P_n,j}(\bar{\theta}_n) - D_j\|\|\lambda\| \\
&\quad + |\mathbb{G}^*_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda|\eta^*_{n,j} + |c^*_n - c_{\pi*}| \\
&= |\mathbb{G}^*_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) - \mathbb{Z}^*_j| + o(1) + \{O_{\mathcal{P}}(1) + O(1)\})\eta^*_{n,j} + o_{\mathcal{P}}(1) \\
&= o_{\mathcal{P}}(1) \qquad \text{(E.126)}
\end{aligned}$$

where the first equality follows from $\|\lambda\| \leqslant \sqrt{d}$, $D_{P_n}(\bar{\theta}_n) \to D$ due to $D_{P_n}(\theta_n) \to D$, Assumption 4.4-(ii), and $\bar{\theta}_n$ being a mean value between $\theta_n$ and $\theta_n + \lambda\rho/\sqrt{n}$. We also note that $\|\mathbb{G}_{n,j}(\theta + \lambda/\sqrt{n})\| = O_{\mathcal{P}}(1)$, $\|D_{P,j}(\theta)\|$ being uniformly bounded for $\theta \in \Theta_I(P)$ (Assumption 4.4-(i)), and $c^*_n \xrightarrow{a.s.} c_{\pi*}$. The last equality follows from $\mathbb{G}^*_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) - \mathbb{Z}^*_j \xrightarrow{a.s.} 0$ and $\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)| = o_{\mathcal{P}}(1)$ by Lemma E.10.

We note that when paired inequalities are merged, for each $j = 1, \ldots, R_1$ such that $\pi^*_{1,j} = 0 = \pi^*_{1,j+R_1}$ 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 $\mu_j$ were defined in (D.11)-(D.12) and (E.10)-(E.11) respectively.

By (E.126) and the fact that $j \in \mathcal{J}^*$, we have

$$\sup_{\lambda\in B^d} |\max_{j\in\mathcal{J}^*}(u^*_{n,j,\theta_n}(\lambda) - c^*_n) - \max_{j\in\mathcal{J}^*}(\mathfrak{w}^*_j(\lambda) - c_{\pi*})| \leqslant \sup_{\lambda\in B^d}\max_{j\in\mathcal{J}^*}|r_{n,j,\theta_n}(\lambda)| = o_{\mathcal{P}}(1). \qquad \text{(E.127)}$$

The conclusion of the lemma then follows from (E.125) and (E.127).

The result in (E.124) follows from similar arguments. □

LEMMA E.5: *Let Assumptions 4.1, 4.2, 4.4, and 4.5 hold. 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*

$$\chi_j(\{Q_n, \vartheta_n\}) \equiv \begin{cases} 0, & \text{if } \lim_{n\to\infty}\kappa_n^{-1}\sqrt{n}\gamma_{1,Q_n,j}(\vartheta_n) = 0, \\ -\infty, & \text{if } \lim_{n\to\infty}\kappa_n^{-1}\sqrt{n}\gamma_{1,Q_n,j}(\vartheta_n) < 0. \end{cases} \qquad \text{(E.128)}$$

*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) $\kappa_n^{-1}\frac{\sqrt{n}\bar{m}_{n,j}(\theta'_n)}{\tilde{\sigma}_{n,j}(\theta'_n)} - \kappa_n^{-1}\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta'_n)]}{\sigma_{P_n,j}(\theta'_n)} = o_{\mathcal{P}}(1)$.*

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

$$\begin{aligned}
&\sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P),\theta'\in\theta+\rho/\sqrt{n}B^d}\left|\frac{\sqrt{n}E_P(m_j(X,\theta))}{\kappa_n\sigma_{P,j}(\theta)} - \frac{\sqrt{n}E_P(m_j(X,\theta'))}{\kappa_n\sigma_{P,j}(\theta')}\right| \\
&\qquad \leqslant \sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P),\theta'\in\theta+\rho/\sqrt{n}B^d}\frac{\sqrt{n}\|D_{P,j}(\tilde{\theta})\|\|\theta'-\theta\|}{\kappa_n} = o(1), \qquad \text{(E.129)}
\end{aligned}$$

where $\tilde{\theta}$ represents a mean value that lies componentwise between $\theta$ and $\theta'$ and where we used the fact that $D_{P,j}(\theta)$ is Lipschitz continuous and $\sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P)}\|D_{P,j}(\theta)\| \leqslant \bar{M}$. Result (ii) then follows immediately from (E.128).

For (iii), note that

$$\begin{aligned}
&\sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d}\left|\kappa_n^{-1}\frac{\sqrt{n}\bar{m}_{n,j}(\theta_n')}{\hat{\sigma}_{n,j}(\theta_n')}-\kappa_n^{-1}\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')}\right| \\
&\leqslant \sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d}\left|\kappa_n^{-1}\frac{\sqrt{n}(\bar{m}_{n,j}(\theta_n')-E_{P_n}[m_j(X_i,\theta_n')])}{\sigma_{n,j}(\theta_n')}(1+\eta_{n,j}(\theta_n'))+\kappa_n^{-1}\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')}\eta_{n,j}(\theta_n')\right| \\
&\leqslant \sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d}|\kappa_n^{-1}\mathbb{G}_n(\theta_n')(1+\eta_{n,j}(\theta_n'))|+\left|\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\kappa_n\sigma_{P_n,j}(\theta_n')}\eta_{n,j}(\theta_n')\right|=o_{\mathcal{P}}(1), \quad \text{(E.130)}
\end{aligned}$$

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 $\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)|=o_{\mathcal{P}}(1)$ by Lemma E.10-(i). □

LEMMA E.6: *Let Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. For any $\theta_n'\in(\theta_n+\rho/\sqrt{n}B^d)\cap\Theta$,*

*(i) For any $\eta>0$, there exist $\delta>0$ such that*

$$\sup_{c\geqslant 0}\Pr(\{\mathfrak{W}(c)\neq\varnothing\}\cap\{\mathfrak{W}^{-\delta}(c)=\varnothing\})<\eta. \quad \text{(E.131)}$$

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

$$\sup_{c\geqslant 0}P_n^*(\{V_n^I(\theta_n',c)\neq\varnothing\}\cap\{V_n^{I,-\delta}(\theta_n',c)=\varnothing\})<\eta,\ \forall n\geqslant N. \quad \text{(E.132)}$$

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

$$\mathfrak{W}^{-\delta}(c)\equiv\left\{\lambda\in\mathfrak{B}_\rho^d:p'\lambda=0\cap\mathfrak{w}_j(\lambda)\leqslant c-\delta,\ \forall j=1,\ldots,J\right\}, \quad \text{(E.133)}$$

*and*

$$V_n^{I,-\delta}(\theta_n',c)\equiv\left\{\lambda\in B_{n,\rho}^d:p'\lambda=0\cap v_{n,j,\theta_n'}^I(\lambda)\leqslant c-\delta,\ \forall j=1,\ldots,J\right\}. \quad \text{(E.134)}$$

*Then for any $\eta>0$, there exists $\delta>0$ such that*

$$\sup_{c\geqslant\underline{c}}\Pr(\{\mathfrak{W}(c)\neq\varnothing\}\cap\{\mathfrak{W}^{-\delta}(c)=\varnothing\})<\eta. \quad \text{(E.135)}$$

*with $\mathfrak{W}^{-\delta}(c)$ defined in (E.133). Moreover, for any $\eta>0$, there exist $\delta>0$ and $N\in\mathbb{N}$ such that*

$$\sup_{c\geqslant\underline{c}}P_n^*(\{V_n^I(\theta_n',c)\neq\varnothing\}\cap\{V_n^{I,-\delta}(\theta_n',c)=\varnothing\})<\eta,\ \forall n\geqslant N, \quad \text{(E.136)}$$

*with $V_n^{-\delta}(\theta_n',c)$ defined in (E.134).*

*Proof.* We first show (E.131). If $\mathcal{J}^*=\varnothing$, with $\mathcal{J}^*$ as defined in (E.29), then the result is immediate. Assume then that $\mathcal{J}^*\neq\varnothing$. Any inequality indexed by $j\notin\mathcal{J}^*$ is satisfied with probability approaching one by similar arguments as in (D.20) (both with $c$ and with $c-\delta$). Hence, one could argue for sets $\mathfrak{W}(c),\mathfrak{W}^{-\delta}(c)$ defined as in equations (E.16) and (E.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 $c \geqslant 0$ be given. Let $g$ be a $J + 2d + 2$ vector with entries

$$g_j = \begin{cases} c - \mathbb{Z}_j, & j = 1, \ldots, J, \\ 1, & j = J+1, \ldots, J+2d, \\ 0, & j = J+2d+1, J+2d+2, \end{cases} \tag{E.137}$$

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

$$\tau_j = \begin{cases} 1, & j = 1, \ldots, J_1, \\ 0, & j = J_1 + 1, \ldots, J + 2d + 2. \end{cases} \tag{E.138}$$

Then we can express the sets of interest as

$$\mathfrak{W}(c) = \{\lambda : K\lambda \leqslant g\}, \tag{E.139}$$

$$\mathfrak{W}^{-\delta}(c) = \{\lambda : K\lambda \leqslant g - \delta\tau\}. \tag{E.140}$$

By Farkas' Lemma, e.g. Rockafellar (1970, Theorem 22.1), a solution to the system of linear inequalities in (E.139) exists if and only if for all $\mu \in \mathbb{R}_+^{J+2d+2}$ such that $\mu' K = 0$, one has $\mu' g \geqslant 0$. Similarly, a solution to the system of linear inequalities in (E.140) exists if and only if for all $\mu \in \mathbb{R}^{J+2d+2}$ such that $\mu' K = 0$, one has $\mu'(g - \delta\tau) \geqslant 0$. Define

$$\mathcal{M} \equiv \{\mu \in \mathbb{R}_+^{J+2d+2} : \mu' K = 0\}. \tag{E.141}$$

Then, one may write

$$\begin{aligned} &\Pr(\{\mathfrak{W}(c) \neq \varnothing\} \cap \{W^{-\delta}(\theta_n', c) = \varnothing\}) \\ &\quad = \Pr(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu'(g - \delta\tau) < 0, \exists \mu \in \mathcal{M}\}) \\ &\quad = \Pr(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu' g < \delta\mu'\tau, \exists \mu \in \mathcal{M}\}). \end{aligned} \tag{E.142}$$

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

$$\mu = b \sum_{t=1}^{T} a_t \nu^t, \tag{E.143}$$

where $b > 0$ and $a_t \geqslant 0,\ t = 1, \ldots, T,\ \sum_{t=1}^{T} a_t = 1$. Hence, if $\mu \in \mathcal{M}$ satisfies $\mu' g < \delta\mu'\tau$, denoting $\nu^{t\prime}$ the transpose of vector $\nu^t$, we have

$$\sum_{t=1}^{T} a_t \nu^{t\prime} g < \delta \sum_{t=1}^{T} a_t \nu^{t\prime} \tau. \tag{E.144}$$

However, due to $a_t \geqslant 0, \forall t$ and $\nu^t \in \mathcal{M}$, this means $\nu^{t\prime} g < \delta\nu^{t\prime}\tau$ for some $t \in \{1, \ldots, T\}$. Furthermore, since $\nu^t \in \mathcal{M}$,

we have $0 \leqslant \nu^{t\prime} g$. Therefore,

$$\Pr\left(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu' g < \delta \mu' \tau, \exists \mu \in \mathcal{M}\}\right)$$
$$\leqslant \Pr\left(0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau, \exists t \in \{1, \cdots, T\}\right) \leqslant \sum_{t=1}^{T} \Pr\left(0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau\right). \quad \text{(E.145)}$$

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

$$\nu^{t\prime} g = \sum_{j=J+1}^{J+2d} \nu_j^t,$$
$$\delta \nu^{t\prime} \tau = \delta \sum_{j=J+1}^{J+2d+2} \nu_j^t \tau_j = 0,$$

where the last equality follows by (E.138). Therefore $\Pr\left(0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau\right) = 0$.

**Case 2.** Consider now any $t = 1, \ldots, T$ such that $\nu^t$ assigns positive weight also to constraints in $\{1, \ldots, J\}$. Recall that indices $j = J_1 + 1, \ldots, J_1 + 2J_2$ correspond to moment equalities, each of which is written as two moment inequalities, therefore yielding a total of $2J_2$ inequalities with $D_{j+J_2} = -D_j$ for $j = J_1 + 1, \ldots, J_1 + J_2$, and:

$$g = \begin{cases} c - \mathbb{Z}_j & j = J_1 + 1, \ldots, J_1 + J_2, \\ c + \mathbb{Z}_{j - J_2} & j = J_1 + J_2 + 1, \ldots, J. \end{cases} \quad \text{(E.146)}$$

For each $\nu^t$, (E.146) implies

$$\sum_{j=J_1+1}^{J_1+2J_2} \nu_j^t g_j = c \sum_{j=J_1+1}^{J_1+2J_2} \nu_j^t + \sum_{j=J_1+1}^{J_1+J_2} (\nu_j^t - \nu_{j+J_2}^t) \mathbb{Z}_j. \quad \text{(E.147)}$$

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

$$\tilde{\nu}_j^t \equiv \begin{cases} \nu_j^t & j = 1, \cdots, J_1 \\ \nu_j^t - \nu_{j+J_2}^t & j = J_1 + 1, \cdots, J_1 + J_2. \end{cases}. \quad \text{(E.148)}$$

We then let $\tilde{\nu}^t \equiv (\tilde{\nu}_{n,1}^t, \cdots, \tilde{\nu}_{n,J_1+J_2}^t)'$ and have

$$\nu^{t\prime} g = \sum_{j=1}^{J_1+J_2} \tilde{\nu}_j^t \mathbb{Z}_j + c \sum_{j=1}^{J} \nu_j^t + \sum_{j=J+1}^{J+2d} \nu_j^t. \quad \text{(E.149)}$$

**Case 2-a.** Suppose $\tilde{\nu}^t \neq 0$. Then, by (E.149), $\frac{\nu^{t\prime} g}{\nu^{t\prime} \tau}$ is a normal random variable with variance $(\tilde{\nu}^{t\prime} \tau)^{-2} \tilde{\nu}'^t \Omega \tilde{\nu}^t$. By Assumption 4.3, there exists a constant $\omega > 0$ such that the smallest eigenvalue of $\Omega$ is bounded from below by $\omega$ for all $\theta_n'$. Hence, letting $\|\cdot\|_p$ denote the $p$-norm in $\mathbb{R}^{J+2d+2}$, we have

$$\frac{\tilde{\nu}'^t \Omega \tilde{\nu}^t}{(\tilde{\nu}^{t\prime} \tau)^2} \geqslant \frac{\omega \|\tilde{\nu}^t\|_2^2}{(J+2d+2)^2 \|\tilde{\nu}^t\|_2^2} \geqslant \frac{\omega}{(J+2d+2)^2}. \quad \text{(E.150)}$$

Therefore, the variance of the normal random variable in (E.145) is uniformly bounded away from 0, which in turn allows one to find $\delta > 0$ such that $\Pr(0 \leqslant \frac{\nu^{t\prime} g}{\nu^{t\prime} \tau} < \delta) \leqslant \eta / T$.

**Case 2-b.** Next, consider the case $\tilde{\nu}^t = 0$. Because we are in the case that $\nu^t$ 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 $\sum_{j=1}^J \nu_j^t g \geqslant 0$, 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 $\nu^t$ assigns positive weight only to constraints in $\{J+1, \ldots, J+2d+2\}$ applies and again $\Pr(0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau) = 0$. This establishes equation (E.131).

To see why equation (E.132) holds, observe that the bootstrap distribution is conditional on $X_1, \ldots, X_n$. Therefore, the matrix $\hat{K}_n$, defined as the matrix in equation (E.57) but with $\hat{D}_n$ replacing $D_P$, can be treated as nonstochastic. This implies that the set $\hat{\mathcal{M}}_n$, defined as the set in equation (E.141) but with $\hat{K}_n$ replacing $K$, can be treated as nonstochastic as well.

By an application of Lemma D.2.8 in Bugni, Canay, and Shi (2015) together with Lemma E.17 (through an argument similar to that following equation (E.87)), $\mathbb{G}_n^b \xrightarrow{d} \mathbb{G}_P$ in $l^\infty(\Theta)$ uniformly in $\mathcal{P}$ conditional on $\{X_1, \cdots, X_n\}$, and by Assumption 4.4 $\hat{D}_n(\theta_n') \xrightarrow{P_3} D$, for almost all sample paths. Set

$$g_{P_n,j}(\theta_n') = \begin{cases} c - \varphi_j^*(\xi_{n,j}(\theta_n')) - \mathbb{G}_{n,j}^b(\theta_n'), & j = 1, \ldots, J, \\ 1, & j = J+1, \ldots, J+2d, \\ 0, & j = J+2d+1, J+2d+2, \end{cases} \tag{E.151}$$

and note that $|\varphi_j^*(\xi_{n,j}(\theta_n'))| < \eta$ for all $j \in \mathcal{J}^*$, and $\mathbb{G}_{n,j}^b(\theta_n') | \{X_i\}_{i=1}^\infty \xrightarrow{d} N(0, \Omega)$. Then one can mimic the argument following (E.137) to conclude (E.132).

The results in (E.135)-(E.136) follow by similar arguments, with proper redefinition of $\tau$ in equation (E.138). □

Lemma E.7: *Let Assumptions 4.3 and 4.5 hold. Let $(P_n, \theta_n)$ have the almost sure representations given in Lemma E.1, let $\mathcal{J}^*$ be defined as in (E.29), and assume that $\mathcal{J}^* \neq \varnothing$. 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 $\mathcal{C}$ equal the first element of $\widetilde{\mathcal{C}}$ s.t. $\det K^C \neq 0$ and $\lambda^C = (K^C)^{-1} g^C \in \mathfrak{W}^{*,-\delta}(0)$ if such an element exists; else, let $\mathcal{C} = \{J + 1, ..., J + d\}$ and $\lambda^C = \mathbf{1}_d$, where $\mathbf{1}_d$ denotes a $d$ vector with each entry equal to 1, and $K$, $g$ and $\mathfrak{W}^{*,-\delta}$ are as defined in Lemma E.2. Then, for any $\eta > 0$, there exist $0 < \varepsilon_\eta < \infty$ and $N \in \mathbb{N}$ s.t. $n \geqslant N$ implies*

$$\mathbf{P}\left(\mathfrak{W}^{*,-\delta}(0) \neq \varnothing, \left|\det K^{\mathcal{C}}\right| \leqslant \varepsilon_\eta\right) \leqslant \eta. \tag{E.152}$$

*Proof.* We bound the probability in (E.152) as follows:

$$\begin{aligned} \mathbf{P}\left(\mathfrak{W}^{*,-\delta}(0) \neq \varnothing, \left|\det K^{\mathcal{C}}\right| \leqslant \varepsilon_\eta\right) &\leqslant \mathbf{P}\left(\exists C \in \widetilde{\mathcal{C}} : \lambda^C \in B^d, \left|\det K^C\right| \leqslant \varepsilon_\eta\right) && \text{(E.153)} \\ &\leqslant \sum_{C \in \widetilde{\mathcal{C}} : \left|\det K^C\right| \leqslant \varepsilon_\eta} \mathbf{P}\left(\lambda^C \in B^d\right) && \text{(E.154)} \\ &\leqslant \sum_{C \in \widetilde{\mathcal{C}} : |\alpha^C| \leqslant \varepsilon_\eta^{2/d}} \mathbf{P}\left(\lambda^C \in B^d\right), && \text{(E.155)} \end{aligned}$$

where $\alpha^C$ denote the smallest eigenvalue of $K^C K^{C\prime}$. Here, the first inequality holds because $\mathfrak{W}^{*,-\delta} \subseteq B^d$ and so the event in the first probability implies the event in the next one; the second inequality is Boolean algebra; the last inequality follows because $|\det K^C| \geqslant |\alpha^C|^{d/2}$. Noting that $\widetilde{\mathcal{C}}$ has $\binom{J+2d+2}{d}$ elements, it suffices to show that

$$\left|\alpha^C\right| \leqslant \varepsilon_\eta^{2/d} \Longrightarrow \mathbf{P}\left(\lambda^C \in B^d\right) \leqslant \overline{\eta} \equiv \frac{\eta}{\binom{J+2d+2}{d}}.$$

Thus, fix $C \in \widetilde{\mathcal{C}}$. Let $q^C$ denote the eigenvector associated with $\alpha^C$ and recall that because $K^C K^{C\prime}$ is symmetric,

$\|q^C\| = 1$. Thus the claim is equivalent to:

$$|q^{C\prime}K^C K^{C\prime}q^C| \leqslant \varepsilon_\eta^{2/d} \implies \mathbf{P}((K^C)^{-1}g^C \in \mathfrak{B}_\rho^d) \leqslant \overline{\eta}. \tag{E.156}$$

Now, if $|q^{C\prime}K^C K^{C\prime}q^C| \leqslant \varepsilon_\eta^{2/d}$ and $(K^C)^{-1}g^C \in \mathfrak{B}_\rho^d$, then the Cauchy-Schwarz inequality yields

$$\left|q^{C\prime}g_{P_n}^C\right| = \left|q^{C\prime}K^C\left(K^C\right)^{-1}g^C\right| < \sqrt{d}\varepsilon_\eta^{1/d}, \tag{E.157}$$

hence

$$\mathbf{P}((K^C)^{-1}g^C \in \mathfrak{B}_\rho^d) \leqslant \mathbf{P}\Big(|q^{C\prime}g^C| < \sqrt{d}\varepsilon_\eta^{1/d}\Big). \tag{E.158}$$

If $q^C$ assigns non-zero weight only to non-stochastic constraints, the result follows immediately. If $q^C$ assigns non-zero weight also to stochastic constraints, Assumptions 4.3 and 4.5 (iii) yield

$$\begin{aligned} &\operatorname{eig}(\tilde{\Omega}) \geqslant \omega \\ \implies &Var_{\mathbf{P}}(q^{C\prime}g^C) \geqslant \omega \\ \implies &\mathbf{P}\left(|q^{C\prime}g^C| < \sqrt{d}\varepsilon_\eta^{1/d}\right) = \mathbf{P}\left(-\sqrt{d}\varepsilon_\eta^{1/d} < q^{C\prime}g^C < \sqrt{d}\varepsilon_\eta^{1/d}\right) \\ &< \frac{2\sqrt{d}\varepsilon_\eta^{1/d}}{\sqrt{2\omega\pi}}, \end{aligned} \tag{E.159}$$

where the result in (E.159) uses that the density of a normal r.v. is maximized at the expected value. The result follows by choosing

$$\varepsilon_\eta = \left(\frac{\overline{\eta}\sqrt{2\omega\pi}}{2\sqrt{d}}\right)^d.$$

□

LEMMA E.8: *Let Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. If $J_2 \geqslant d$, then $\exists \underline{c} > 0$ s.t.*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(c_n^I(\theta) \geqslant \underline{c}) = 1.$$

*Proof.* Fix any $c \geqslant 0$ 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 $\mathbb{G}_{n,j}^b(\theta)$ on $\theta$. Also, if $C$ is an index set picking certain constraints, then $\hat{D}_n^C$ is the matrix collecting the corresponding estimated gradients, and similarly for $\mathbb{G}_n^{b,C}$.

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 $\hat{D}_P^{\bar{C}}$. (This choice ensures that a Karush-Kuhn-Tucker condition holds, justifying the step from (E.160) to (E.161) below.) Then the coverage

probability $P^*(V_n^I(\theta,c) \neq \varnothing)$ is asymptotically bounded above by

$$P^*\Big(\sup_{\lambda\in\rho B^d_{n,\rho}}\Big\{p'\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j\in\mathcal{J}^*\Big\} \geqslant 0\Big) \leqslant P^*\Big(\sup_{\lambda\in\mathbb{R}^d}\Big\{p'\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j\in\bar{C}\Big\} \geqslant 0\Big) \tag{E.160}$$

$$= P^*\Big(p'(\hat{D}^{\bar{C}}_n)^{-1}(c\mathbf{1}_d - \mathbb{G}^{b,\bar{C}}_n) \geqslant 0\Big) \tag{E.161}$$

$$= P^*\left(\frac{p'(\hat{D}^{\bar{C}}_n)^{-1}(c\mathbf{1}_d - \mathbb{G}^{b,\bar{C}}_n)}{\sqrt{p'(\hat{D}^{\bar{C}}_n)^{-1}\Omega^C_P(\hat{D}^{\bar{C}}_n)^{-1}p}} \geqslant 0\right) \tag{E.162}$$

$$= P^*\left(\frac{p'adj(\hat{D}^{\bar{C}}_n)(c\mathbf{1}_d - \mathbb{G}^{b,\bar{C}}_n)}{\sqrt{p'(adj(\hat{D}^{\bar{C}}_n)\Omega^C_P adj(\hat{D}^{\bar{C}}_n)p}} \geqslant 0\right) \tag{E.163}$$

$$= \Phi\left(\frac{p'adj(\hat{D}^{\bar{C}}_n)c\mathbf{1}_d}{\sqrt{p'(adj(\hat{D}^{\bar{C}}_n)\Omega^C_P adj(\hat{D}^{\bar{C}}_n)p}}\right) + o_{\mathcal{P}}(1) \tag{E.164}$$

$$\leqslant \Phi(d\omega^{-1/2}c) + o_{\mathcal{P}}(1). \tag{E.165}$$

Here, (E.160) removes constraints and hence enlarges the feasible set; (E.161) solves in closed form; (E.162) divides through by a positive scalar; (E.163) 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; (E.164) follows by Assumption 4.5, using that the term multiplying $\mathbb{G}^{b,\bar{C}}_n$ is $O_{\mathcal{P}}(1)$; and (E.165) uses that by Assumption 4.3, there exists a constant $\omega > 0$ that does not depend on $\theta$ such that the smallest eigenvalue of $\Omega_P$ is bounded from below by $\omega$. 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

$$P^*\Big(\sup_{\lambda\in\rho B^d_{n,\rho}}\Big\{p'\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j\in\bar{C}\cup\{J_1+J_2+1\}\Big\} \geqslant 0\Big) \tag{E.166}$$

$$\leqslant P^*\Big(\exists\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j\in\bar{C}\cup\{J_1+J_2+1\}\Big) \tag{E.167}$$

$$\leqslant P^*\Big(\sup_{\lambda\in\rho B^d_{n,\rho}}\Big\{\hat{D}_{n,J_1+1}\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j\in\bar{C}\Big\} \geqslant \inf_{\lambda\in\rho B^d_{n,\rho}}\Big\{\hat{D}_{n,J_1+1}\lambda : \hat{D}_{n,J_1+J_2+1}\lambda \leqslant c - \mathbb{G}^b_{n,J_1+J_2+1}\Big\}\Big) \tag{E.168}$$

$$= P^*\Big(\hat{D}_{n,J_1+1}\hat{D}^{\bar{C}\prime}_n(\hat{D}^{\bar{C}}_n\hat{D}^{\bar{C}\prime}_n)^{-1}(c\mathbf{1}_{\bar{d}} - \mathbb{G}^{b,\bar{C}}_n) \geqslant -c + \mathbb{G}^b_{n,J_1+J_2+1}\Big). \tag{E.169}$$

Here, the reasoning from (E.166) to (E.168) holds because we evaluate the probability of increasingly larger events; in particular, if the event in (E.168) 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 (E.169) resembles (E.162), and the argument can be concluded in analogy to (E.163)-(E.165). □

LEMMA E.9: *Let Assumptions 4.1, 4.2, 4.3-(II), 4.4, and 4.5 hold. Suppose that both $\pi_{1,j}$ and $\pi_{1,j+R_1}$ are finite, with $\pi_{1,j}$, $j = 1,\dots,J$, defined in (D.4). Let $(P_n,\theta_n)$ be the sequence satisfying the conditions of Lemma E.3. Then for any $\theta'_n \in (\theta_n + \rho/\sqrt{n}B^d)\cap\Theta$,*

*(1)* $\sigma^2_{P_n,j}(\theta'_n)/\sigma^2_{P_n,j+R_1}(\theta'_n) \to 1$ *for* $j = 1,\cdots,R_1$.

*(2)* $Corr_{P_n}(m_j(X_i,\theta'_n), m_{j+R_1}(X_i,\theta'_n)) \to -1$ *for* $j = 1, \cdots, R_1$.

*(3)* $|\mathbb{G}_{n,j}(\theta'_n) + \mathbb{G}_{n,j+R_1}(\theta'_n)| \xrightarrow{P_n} 0$, *and* $|\mathbb{G}^b_{n,j}(\theta'_n) + \mathbb{G}^b_{n,j+R_1}(\theta'_n)| \xrightarrow{P^*_n} 0$ *for almost all* $\{X_i\}_{i=1}^\infty$.

*(4)* $\rho\|D_{P_n,j+R_1}(\theta'_n) + D_{P_n,j}(\theta'_n)\| \to 0$.

*Proof.* By Lemma E.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 $\pi_{1,j}, \pi_{1,j+R_1}$ 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.**

$$\lim_{n\to\infty} \frac{\kappa_n}{\sqrt{n}} \sigma_{P_n,j}(\theta'_n) > 0, \tag{E.170}$$

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

$$\frac{\sigma^2_{P_n,j+R_1}(\theta'_n)}{\sigma^2_{P_n,j}(\theta'_n)} = \frac{\sigma^2_{P_n,j}(\theta'_n) + Var_{P_n}(t_j(X_i,\theta'_n)) + 2Cov_{P_n}(m_j(X_i,\theta'_n), t_j(X_i,\theta'_n))}{\sigma^2_{P_n,j}(\theta'_n)} \to 1, \tag{E.171}$$

where the convergence follows because $Var_{P_n}(t_j(X_i,\theta'_n))$ is bounded due to Assumption 4.3-(II),

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

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

**Case 2.**

$$\lim_{n\to\infty} \frac{\kappa_n}{\sqrt{n}} \sigma_{P_n,j}(\theta'_n) = 0. \tag{E.172}$$

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 (E.171), it also follows that

$$\lim_{n\to\infty} \frac{\kappa_n}{\sqrt{n}} \sigma_{P_n,j+R_1}(\theta'_n) = 0, \tag{E.173}$$

and hence that $E_{P_n}(t_j(X_i,\theta'_n)) \to 0$. We then have, using Assumption 4.3-(II) again,

$$\begin{aligned} Var_{P_n}(t_j(X_i,\theta'_n)) &= \int t_j(x,\theta'_n)^2 dP_n(x) - E_{P_n}[t_j(X_i,\theta'_n)]^2 \\ &\leqslant M \int t_j(x,\theta'_n) dP_n(x) - E_{P_n}[t_j(X_i,\theta'_n)]^2 \to 0. \end{aligned} \tag{E.174}$$

Hence,

$$\begin{aligned} \frac{\sigma^2_{P_n,j+R_1}(\theta'_n)}{\sigma^2_{P_n,j}(\theta'_n)} &= \frac{\sigma^2_{P_n,j}(\theta'_n) + Var_{P_n}(t_j(X_i,\theta'_n)) + 2Cov_{P_n}(m_j(X_i,\theta'_n), t_j(X_i,\theta'_n))}{\sigma^2_{P_n,j}(\theta'_n)} \\ &\leqslant \frac{\sigma^2_{P_n,j}(\theta'_n) + Var_{P_n}(t_j(X_i,\theta'_n))}{\sigma^2_{P_n,j}(\theta'_n)} + \frac{2(Var_{P_n}(t_j(X_i,\theta'_n)))^{1/2}}{\sigma_{P_n,j}(\theta'_n)} \\ &\to 1, \end{aligned} \tag{E.175}$$

and the first claim follows.

To obtain claim 2, note that

$$Corr_{P_n}(m_j(X_i,\theta_n'),m_{j+R_1}(X_i,\theta_n')) = \frac{-\sigma^2_{P_n,j}(\theta_n') - Cov_{P_n}(m_j(X_i,\theta_n'),t_j(X_i,\theta_n'))}{\sigma_{P_n,j}(\theta_n')\sigma_{P_n,j+R_1}(\theta_n')}$$
$$\to -1, \tag{E.176}$$

where the result follows from (E.174) and (E.175).

To establish Claim 3, consider $\mathbb{G}_n$ below. Note that, for $j = 1, \cdots, R_1$,

$$\begin{bmatrix} \mathbb{G}_{n,j}(\theta_n') \\ \mathbb{G}_{n,j+R_1}(\theta_n') \end{bmatrix} = \begin{bmatrix} \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n (m_j(X_i,\theta_n') - E_{P_n}[m_j(X_i,\theta_n')])}{\sigma_{P_n,j}(\theta_n')} \\ -\frac{\sum_{i=1}^n (m_j(X_i,\theta_n') - E_{P_n}[m_j(X_i,\theta_n')]) + \frac{1}{\sqrt{n}}\sum_{i=1}^n (t_j(X_i,\theta_n') - E_{P_n}[t_j(X_i,\theta_n')])}{\sigma_{P_n,j+R_1}(\theta_n')} \end{bmatrix}. \tag{E.177}$$

Under the conditions of Case 1 above, we immediately obtain

$$|\mathbb{G}_{n,j}(\theta_n') + \mathbb{G}_{n,j+R_1}(\theta_n')| \xrightarrow{P_n} 0. \tag{E.178}$$

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, (E.178) obtains again. These results imply that $\mathbb{Z}_j + \mathbb{Z}_{j+R_1} = 0, a.s.$ By Lemma E.15, $\{\mathbb{G}_n^b\}$ converges in law to the same limit as $\{\mathbb{G}_n\}$ for almost all sample paths $\{X_i\}_{i=1}^\infty$. This and (E.178) then imply the second half of Claim 3.

To establish Claim 4, finiteness of $\pi_{1,j}$ and $\pi_{1,j+R_1}$ implies that

$$E_{P_n}\left(\frac{m_j(X,\theta_n')}{\sigma_{P_n,j}(\theta_n')} + \frac{m_{j+R_1}(X,\theta_n')}{\sigma_{P_n,j+R_1}(\theta_n')}\right) = O_{\mathcal{P}}\left(\frac{\kappa_n}{\sqrt{n}}\right). \tag{E.179}$$

Define the $1 \times d$ vector

$$q_n \equiv D_{P_n,j+R_1}(\theta_n') + D_{P_n,j}(\theta_n'). \tag{E.180}$$

Suppose by contradiction that

$$\rho q_n \to \varsigma \neq 0,$$

where $\|\varsigma\|$ might be infinite. Write

$$\tilde{r}_n = \frac{q_n'}{\|q_n\|}. \tag{E.181}$$

Let

$$r_n = \tilde{r}_n \rho \kappa_n^2/\sqrt{n}. \tag{E.182}$$

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 $\theta_n'$ and $\theta_n' + r_n$) we obtain

$$\begin{aligned} &E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)} + \frac{m_{j+R_1}(X,\theta_n'+r_n)}{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}\right) = E_{P_n}\left(\frac{m_j(X,\theta_n')}{\sigma_{P_n,j}(\theta_n')} + \frac{m_{j+R_1}(X,\theta_n')}{\sigma_{P_n,j+R_1}(\theta_n')}\right) + \left(D_{P_n,j}(\bar{\theta}_n) + D_{P_n,j+R_1}(\tilde{\theta}_n)\right) r_n \\ =&O_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}}) + \left(D_{P_n,j}(\theta_n') + D_{P_n,j+R_1}(\theta_n')\right) r_n + \left(D_{P_n,j}(\bar{\theta}_n) - D_{P_n,j}(\theta_n')\right) r_n + \left(D_{P_n,j+R_1}(\tilde{\theta}_n) - D_{P_n,j+R_1}(\theta_n')\right) r_n \\ =&O_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}}) + \frac{\rho\kappa_n^2}{\sqrt{n}} + O_{\mathcal{P}}(\frac{\rho^2\kappa_n^4}{n}). \end{aligned} \tag{E.183}$$

It then follows that there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$, the right hand side in (E.183) is strictly greater

than zero.

Next, observe that

$$\begin{aligned}
&E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+R_1}(X,\theta_n'+r_n)}{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}\right)\\
=&E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+R_1}(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}\right)-\left(\frac{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}-1\right)\frac{E_{P_n}(m_{j+R_1}(X,\theta_n'+r_n))}{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}\\
=&E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+R_1}(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}\right)-o_{\mathcal{P}}(\frac{\rho\kappa_n^2}{\sqrt{n}}).
\end{aligned}\tag{E.184}$$

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 4.3-(II), we have

$$\frac{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}-1=\frac{\sigma_{P_n,j+R_1}(\theta_n')}{\sigma_{P_n,j}(\theta_n')}-1+o_{\mathcal{P}}(1)=o_{\mathcal{P}}(1),\tag{E.185}$$

where we used Claim 1. Second, using Assumption 4.4, we have that

$$\frac{E_{P_n}(m_{j+R_1}(X,\theta_n'+r_n))}{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}=\frac{E_{P_n}(m_{j+R_1}(X,\theta_n'))}{\sigma_{P_n,j+R_1}(\theta_n')}+D_{P_n,j+R_1}(\tilde{\theta}_n)r_n=O_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}})+O_{\mathcal{P}}(\frac{\rho\kappa_n^2}{\sqrt{n}}).\tag{E.186}$$

The product of (E.185) and (E.186) is therefore $o_{\mathcal{P}}(\frac{\rho\kappa_n^2}{\sqrt{n}})$ and (E.184) follows.

To conclude the argument, note that for $n$ large enough, $m_{j+R_1}(X,\theta_n'+r_n)\leqslant -m_j(X,\theta_n'+r_n)$ *a.s.* because for any $\theta_n\in\Theta_I(P_n)$ and $\theta_n'\in(\theta_n+\rho/\sqrt{n}B^d)\cap\Theta$ for $n$ large enough, $\theta_n'+r_n\in\Theta^\epsilon$ and Assumption 4.3-(II) applies. Therefore, there exists $N\in\mathbb{N}$ such that for all $n\geqslant N$, the left hand side in (E.183) is strictly less than the right hand side, yielding a contradiction. □

Below, we let $\mathcal{R}_1=\{1,\cdots,R_1\}$ and $\mathcal{R}_2=\{R_1+1,\cdots,2R_1\}$.

LEMMA E.10: *Suppose Assumptions 4.1, 4.2, and 4.5 hold. For each $\theta\in\Theta$, let $\eta_{n,j}(\theta)=\sigma_{P,j}(\theta)/\hat{\sigma}_{n,j}(\theta)-1$. Then, (i) for each $j=1,\ldots,J_1+J_2$*

$$\inf_{P\in\mathcal{P}}P\Big(\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)|\to 0\Big)=1.\tag{E.187}$$

*(ii) For any $j=1,\ldots,R_1$ let*

$$\hat{\sigma}_{n,j}^M(\theta)=\hat{\sigma}_{n,j+R_1}^M(\theta)\equiv\hat{\mu}_{n,j}(\theta)\hat{\sigma}_{n,j}(\theta)+(1-\hat{\mu}_{n,j}(\theta))\hat{\sigma}_{n,j+R_1}(\theta).\tag{E.188}$$

*Let $(P_n,\theta_n)$ be a sequence such that $P_n\in\mathcal{P}$, $\theta_n\in\Theta$ for all $n$, and $\kappa_n^{-1}\sqrt{n}\gamma_{1,P_n,j}(\theta_n)\to\pi_{1j}\in\mathbb{R}_{[-\infty]}$. Let $\mathcal{J}^*$ be defined as in (E.29). Then, for any $\eta>0$, there exists $N\in\mathbb{N}$ such that*

$$P_n\Big(\max_{j\in(\mathcal{R}_1\cup\mathcal{R}_2)\cap\mathcal{J}^*}\Big|\frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}_{n,j}^M(\theta_n)}-1\Big|>\eta\Big)<\eta\tag{E.189}$$

*for all $n\geqslant N$.*

*Proof.* We first show that, for any $\epsilon>0$ and for any $j=1,\ldots,J_1+J_2$,

$$\inf_{P\in\mathcal{P}}P\Big(\sup_{m\geqslant n}\sup_{\theta\in\Theta}\Big|\frac{\hat{\sigma}_{n,j}(\theta)}{\sigma_{P,j}(\theta)}-1\Big|\leqslant\epsilon\Big)\to 1.\tag{E.190}$$

For this, define the following sets:

$$\mathcal{M}_j \equiv \{m_j(\cdot,\theta)/\sigma_{P,j}(\theta) : \theta \in \Theta, P \in \mathcal{P}\} \tag{E.191}$$

$$\mathcal{S}_j \equiv \{(m_j(\cdot,\theta)/\sigma_{P,j}(\theta))^2 : \theta \in \Theta, P \in \mathcal{P}\}. \tag{E.192}$$

By Assumptions 4.1-(a), 4.1 (iv), 4.5 (i), (iii), and arguing as in the proof of Lemma D.2.2 (and D.2.1) in Bugni, Canay, and Shi (2015), it follows that $\mathcal{S}_j$ and $\mathcal{M}_j$ are Glivenko-Cantelli (GC) classes uniformly in $P \in \mathcal{P}$ (in the sense of van der Vaart and Wellner, 2000, page 167).

Therefore, for any $\epsilon > 0$,

$$\inf_{P\in\mathcal{P}} P\Big( \sup_{m\geqslant n} \sup_{\theta\in\Theta} \Big| \frac{n^{-1}\sum_{i=1}^n m_j(X_i,\theta)^2}{\sigma^2_{P,j}(\theta)} - \frac{E_P[m_j(X,\theta)^2]}{\sigma^2_{P,j}(\theta)} \Big| \leqslant \epsilon \Big) \to 1 \tag{E.193}$$

$$\inf_{P\in\mathcal{P}} P\Big( \sup_{m\geqslant n} \sup_{\theta\in\Theta} \Big| \frac{\bar{m}_{n,j}(\theta) - E_P[m_j(X,\theta)]}{\sigma_{P,j}(\theta)} \Big| \leqslant \epsilon \Big) \to 1. \tag{E.194}$$

Note that, by Assumption 4.1 (iv), $|E_P[m_j(X,\theta)]/\sigma_{P,j}(\theta)| \leqslant M$ for some constant $M > 0$ that does not depend on $P$ and $(x^2 - y^2) \leqslant |x+y||x-y| \leqslant 2M|x-y|$ for all $x, y \in [-M, M]$. By (E.194), for any $\epsilon > 0$, it follows that

$$\inf_{P\in\mathcal{P}} P\Big( \sup_{m\geqslant n} \sup_{\theta\in\Theta} \Big| \frac{\bar{m}_{n,j}(\theta)^2 - E_P[m_j(X,\theta)]^2}{\sigma^2_{P,j}(\theta)} \Big| \leqslant \epsilon \Big) \to 1. \tag{E.195}$$

By the uniform continuity of $x \mapsto \sqrt{x}$ on $\mathbb{R}_+$, for any $\epsilon > 0$, there is a constant $\eta > 0$ such that

$$\Big| \frac{\hat{\sigma}^2_{n,j}(\theta)}{\sigma^2_{P,j}(\theta)} - 1 \Big| \leqslant \eta \;\Rightarrow \Big| \frac{\hat{\sigma}_{n,j}(\theta)}{\sigma_{P,j}(\theta)} - 1 \Big| \leqslant \epsilon. \tag{E.196}$$

By the definition of $\sigma^2_{P,j}(\theta)$ and the triangle inequality,

$$\Big| \frac{\hat{\sigma}^2_{n,j}(\theta)}{\sigma^2_{P,j}(\theta)} - 1 \Big| \leqslant \Big| \frac{n^{-1}\sum_{i=1}^n m(X_i,\theta)^2 - E[m_j(X_i,\theta)^2]}{\sigma^2_{P,j}(\theta)} \Big| + \Big| \frac{\bar{m}_{n,j}(\theta)^2 - E[m_j(X_i,\theta)]^2}{\sigma^2_{P,j}(\theta)} \Big|. \tag{E.197}$$

By (E.196)-(E.197), bounding each of the terms on the right hand side of (E.197) by $\eta/2$ implies $|\hat{\sigma}_{n,j}(\theta)/\sigma_{P,j}(\theta) - 1| \leqslant \epsilon$. This, together with (E.193) and (E.195), ensures that, for any $\epsilon > 0$, (E.190) holds.

Note that $|\hat{\sigma}_{n,j}(\theta)/\sigma_{P,j}(\theta) - 1| \leqslant \epsilon$ implies $\hat{\sigma}_{n,j}(\theta) > 0$, and argue as in the proof of Lemma D.2.4 in Bugni, Canay, and Shi (2015) to conclude that

$$\inf_{P\in\mathcal{P}} P\Big( \sup_{m\geqslant n} \sup_{\theta\in\Theta} \Big| \frac{\sigma_{P,j}(\theta)}{\hat{\sigma}_{n,j}(\theta)} - 1 \Big| \leqslant \epsilon \Big) \to 1. \tag{E.198}$$

Finally, recall that $\eta_{n,j}(\theta) = \sigma_{P,j}(\theta)/\hat{\sigma}_{n,j}(\theta) - 1$ and note that for any $\epsilon > 0$,

$$\begin{aligned}
1 &= \lim_{n\to\infty} \inf_{P\in\mathcal{P}} P\Big( \sup_{m\geqslant n} \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant \epsilon \Big) \\
&\leqslant \inf_{P\in\mathcal{P}} \lim_{n\to\infty} P\Big( \bigcap_{m\geqslant n} \big\{ \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant \epsilon \big\} \Big) \\
&= \inf_{P\in\mathcal{P}} P\Big( \lim_{n\to\infty} \bigcap_{m\geqslant n} \{ \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant \epsilon \} \Big) \\
&= \inf_{P\in\mathcal{P}} P\Big( \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant \epsilon, \text{ for almost all } n \Big),
\end{aligned} \tag{E.199}$$

where the second equality is due to the continuity of probability with respect to monotone sequences. Therefore, the first conclusion of the lemma follows.

(ii) We first give the limit of $\hat{\mu}_{n,j}(\theta_n)$. Recall the definitions of $\hat{\mu}_{n,j+R_1}$ and $\hat{\mu}_{n,j}(\theta_n)$ in (E.14)-(E.15). Note that

$$\begin{aligned}
&\sup_{\theta_n' \in \theta_n + \rho/\sqrt{n}B^d} \left| \kappa_n^{-1} \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n')}{\hat{\sigma}_{n,j}(\theta_n')} - \kappa_n^{-1} \frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')} \right| \\
&\leqslant \sup_{\theta_n' \in \theta_n + \rho/\sqrt{n}B^d} \left| \kappa_n^{-1} \frac{\sqrt{n}(\bar{m}_{n,j}(\theta_n') - E_{P_n}[m_j(X_i,\theta_n')])}{\sigma_{n,j}(\theta_n')}(1+\eta_{n,j}(\theta_n')) + \kappa_n^{-1} \frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')}\eta_{n,j}(\theta_n') \right| \\
&\leqslant \sup_{\theta_n' \in \theta_n + \rho/\sqrt{n}B^d} |\kappa_n^{-1}\mathbb{G}_n(\theta_n')(1+\eta_{n,j}(\theta_n'))| + \left| \frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\kappa_n \sigma_{P_n,j}(\theta_n')}\eta_{n,j}(\theta_n') \right| = o_{\mathcal{P}}(1), \quad \text{(E.200)}
\end{aligned}$$

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 $\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)| = o_{\mathcal{P}}(1)$ by part (i) of this Lemma. Hence,

$$\hat{\mu}_{n,j}(\theta_n) \xrightarrow{P_n} 1 - \min\Big\{ \max(0, \frac{\pi_{1,j}}{\pi_{1,j+R_1} + \pi_{1,j}}), 1 \Big\}, \quad \text{(E.201)}$$

unless $\pi_{1,j+R_1} + \pi_{1,j} = 0$ (this case is considered later). This implies that if $\pi_{1,j} \in (-\infty, 0]$ and $\pi_{1,j+R_1} = -\infty$, one has

$$\hat{\mu}_{n,j}(\theta_n) \xrightarrow{P_n} 1. \quad \text{(E.202)}$$

Similarly, if $\pi_{1,j} = -\infty$ and $\pi_{1,j+R_1} \in (-\infty, 0]$, one has

$$\hat{\mu}_{n,j+R_1}(\theta_n) \xrightarrow{P_n} 1. \quad \text{(E.203)}$$

Now, one may write

$$\frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1 = \frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}_{n,j}(\theta_n)}\Big(\frac{\hat{\sigma}_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1\Big) + \Big(\frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}_{n,j}(\theta_n)} - 1\Big) = O_{P_n}(1)\Big(\frac{\hat{\sigma}_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1\Big) + o_{P_n}(1), \quad \text{(E.204)}$$

where the second equality follows from the first conclusion of the lemma. Hence, for the second conclusion of the lemma, it suffices to show $\hat{\sigma}_{n,j}(\theta_n)/\hat{\sigma}^M_{n,j}(\theta_n) - 1 = o_{\mathcal{P}}(1)$. For this, we consider three cases.

Suppose first $j \in \mathcal{R}_1 \cap \mathcal{J}^*$ and $j + R_1 \notin \mathcal{J}^*$. Then, $\pi^*_{1,j} = 0$ and $\pi^*_{1,j+R_1} = -\infty$. Then,

$$\begin{aligned}
\hat{\sigma}^M_{n,j}(\theta_n) &= \hat{\mu}_{n,j}(\theta_n)\hat{\sigma}_{n,j}(\theta_n) + (1 - \hat{\mu}_{n,j}(\theta_n))\hat{\sigma}_{n,j+R_1}(\theta_n) \quad \text{(E.205)} \\
&= (1 + o_{P_n}(1))\hat{\sigma}_{n,j}(\theta_n) + (1 - \hat{\mu}_{n,j}(\theta_n))O_{P_n}(\hat{\sigma}_{n,j}(\theta_n)), \quad \text{(E.206)}
\end{aligned}$$

where the second equality follows from (E.202) and the fact that

$$\begin{aligned}
\hat{\sigma}_{n,j+R_1}(\theta_n) &= \Big(\hat{\sigma}^2_{n,j}(\theta_n) + 2\widehat{Cov}_n(m_j(X_i,\theta), t_j(X_i,\theta)) + \widehat{Var}_n(t_j(X_i,\theta))\Big)^{1/2} \\
&= \Big(\hat{\sigma}^2_{n,j}(\theta_n) + O_{P_n}(\hat{\sigma}_{n,j}(\theta_n)) + O_{P_n}(1)\Big)^{1/2} = O_{P_n}(\hat{\sigma}_{n,j}(\theta_n)), \quad \text{(E.207)}
\end{aligned}$$

where the second equality follows from, $Var_{P_n}(t_j(X_i,\theta))$ being bounded by Assumption 4.3-(II) and

$$\begin{aligned}
\widehat{Var}_n(t_j(X_i,\theta)) &= Var_{P_n}(t_j(X_i,\theta)) + o_{P_n}(1) \quad \text{(E.208)} \\
\widehat{Cov}_n(m_j(X_i,\theta), t_j(X_i,\theta)) &\leqslant \hat{\sigma}_{n,j}(\theta_n)\widehat{Var}_n(t_j(X_i,\theta))^{1/2}, \quad \text{(E.209)}
\end{aligned}$$

where the last inequality is due to the Cauchy-Schwarz inequality.

Therefore,

$$\frac{\hat{\sigma}_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1 = \frac{\hat{\sigma}_{n,j}(\theta_n) - \hat{\sigma}^M_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} = \frac{(1-\hat{\mu}_{n,j}(\theta_n))O_{P_n}(\hat{\sigma}_{n,j}(\theta_n))}{(1+o_{P_n}(1))\hat{\sigma}_{n,j}(\theta_n) + (1-\hat{\mu}_{n,j}(\theta_n))O_{P_n}(\hat{\sigma}_{n,j}(\theta_n))} = o_{P_n}(1), \tag{E.210}$$

where we used $\hat{\sigma}^{-1}_{n,j}(\theta_n) = O_{P_n}(1)$ by equation (4.3) and part (i) of the lemma. By (E.204) and (E.210), $\sigma_{P_n,j}(\theta_n)/\hat{\sigma}^M_{n,j}(\theta_n) - 1 = o_{P_n}(1)$. Using a similar argument, the same conclusion follows when $j \in \mathcal{R}_1, j \notin \mathcal{J}^*$, but $j + R_1 \in \mathcal{R}_2 \cap \mathcal{J}^*$.

Now consider the case $j \in \mathcal{R}_1 \cap \mathcal{J}^*$ and $j + R_1 \in \mathcal{R}_2 \cap \mathcal{J}^*$. Then, $\pi^*_{1,j} = 0$ and $\pi^*_{1,j+R_1} = 0$. In this case, $\hat{\mu}_{n,j}(\theta_n) \in [0,1]$ for all $n$ and by Lemma E.9 (1),

$$\left|\frac{\sigma_{P_n,j}(\theta_n)}{\sigma_{P_n,j+R_1}(\theta_n)} - 1\right| = o_{P_n}(1), \quad \text{for } j = 1, \cdots, R_1, \tag{E.211}$$

and therefore,

$$\begin{aligned}
\frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1 &= \frac{\sigma_{P_n,j}(\theta_n) - \hat{\sigma}^M_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} \\
&= \frac{[\hat{\mu}_{n,j}(\theta_n) + (1-\hat{\mu}_{n,j}(\theta_n))]\sigma_{P_n,j}(\theta_n) - [\hat{\mu}_{n,j}(\theta_n)\hat{\sigma}_{n,j}(\theta_n) + (1-\hat{\mu}_{n,j}(\theta_n))\hat{\sigma}_{n,j+R_1}(\theta_n)]}{\hat{\sigma}^M_{n,j}(\theta_n)} \\
&= \frac{\hat{\mu}_{n,j}(\theta_n)[\sigma_{P_n,j}(\theta_n) - \hat{\sigma}_{n,j}(\theta_n)]}{\hat{\sigma}^M_{n,j}(\theta_n)} + \frac{(1-\hat{\mu}_{n,j}(\theta_n))[\sigma_{P_n,j+R_1}(\theta_n) - \hat{\sigma}_{n,j+R_1}(\theta_n) + o_{P_n}(1)]}{\hat{\sigma}^M_{n,j}(\theta_n)},
\end{aligned} \tag{E.212}$$

where the second equality follows from the definition of $\hat{\sigma}^M_{n,j}(\theta_n)$, and the third equality follows from (E.211) and $\sigma_{P_n,j+R_1}$ bounded away from 0 due to (4.3). Note that

$$\frac{\hat{\mu}_{n,j}(\theta_n)[\sigma_{P_n,j}(\theta_n) - \hat{\sigma}_{n,j}(\theta_n)]}{\hat{\sigma}^M_{n,j}(\theta_n)} = \hat{\mu}_{n,j}(\theta_n)\frac{\hat{\sigma}_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)}\left(\frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}_{n,j}(\theta_n)} - 1\right) = o_{P_n}(1), \tag{E.213}$$

where the second equality follows from the first conclusion of the lemma. Similarly,

$$\begin{aligned}
&\frac{(1-\hat{\mu}_{n,j}(\theta_n))[\sigma_{P_n,j+R_1}(\theta_n) - \hat{\sigma}_{n,j+R_1}(\theta_n) + o_{P_n}(1)]}{\hat{\sigma}^M_{n,j}(\theta_n)} \\
&\qquad = (1-\hat{\mu}_{n,j}(\theta_n))\frac{\hat{\sigma}_{n,j+R_1}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)}\left(\frac{\sigma_{P_n,j+R_1}(\theta_n)}{\hat{\sigma}_{n,j+R_1}(\theta_n)} - 1 + o_{P_n}(1)\right) = o_{P_n}(1).
\end{aligned} \tag{E.214}$$

By (E.212)-(E.214), it follows that $\sigma_{P_n,j}(\theta_n)/\hat{\sigma}^M_{n,j}(\theta_n) - 1 = o_{P_n}(1)$. Therefore, the second conclusion holds for all subcases. □

## E.2 Lemmas Used to Prove Theorem B.1

Let $\{X^b_i\}_{i=1}^n$ denote a bootstrap sample drawn randomly from the empirical distribution. Define

$$\begin{aligned}
\mathfrak{G}^b_{n,j}(\theta) &\equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(m_j(X^b_i,\theta) - \bar{m}_n(\theta)\right)/\sigma_{P,j}(\theta) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n (M_{n,i}-1)m_j(X_i,\theta)/\sigma_{P,j}(\theta),
\end{aligned} \tag{E.215}$$

where $\{M_{n,i}\}_{i=1}^n$ denotes the multinomial weights on the original sample, and we let $P^*_n$ denote the conditional distribution of $\{M_{n,i}\}_{i=1}^n$ given the sample path $\{X_i\}_{i=1}^\infty$ (see Appendix E.3 for details on the construction of the bootstrapped empirical process).

LEMMA E.11: *(i) Let* $\mathcal{M}_P \equiv \{f : \mathcal{X} \to \mathbb{R} : f(\cdot) = \sigma_{P,j}(\theta)^{-1} m_j(\cdot,\theta), \theta \in \Theta, j = 1,\cdots,J\}$ *and let* $F$ *be its envelope. Suppose that (i) there exist constants* $K, v > 0$ *that do not depend on* $P$ *such that*

$$\sup_Q N(\epsilon \|F\|_{L^2_Q}, \mathcal{M}_P, L^2_Q) \leqslant K\epsilon^{-v}, \ 0 < \epsilon < 1, \tag{E.216}$$

*where the supremum is taken over all discrete distributions; (ii) There exists a positive constant* $\gamma > 0$ *such that*

$$\|(\theta_1, \tilde{\theta}_1) - (\theta_2, \tilde{\theta}_2)\| \leqslant \delta \quad \Rightarrow \quad \sup_{P \in \mathcal{P}} \|Q_P(\theta_1, \tilde{\theta}_1) - Q_P(\theta_2, \tilde{\theta}_2)\| \leqslant M\delta^\gamma. \tag{E.217}$$

*Let* $\delta_n$ *be a positive sequence tending to 0 and let* $\epsilon_n$ *be a positive sequence such that* $\epsilon_n / |\delta_n^\gamma \ln \delta_n| \to \infty$ *as* $n \to \infty$. *Then,*

$$\sup_{P \in \mathcal{P}} P\left( \sup_{\|\theta - \theta'\| \leqslant \delta_n} \|\mathbb{G}_n(\theta) - \mathbb{G}_n(\theta'))\| > \epsilon_n \right) = o(1). \tag{E.218}$$

*Further,*

$$\lim_{n \to \infty} P_n^* \left( \sup_{\|\theta - \theta'\| \leqslant \delta_n} \|\mathfrak{G}_n^b(\theta) - \mathfrak{G}_n^b(\theta'))\| > \epsilon_n | \{X_i\}_{i=1}^\infty \right) = 0. \tag{E.219}$$

*for almost all sample paths* $\{X_i\}_{i=1}^\infty$ *uniformly in* $P \in \mathcal{P}$.

*Proof.* For the first conclusion of the lemma, it suffices to show that there is a sequence $\{\epsilon_n\}$ such that, uniformly in $P$:

$$P\left( \sup_{\|\theta - \theta'\| \leqslant \delta_n} \max_{j=1,\cdots,J} |\mathbb{G}_{n,j}(\theta) - \mathbb{G}_{n,j}(\theta')| > \epsilon_n \right) = o(1). \tag{E.220}$$

For this purpose, we mostly mimic the argument required to show the stochastic equicontinuity of empirical processes (see e.g. van der Vaart and Wellner, 2000, Ch.2.5). Before doing so, note that, arguing as in the proof of Lemma D.1 (Part 1) in Bugni, Canay, and Shi (2015), one has

$$\|\theta - \theta'\| \leqslant \delta_n \quad \Rightarrow \quad \varrho_P(\theta, \theta') \leqslant \tilde{\delta}_n, \tag{E.221}$$

where $\tilde{\delta}_n = O(\delta_n^\gamma)$ by assumption. Define

$$\mathcal{M}_{P,\tilde{\delta}_n} = \{\sigma_{P,j}(\theta)^{-1} m_j(\cdot,\theta) - \sigma_{P,j}(\theta')^{-1} m_j(\cdot,\theta') | \theta, \theta' \in \Theta, \varrho_P(\theta, \tilde{\theta}) < \tilde{\delta}_n, j = 1,\cdots,J\}. \tag{E.222}$$

Define $Z_n(\tilde{\delta}_n) \equiv \sup_{f \in \mathcal{M}_{\tilde{\delta}_n}} |\sqrt{n}(\mathbb{P}_n - P)f|$. Then, by (E.221), one has

$$P\left( \sup_{\|\theta - \theta'\| \leqslant \delta_n} \max_{j=1,\cdots,J} |\mathbb{G}_{n,j}(\theta) - \mathbb{G}_{n,j}(\theta'))| > \epsilon_n \right) \leqslant P(Z_n(\tilde{\delta}_n) > \epsilon_n \Big). \tag{E.223}$$

From here, we deal with the supremum of empirical processes though symmetrization and an application of a maximal inequality. By Markov's inequality and Lemma 2.3.1 (symmetrization lemma) in van der Vaart and Wellner (2000), one has

$$P(Z_n(\tilde{\delta}_n) > \epsilon_n) \leqslant \frac{2}{\epsilon_n} E_{P \times P^W} \left[ \sup_{f \in \mathcal{M}_{P,\tilde{\delta}_n}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i f(X_i) \right| \right], \tag{E.224}$$

where $\{W_i\}_{i=1}^n$ are i.i.d. Rademacher random variables independent of $\{X_i\}_{i=1}^\infty$ whose law is denoted by $P^W$. Now,

fix the sample path $\{X_i\}_{i=1}^n$, and let $\hat{P}_n$ be the empirical distribution. By Hoeffding's inequality, the stochastic process $f \mapsto \{n^{-1/2}\sum_{i=1}^n W_i f(X_i)\}$ is sub-Gaussian for the $L^2_{\hat{P}_n}$ seminorm $\|f\|_{L^2_{\hat{P}_n}} = (n^{-1}\sum_{i=1}^n f(X_i)^2)^{1/2}$. By the maximal inequality (Corollary 2.2.8) and arguing as in the proof of Theorem 2.5.2 in in van der Vaart and Wellner (2000), one then has

$$\begin{aligned} E_{P^W}\left[\sup_{f\in\mathcal{M}_{\tilde{\delta}_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i f(X_i)\right|\right] &\leqslant K\int_0^{\tilde{\delta}_n}\sqrt{\ln N(\epsilon,\mathcal{M}_{P,\tilde{\delta}_n},L^2_{\hat{P}_n})}d\epsilon \\ &\leqslant K\int_0^{\tilde{\delta}_n/\|F\|_{L^2_{\hat{Q}}}}\sup_Q\sqrt{\ln N(\epsilon\|F\|_{L^2_{\hat{Q}}},\mathcal{M}_P,L^2_Q)}d\epsilon \\ &\leqslant K'\int_0^{\tilde{\delta}_n/\|F\|_{L^2_{\hat{Q}}}}\sqrt{-v\ln\epsilon}d\epsilon, \end{aligned} \tag{E.225}$$

for some $K' > 0$, where the last inequality follows from (E.216). Note that $\sqrt{-\ln\epsilon} \leqslant -\ln\epsilon$ for $\epsilon \leqslant \tilde{\delta}_n/\|F\|_{L^2_{\hat{Q}}}$ with $n$ sufficiently large. Hence,

$$E_{P^W}\left[\sup_{f\in\mathcal{M}_{\tilde{\delta}_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i f(X_i)\right|\right] \leqslant K'v^{1/2}\int_0^{\tilde{\delta}_n/\|F\|_{L^2_{\hat{Q}}}}(-\ln\epsilon)d\epsilon = K'v^{1/2}(\tilde{\delta}_n - \tilde{\delta}_n\ln(\tilde{\delta}_n)). \tag{E.226}$$

By (E.224) and taking expectations with respect to $P$ in (E.226), it follows that

$$P(Z_n(\tilde{\delta}_n) > \epsilon_n) \leqslant 2K'v^{1/2}(\tilde{\delta}_n - \tilde{\delta}_n\ln(\tilde{\delta}_n))/\epsilon_n = O(\delta_n^\gamma/\epsilon_n) + O(|\delta_n^\gamma\ln(\delta_n)|/\epsilon_n) = o(1), \tag{E.227}$$

where the last equality follows from the rate condition on $\epsilon_n$. By (E.223) and (E.227), conclude that the first claim of the lemma holds.

For the second claim, define $Z_n^*(\tilde{\delta}_n) \equiv \sup_{f\in\mathcal{M}_{\tilde{\delta}_n}}|\sqrt{n}(\hat{P}_n^* - \hat{P}_n)f|$, where $\hat{P}_n^*$ is the empirical distribution of $\{X_i^b\}_{i=1}^n$. Then, by (E.221), one has

$$P_n^*\left(\sup_{\|\theta-\theta'\|\leqslant\delta_n}\max_{j=1,\cdots,J}|\mathfrak{G}_{n,j}^b(\theta) - \mathfrak{G}_{n,j}^b(\theta')| > \epsilon_n\middle|\{X_i\}_{i=1}^\infty\right) \leqslant P_n^*\left(Z_n^*(\tilde{\delta}_n) > \epsilon_n\middle|\{X_i\}_{i=1}^\infty\right). \tag{E.228}$$

By Markov's inequality and Lemma 2.3.1 (symmetrization lemma) in van der Vaart and Wellner (2000), one has

$$P_n^*\left(Z_n^*(\tilde{\delta}_n) > \epsilon_n\middle|\{X_i\}_{i=1}^\infty\right) \leqslant \frac{2}{\epsilon_n}E_{P_n^*\times P^W}\left[\sup_{f\in\mathcal{M}_{P,\tilde{\delta}_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i f(X_i^b)\right|\middle|\{X_i\}_{i=1}^\infty\right] \tag{E.229}$$

$$= \frac{2}{\epsilon_n}E_{P_n^*}\left[E_{P^W}\left[\sup_{f\in\mathcal{M}_{P,\tilde{\delta}_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i f(X_i^b)\right|\middle|\{X_i^b\},\{X_i\}_{i=1}^\infty\right]\middle|\{X_i\}_{i=1}^\infty\right], \tag{E.230}$$

where $\{W_i\}_{i=1}^n$ are i.i.d. Rademacher random variables independent of $\{X_i\}_{i=1}^\infty$ and $\{M_{n,i}\}_{i=1}^n$. Argue as in (E.224)-(E.227). Then, it follows that

$$P_n^*(Z_n^*(\tilde{\delta}_n) > \epsilon_n|\{X_i\}_{i=1}^\infty) = O(\delta_n^\gamma/\epsilon_n) + O(-\delta_n^\gamma\ln(\delta_n)/\epsilon_n) = o(1),$$

for almost all sample paths. Hence, the second claim of the lemma follows. □

LEMMA E.12: *Suppose Assumptions 4.1, 4.2, and 4.5 hold. Let* $\mathcal{S}_P \equiv \{f : \mathcal{X} \to \mathbb{R} : f(\cdot) = \sigma_{P,j}(\theta)^{-2}m_j^2(\cdot,\theta), \theta \in \Theta, j = 1,\cdots,J\}$ *and let* $F$ *be its envelope. (i) If* $\mathcal{S}_P$ *is Donsker and pre-Gaussian uniformly in* $P \in \mathcal{P}$*, then*

$$\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)|^* = O_{\mathcal{P}}(1/\sqrt{n}); \tag{E.231}$$

*(ii) If* $|\sigma_{P,j}(\theta)^{-1}m_j(x,\theta) - \sigma_{P,j}(\theta')^{-1}m_j(x,\theta')| \leqslant \bar{M}(x)\|\theta-\theta'\|$ *with* $E_P[\bar{M}(X)^2] < M$ *for all* $\theta,\theta' \in \Theta$, $x \in \mathcal{X}$, $j = 1,\cdots,J$, *and* $P \in \mathcal{P}$, *then, for any* $\eta > 0$, *there exists a constant* $C > 0$ *such that*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} P\Big( \max_{j=1,\cdots,J} \sup_{\|\theta-\theta'\|<\delta} |\eta_{n,j}(\theta) - \eta_{n,j}(\theta')| > C\delta \Big) < \eta. \tag{E.232}$$

*Proof.* We show the claim by first showing that, for any $\delta > 0$, there exist $M > 0$ and $N \in \mathbb{N}$ such that

$$\inf_{P\in\mathcal{P}} P^\infty\Big( \sup_{\theta\in\Theta} \Big| \frac{\hat{\sigma}_{n,j}(\theta)}{\sigma_{P,j}(\theta)} - 1 \Big| \leqslant M/\sqrt{n} \Big) \geqslant 1-\delta, \ \forall n \geqslant N. \tag{E.233}$$

By Assumptions 4.1 (iv), 4.5 and Theorem 2.8.2 in van der Vaart and Wellner (2000), $\mathcal{M}_P$ is a Donsker class uniformly in $P \in \mathcal{P}$. By hypothesis, $\mathcal{S}_P$ is a Donsker class uniformly in $P \in \mathcal{P}$.

Therefore, by the continuous mapping theorem, for any $\epsilon > 0$,

$$\Big| P\Big( \sqrt{n} \sup_{\theta\in\Theta} \Big| \frac{n^{-1}\sum_{i=1}^n m_j(X_i,\theta)^2}{\sigma_{P,j}^2(\theta)} - \frac{E_P[m_j(X,\theta)^2]}{\sigma_{P,j}^2(\theta)} \Big| \leqslant C_1 \Big) - \Pr(\sup_{\theta\in\Theta} |\mathbb{H}_{P,j}(\theta)| \leqslant C_1) \Big| \leqslant \epsilon \tag{E.234}$$

$$\Big| P\Big( \sqrt{n} \sup_{\theta\in\Theta} \Big| \frac{\bar{m}_{n,j}(\theta) - E_P[m_j(X,\theta)]}{\sigma_{P,j}(\theta)} \Big| \leqslant C_2 \Big) - \Pr(\sup_{\theta\in\Theta} |\mathbb{G}_{P,j}(\theta)| \leqslant C_2) \Big| \leqslant \epsilon. \tag{E.235}$$

for $n$ sufficiently large uniformly in $P \in \mathcal{P}$, where $\mathbb{H}_{P,j}$ and $\mathbb{G}_{P,j}$ are tight Gaussian processes, and $C_1$ and $C_2$ are the continuity points of the distributions of $\sup_{\theta\in\Theta}|\mathbb{H}_{P,j}(\theta)|$ and $\sup_{\theta\in\Theta}|\mathbb{G}_{P,j}(\theta)|$ respectively. As in the proof of Lemma E.10 (i), bounding each term of the right hand side of (E.197) by $C_1/\sqrt{n}$ and $C_2/\sqrt{n}$ implies that $\sup_{\theta\in\Theta}\Big|\frac{\hat{\sigma}_{n,j}^2(\theta)}{\sigma_{P,j}^2(\theta)} - 1\Big| \leqslant C/\sqrt{n}$ for some constant $C > 0$. Now choose $C_1 > 0$ and $C_2 > 0$ so that

$$\Pr(\sup_{\theta\in\Theta} |\mathbb{H}_{P,j}(\theta)| \leqslant C_1) > 1-\delta/3, \quad \text{and} \quad \Pr(\sup_{\theta\in\Theta} |\mathbb{G}_{P,j}(\theta)| \leqslant C_2) > 1-\delta/3, \tag{E.236}$$

and set $\epsilon > 0$ sufficiently small so that $1 - 2\delta/3 - 2\epsilon \geqslant 1-\delta$. The existence of such continuity points $C_1, C_2 > 0$ is due to Theorem 11.1 in Davydov, Lifshitz, and Smorodina (1995) applied to $\sup_{\theta\in\Theta}|\mathbb{H}_{P,j}(\theta)|$ and $\sup_{\theta\in\Theta}|\mathbb{G}_{P,j}(\theta)|$ respectively. Then, for sufficiently large $n$,

$$\begin{aligned} 1-\delta &\leqslant P\Big( \sqrt{n} \sup_{\theta\in\Theta} \Big| \frac{n^{-1}\sum_{i=1}^n m_j(X_i,\theta)^2}{\sigma_{P,j}^2(\theta)} - \frac{E_P[m_j(X,\theta)^2]}{\sigma_{P,j}^2(\theta)} \Big| \leqslant C_1, \\ &\qquad\qquad \sqrt{n} \sup_{\theta\in\Theta} \Big| \frac{\bar{m}_{n,j}(\theta) - E_P[m_j(X,\theta)]}{\sigma_{P,j}(\theta)} \Big| \leqslant C_2 \Big) \\ &\leqslant P\Big( \sup_{\theta\in\Theta} \Big| \frac{\hat{\sigma}_{n,j}^2(\theta)}{\sigma_{P,j}^2(\theta)} - 1 \Big| \leqslant C/\sqrt{n} \Big), \end{aligned} \tag{E.237}$$

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

Next, note that, for $x > 0$ and $0 < \eta < 1$, $|x^2 - 1| \leqslant \eta$ implies $|x-1| \leqslant 1 - (1-\eta)^{1/2} \leqslant \eta$, and hence by (E.237), for sufficiently large $n$,

$$1-\delta \leqslant P\Big( \sup_{\theta\in\Theta} \Big| \frac{\hat{\sigma}_{n,j}(\theta)}{\sigma_{P,j}(\theta)} - 1 \Big| \leqslant C/\sqrt{n} \Big), \tag{E.238}$$

uniformly in $P \in \mathcal{P}$. Finally, note again that $|\hat{\sigma}_{n,j}(\theta)/\sigma_{P,j}(\theta) - 1| \leqslant \epsilon$ implies $\hat{\sigma}_{n,j}(\theta) > 0$, and by the local Lipshitz continuity of $x \mapsto 1/x$ on a neighborhood around 1, there is a constant $C'$ such that

$$P\Big( \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant C'/\sqrt{n} \Big) \geqslant 1-\delta, \tag{E.239}$$

uniformly in $P \in \mathcal{P}$ for all $n$ sufficiently large. This establishes the first claim of the lemma.

(ii) First, consider

$$\frac{\hat{\sigma}_{n,j}^2(\theta)}{\sigma_{P,j}^2(\theta)} = n^{-1}\sum_{i=1}^n \left(\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right)^2 - \left(n^{-1}\sum_{i=1}^n \frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right)^2. \tag{E.240}$$

We claim that this function is Lipschitz with probability approaching 1. To see this, note that, for any $\theta, \theta' \in \Theta$,

$$\begin{aligned}
&\left| n^{-1}\sum_{i=1}^n \left(\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right)^2 - n^{-1}\sum_{i=1}^n \left(\frac{m(X_i,\theta')}{\sigma_{P,j}(\theta')}\right)^2 \right| \\
&\qquad = \left| n^{-1}\sum_{i=1}^n \left(\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)} + \frac{m(X_i,\theta')}{\sigma_{P,j}(\theta')}\right)\left(\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)} - \frac{m(X_i,\theta')}{\sigma_{P,j}(\theta')}\right)\right| \\
&\qquad \leqslant n^{-1}\sum_{i=1}^n 2\sup_{\theta\in\Theta}\left|\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right| \bar{M}(X_i)\|\theta-\theta'\|.
\end{aligned} \tag{E.241}$$

Define $B_n \equiv n^{-1}\sum_{i=1}^n 2\sup_{\theta\in\Theta}\left|\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right|\bar{M}(X_i)$. By Markov and Cauchy-Schwarz inequalities,

$$P(B_n > C) \leqslant \frac{E[B_n]}{C} \leqslant \frac{2E_P\left[\sup_{\theta\in\Theta}\left|\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right|^2\right]^{1/2} E_P\left[\bar{M}(X_i)^2\right]^{1/2}}{C} \leqslant \frac{2M}{C}, \tag{E.242}$$

where the third inequality is due to Assumptions 4.1 (iv) and the assumption on $\bar{M}$. Hence, for any $\eta > 0$, one may find $C > 0$ such that $\sup_{P\in\mathcal{P}} P(B_n > C) < \eta$ for all $n$.

Similarly, for any $\theta, \theta' \in \Theta$,

$$\begin{aligned}
&\left| \left(n^{-1}\sum_{i=1}^n \frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right)^2 - \left(n^{-1}\sum_{i=1}^n \frac{m(X_i,\theta')}{\sigma_{P,j}(\theta')}\right)^2 \right| \\
&\qquad = \left| n^{-1}\sum_{i=1}^n \frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)} + n^{-1}\sum_{i=1}^n \frac{m(X_i,\theta')}{\sigma_{P,j}(\theta')}\right| \left| n^{-1}\sum_{i=1}^n \frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)} - n^{-1}\sum_{i=1}^n \frac{m(X_i,\theta')}{\sigma_{P,j}(\theta')}\right| \\
&\qquad \leqslant n^{-1}\sum_{i=1}^n 2\sup_{\theta\in\Theta}\left|\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right| \left| n^{-1}\sum_{i=1}^n \bar{M}(X_i)\right| \|\theta-\theta'\|.
\end{aligned} \tag{E.243}$$

Define $\tilde{B}_n \equiv n^{-1}\sum_{i=1}^n 2\sup_{\theta\in\Theta}\left|\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right| \left| n^{-1}\sum_{i=1}^n \bar{M}(X_i)\right|$. By Markov, Cauchy-Schwarz, and Jensen's inequalities,

$$\begin{aligned}
P(\tilde{B}_n > C) \leqslant \frac{E[\tilde{B}_n]}{C} &\leqslant \frac{2E_P\left[\left(n^{-1}\sum \sup_{\theta\in\Theta}\left|\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right|\right)^2\right]^{1/2} E_P\left[\left(n^{-1}\sum \bar{M}(X_i)\right)^2\right]^{1/2}}{C} \\
&\leqslant \frac{2E_P\left[\sup_{\theta\in\Theta}\left|\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right|^2\right]^{1/2} E_P[\bar{M}(X_i)^2]^{1/2}}{C} \leqslant \frac{2M}{C},
\end{aligned} \tag{E.244}$$

where the last inequality is due to Assumptions 4.1 (iv) and the assumption on $\bar{M}$. Hence, for any $\eta > 0$, one may find $C > 0$ such that $\sup_{P\in\mathcal{P}} P(\tilde{B}_n > C) < \eta$ for all $n$.

Finally, let $g(y) \equiv y^{-1/2} - 1$ and note that $|g(y) - g(y')| \leqslant \frac{1}{2}\sup_{\bar{y}\in(1-\epsilon,1+\epsilon)} |\bar{y}|^{-3/2}|y - y'|$ on $(1-\epsilon, 1+\epsilon)$. As shown in (E.238), $\hat{\sigma}_{n,j}^2(\theta)/\sigma_{P,j}^2(\theta)$ converges to 1 in probability, and $g$ is locally Lipschitz on a neighborhood of 1. Combining this with (E.240)-(E.244) yields the desired result. □

LEMMA E.13: *Suppose Assumption 4.1 holds. Suppose further that* $|\sigma_{P,j}(\theta)^{-1}m_j(x,\theta) - \sigma_{P,j}(\theta')^{-1}m_j(x,\theta')| \leqslant$

$\bar{M}(x)\|\theta-\theta'\|$ *with* $E_P[\bar{M}(X)^2]<M$ *for all* $\theta,\theta'\in\Theta$, $x\in\mathcal{X}$, $j=1,\cdots,J$, *and* $P\in\mathcal{P}$.

*Then,*

$$\sup_{P\in\mathcal{P}}\|Q_P(\theta_1,\tilde{\theta}_1)-Q_P(\theta_2,\tilde{\theta}_2)\|\leqslant M\|(\theta_1,\tilde{\theta}_1)-(\theta_2,\tilde{\theta}_2)\|, \tag{E.245}$$

*for some* $M>0$ *and for all* $\theta_1,\tilde{\theta}_1,\theta_2,\tilde{\theta}_2\in\Theta$.

*Proof.* Recall that

$$[Q_P(\theta_1,\tilde{\theta}_1)]_{j,k}=E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}\frac{m_k(X_i,\tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)}\Big]-E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}\Big]E_P\Big[\frac{m_k(X_i,\tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)}\Big]. \tag{E.246}$$

For any $\theta_1,\tilde{\theta}_1,\theta_2,\tilde{\theta}_2\in\Theta$,

$$\begin{aligned}
&\Big|E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}\frac{m_k(X_i,\tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)}\Big]-E_P\Big[\frac{m_j(X_i,\theta_2)}{\sigma_{P,j}(\theta_2)}\frac{m_k(X_i,\tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)}\Big]\Big|\\
&\leqslant\Big|E_P\Big[\Big(\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}-\frac{m_j(X_i,\theta_2)}{\sigma_{P,j}(\theta_2)}\Big)\frac{m_k(X_i,\tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)}\Big]\Big|+\Big|E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}\Big(\frac{m_k(X_i,\tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)}-\frac{m_k(X_i,\tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)}\Big)\Big]\Big|\\
&\leqslant E_P\Big[\sup_{\theta\in\Theta}\Big|\frac{m_k(X_i,\theta)}{\sigma_{P,k}(\theta)}\Big|\bar{M}(X_i)\Big]\|\theta_1-\theta_2\|+E_P\Big[\sup_{\theta\in\Theta}\Big|\frac{m_j(X_i,\theta)}{\sigma_{P,j}(\theta)}\Big|\bar{M}(X_i)\Big]\|\tilde{\theta}_1-\tilde{\theta}_2\|\\
&\leqslant M(\|\theta_1-\theta_2\|+\|\tilde{\theta}_1-\tilde{\theta}_2\|),
\end{aligned} \tag{E.247}$$

where the last inequality is due to the Cauchy-Schwarz inequality, Assumption 4.1 (iv), and the assumption on $\bar{M}$.

Similarly, for any $\theta_1,\tilde{\theta}_1,\theta_2,\tilde{\theta}_2\in\Theta$,

$$\begin{aligned}
&\Big|E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}\Big]E_P\Big[\frac{m_k(X_i,\tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)}\Big]-E_P\Big[\frac{m_j(X_i,\theta_2)}{\sigma_{P,j}(\theta_2)}\Big]E_P\Big[\frac{m_k(X_i,\tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)}\Big]\Big|\\
&\leqslant\Big|E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}-\frac{m_j(X_i,\theta_2)}{\sigma_{P,j}(\theta_2)}\Big]\Big|\Big|E_P\Big[\frac{m_k(X_i,\tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)}\Big]\Big|+\Big|E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}\Big]\Big|\Big|E_P\Big[\frac{m_k(X_i,\tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)}-\frac{m_k(X_i,\tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)}\Big]\Big|\\
&\leqslant E_P\Big[\sup_{\theta\in\Theta}\Big|\frac{m_k(X_i,\theta)}{\sigma_{P,k}(\theta)}\Big|\Big]E_P[\bar{M}(X_i)]\|\theta_1-\theta_2\|+E_P\Big[\sup_{\theta\in\Theta}\Big|\frac{m_j(X_i,\theta)}{\sigma_{P,j}(\theta)}\Big|\Big]E_P[\bar{M}(X_i)]\|\tilde{\theta}_1-\tilde{\theta}_2\|\\
&\leqslant M(\|\theta_1-\theta_2\|+\|\tilde{\theta}_1-\tilde{\theta}_2\|),
\end{aligned} \tag{E.248}$$

where the last inequality is due to the Cauchy-Schwarz inequality, Assumption 4.1 (iv), and the assumption on $\bar{M}$. The conclusion of the lemma then follows from (E.246)-(E.248). □

## E.3 Almost Sure Representation Lemma and Related Results

In this appendix, we provide details on the almost sure representation used in Lemmas E.3, E.4, E.6, and E.9. We start with stating a uniform version of the bootstrap consistency in van der Vaart and Wellner (2000). For this, we define the original sample $X^\infty=(X_1,X_2,\cdots)$ and a $n$-dimensional multinomial vector $M_n$ on a common probability space $(\mathcal{X}^\infty,\mathcal{A}^\infty,P^\infty)\times(\mathcal{Z},\mathcal{C},Q)$. We then view $X^\infty$ as the coordinate projection on the first $\infty$ coordinates of the probability space above. Similarly, we view $M_n$ as the coordinate projection on $\mathcal{Z}$. Here, $M_n$ follows a multinomial distribution with parameter $(n;1/n,\cdots,1/n)$ and is independent of $X^\infty$. We then let $E_M[\cdot|X^\infty=x^\infty]$ denote the conditional expectation of $M_n$ given $X^\infty=x^\infty$. Throughout, we let $\ell^\infty(\Theta,\mathbb{R}^J)$ denote uniformly bounded $\mathbb{R}^J$-valued functions on $\Theta$. We simply write $\ell^\infty(\Theta)$ when $J=1$.

Using the multinomial weight, we rewrite the empirical bootstrap process as

$$\mathbb{G}_{n,j}^b(\cdot) = g_j(X^\infty, M_n) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{n,i} - 1) m_j(X_i, \cdot)/\hat{\sigma}_{n,j}(\cdot), \ j = 1, \cdots, J, \tag{E.249}$$

where $g_j : \mathcal{X}^\infty \times \mathcal{Z} \to \ell^\infty(\Theta)$ is a function that maps the sample path and the multinomial weight $(X^\infty, M_n)$ to the empirical bootstrap process $\mathbb{G}_{n,j}^b$. We then let $g : \mathcal{X}^\infty \times \mathcal{Z} \to \ell^\infty(\Theta, \mathbb{R}^J)$ be defined by $g = (g_1, \cdots, g_J)'$. For any function $f : \ell^\infty(\Theta, \mathbb{R}^J) \to \mathbb{R}$, the conditional expectation of $f(\mathbb{G}_n^b)$ given the sample path $X^\infty$ is

$$E_M[f(\mathbb{G}_n^b)|X^\infty = x^\infty] = \int f \circ g(x^\infty, m_n) dQ(m_n), \tag{E.250}$$

where, with a slight abuse of notation, we use $Q$ for the induced law of $M_n$.

Let $\mathcal{F}$ be the function space $\{f(\cdot) = (m_1(\cdot, \theta)/\sigma_{P,1}(\theta), \cdots, m_J(\cdot, \theta)/\sigma_{P,J}(\theta)), \theta \in \Theta, P \in \mathcal{P}\}$. For each $j$, define a bootstrapped empirical process standardized by $\sigma_{P,j}$ as follows:

$$\begin{aligned} \mathfrak{G}_{n,j}^b(\theta) &\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( m_j(X_i^b, \theta) - \bar{m}_n(\theta) \right) / \sigma_{P,j}(\theta) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{n,i} - 1) m_j(X_i, \theta)/\sigma_{P,j}(\theta). \end{aligned} \tag{E.251}$$

The following result was shown in the proof of Lemma D.2.8 in Bugni, Canay, and Shi (2015), which is a uniform version of (a part of) Theorem 3.6.2 in van der Vaart and Wellner (2000). For the definition of a uniform version of Donskerness and pre-Gaussianity, we refer to van der Vaart and Wellner (2000) pages 168-169. Below, we let $P^*$ denote the outer probability of $P$ and let $T^*$ denote the minimal measurable majorant of any (not necessarily measurable) random element $T$.

LEMMA E.14: *Let $\mathcal{F}$ be a class of measurable functions with finite envelope function. Suppose $\mathcal{F}$ is such that (i) $\mathcal{F}$ is Donsker and pre-Gaussian uniformly in $P \in \mathcal{P}$; and (ii) $\sup_{P \in \mathcal{P}} P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$. Then,*

$$\sup_{h \in BL_1} |E_M[h(\mathfrak{G}_n^b)|X^\infty] - E[h(\mathbb{G}_P)]| \overset{as*}{\to} 0, \tag{E.252}$$

*uniformly in $P \in \mathcal{P}$.*

The result above gives uniform consistency of the standardized bootstrap process $\mathfrak{G}_n^b$. We now extend this to the studentized bootstrap process $\mathbb{G}_n^b$.

LEMMA E.15: *Suppose Assumptions 4.1, 4.2, and 4.5 hold. Then,*

$$\sup_{h \in BL_1} |E_M[h(\mathbb{G}_n^b)|X^\infty] - E[h(\mathbb{G}_P)]| \overset{as*}{\to} 0, \tag{E.253}$$

*uniformly in $P \in \mathcal{P}$.*

*Proof.* By Assumptions 4.1 (iv) and 4.5, Assumptions A.1-A.4 in Bugni, Canay, and Shi (2015) hold, which in turn implies that, by their Lemma D.1.2, $\mathcal{F}$ is Donsker and pre-Gaussian uniformly in $P \in \mathcal{P}$. Further, by Assumption 4.1 (iv) again, $\sup_{P \in \mathcal{P}} P^* \|f - Pf\|_{\mathcal{F}} < \infty$. Hence, by Lemma E.14,

$$\inf_{P \in \mathcal{P}} P^\infty \Big( \sup_{h \in BL_1} |E_M[h(\mathfrak{G}_n^b)|X^\infty] - E[h(\mathbb{G}_P)]|^* \to 0 \Big) = 1. \tag{E.254}$$

For later use, we define the following set of sample paths, which has probability 1 uniformly in $P \in \mathcal{P}$.

$$A \equiv \Big\{ x^\infty \in \mathcal{X}^\infty : \sup_{h \in BL_1} |E_M[h(\mathfrak{G}_n^b)|X^\infty = x^\infty] - E[h(\mathbb{G}_P)]|^* \to 0 \Big\}. \tag{E.255}$$

Note that $\mathbb{G}_{n,j}^b$ and $\mathfrak{G}_{n,j}^b$ are related to each other by the following relationship:

$$\mathbb{G}_{n,j}^b(\theta) - \mathfrak{G}_{n,j}^b(\theta) = \mathfrak{G}_{n,j}^b(\theta) \left( \frac{\sigma_{P,j}(\theta)}{\hat{\sigma}_{n,j}(\theta)} - 1 \right) = \mathfrak{G}_{n,j}^b(\theta) \eta_{n,j}(\theta), \ \theta \in \Theta. \tag{E.256}$$

By Assumptions 4.1, 4.2, and 4.5, Lemma E.10 applies. Hence,

$$\inf_{P \in \mathcal{P}} P^\infty \Big( \sup_{\theta \in \Theta} |\eta_{n,j}(\theta)|^* \to 0 \Big) = 1. \tag{E.257}$$

Define the following set of sample paths:

$$B \equiv \Big\{ x^\infty \in \mathcal{X}^\infty : \sup_{\theta \in \Theta} |\eta_{n,j}(\theta)|^* \to 0, \forall j = 1, \cdots, J \Big\}. \tag{E.258}$$

For any $x^\infty \in A \cap B$, it then follows that

$$\sup_{h \in BL_1} \big| E_M[h(\mathbb{G}_n^b)|X^\infty = x^\infty] - E[h(\mathbb{G}_P)] \big|^* \to 0, \tag{E.259}$$

due to (E.254) and (E.256), $h$ being Lipschitz, $\mathfrak{G}_{n,j}^b$ being bounded (given $x^\infty$), and $\sup_{\theta \in \Theta} |\eta_{n,j}(\theta)|^* \to 0$ for all $j$. Finally, note that $\inf_{P \in \mathcal{P}} P^\infty(A \cap B) = 1$ due to (E.254), (E.257), and De Morgan's law. This establishes the conclusion of the lemma. □

The following lemma shows that, for almost all sample path $x^\infty$, one can find an almost sure representation of the bootstrapped empirical process that is convergent.

LEMMA E.16: *Suppose Assumptions 4.1, 4.2, and 4.5 hold. Then, for each $x^\infty \in \mathcal{X}^\infty$, there exists a sequence $\{\tilde{G}_{n,x^\infty} \in \ell(\Theta, \mathbb{R}^J), n \geqslant 1\}$ and a random element $\tilde{G}_{P,x^\infty} \in \ell(\Theta, \mathbb{R}^J)$ defined on some probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathbf{P}})$ such that*

$$\int h \circ g(x^\infty, m_n) dQ(m_n) = \int h(\tilde{G}_{n,x^\infty}(\tilde{\omega})) d\tilde{\mathbf{P}}^*(\tilde{\omega}), \ \forall h \in BL_1 \tag{E.260}$$

$$\int h(\mathbb{G}_P(\omega)) dP(\omega) = \int h(\tilde{G}_{P,x^\infty}(\tilde{\omega})) d\tilde{\mathbf{P}}^*(\tilde{\omega}), \ \forall h \in BL_1, \tag{E.261}$$

*for all $x^\infty \in C$ for some set $C \subset \mathcal{X}^\infty$ such that $\inf_{P \in \mathcal{P}} P^\infty(C) = 1$ and*

$$\inf_{P \in \mathcal{P}} P^\infty \Big( \big\{ x^\infty \in \mathcal{X}^\infty : \tilde{G}_{n,x^\infty} \overset{\tilde{\mathbf{P}}-as*}{\to} \tilde{G}_{P,x^\infty} \big\} \Big) = 1. \tag{E.262}$$

*Proof.* Define the following set of sample paths:

$$C \equiv \Big\{ x^\infty \in \mathcal{X}^\infty : \sup_{h \in BL_1} |E_M[h(\mathbb{G}_{n,j}^b)|X^\infty = x^\infty] - E[h(\mathbb{G}_P)]|^* \to 0 \Big\}. \tag{E.263}$$

By Lemma E.15, $\inf_{P \in \mathcal{P}} P^\infty(C) = 1$.

For each fixed sample path $x^\infty \in C$, consider the bootstrap empirical process $g(x^\infty, M_n)$ in (E.249). This is a random element in $\ell^\infty(\Theta, \mathbb{R}^J)$ with a law governed by $Q$. For each $x^\infty \in C$, by Lemma E.15,

$$\sup_{h \in BL_1} \left| \int h \circ g(x^\infty, m_n) dQ(m_n) - E[h(\mathbb{G}_P)] \right|^* \to 0. \tag{E.264}$$

Hence, by Theorem 1.10.4 in van der Vaart and Wellner (2000), for each $x^\infty \in C$, one may find an almost sure representation $\tilde{G}_{n,x^\infty}$ of $g(x^\infty, M_n)$ on some probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathbf{P}})$ such that

$$\int h \circ g(x^\infty, m_n) dQ(m_n) = \int h(\tilde{G}_{n,x^\infty}(\tilde{\omega})) d\tilde{\mathbf{P}}^*(\tilde{\omega}), \ \forall h \in BL_1. \tag{E.265}$$

In particular, the proof of Theorem 1.10.4 in van der Vaart and Wellner (2000) (see also Addendum 1.10.5) allows us to take $\tilde{G}_{n,x^\infty}$ to be defined for each $\tilde{\omega} \in \tilde{\Omega}$ as

$$\tilde{G}_{n,x^\infty}(\tilde{\omega}) = g(x^\infty, M_n(\phi_n(\tilde{\omega}))), \tag{E.266}$$

for some perfect map $\phi_n : \tilde{\Omega} \to \mathcal{Z}$ (see the construction of $\phi_\alpha$ in the middle of page 61 in VW). One may define $\tilde{G}_{n,x^\infty}$ arbitrarily for any $x^\infty \notin C$. The almost sure representation $\tilde{G}_{P,x^\infty}$ of $\mathbb{G}_{P,j}$ is defined similarly.

By Theorem 1.10.4 in van der Vaart and Wellner (2000), Eq. (E.259), and $\inf_{P\in\mathcal{P}} P(C) = 1$, it follows that

$$\inf_{P\in\mathcal{P}} P^\infty\Big(\{x^\infty \in \mathcal{X}^\infty : \tilde{G}_{n,x^\infty} \overset{\tilde{\mathbf{P}}-as*}{\to} \tilde{G}_{P,x^\infty}\}\Big) = 1. \tag{E.267}$$

This establishes the claim of the lemma. □

Lemma E.17: *Suppose Assumptions 4.1, 4.2, and 4.5 hold. Let $W_n \equiv (\mathbb{G}_n^b, Y_n)$ be a sequence in $\mathcal{W} \equiv \ell(\Theta, \mathbb{R}^J) \times \mathbb{R}^{d_Y}$ such that $Y_n = \tilde{g}(X^\infty, M_n)$ for some map $\tilde{g} : \mathcal{X}^\infty \times \mathcal{Z} \to \mathbb{R}^{d_Y}$ and*

$$\inf_{P\in\mathcal{P}} P^\infty\big(\sup_{h\in BL_1} |E_M[h(W_n)|X^\infty = x^\infty] - E[h(W)]|^* \to 0\big) = 1, \tag{E.268}$$

*where $W = (\mathbb{G}, Y)$ is a Borel measurable random element in $\mathcal{W}$.*

*Then, for each $x^\infty \in \mathcal{X}^\infty$, there exists a sequence $\{W^*_{n,x^\infty} \in \mathcal{W}, n \geqslant 1\}$ and a random element $W^*_{x^\infty} \in \mathcal{W}$ defined on some probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathbf{P}})$ such that*

$$E_M[h(W_n)|X^\infty = x^\infty] = \int h(W^*_{n,x^\infty}(\tilde{\omega})) d\tilde{\mathbf{P}}^*(\tilde{\omega}), \ \forall h \in BL_1 \tag{E.269}$$

$$E[h(W)] = \int h(W^*_{x^\infty}(\tilde{\omega})) d\tilde{\mathbf{P}}^*(\tilde{\omega}), \ \forall h \in BL_1, \tag{E.270}$$

*for all $x^\infty \in C$ for some set $C \subset \mathcal{X}^\infty$ such that $\inf_{P\in\mathcal{P}} P^\infty(C) = 1$, and*

$$\inf_{P\in\mathcal{P}} P^\infty\Big(\{x^\infty \in \mathcal{X}^\infty : W^*_{n,x^\infty} \overset{\tilde{\mathbf{P}}-as*}{\to} \tilde{W}^*_{x^\infty}\}\Big) = 1. \tag{E.271}$$

*Proof.* Let $C \equiv \{x^\infty : \sup_{h\in BL_1} |E_M[h(W_n)|X^\infty = x^\infty] - E[h(W)]|^* \to 0\}$. The rest of the proof is the same as the one for Lemma E.16 and is therefore omitted. □

Remark E.1: When called by the Lemmas in Appendix E, Lemma E.17 is applied, for example, with $Y_n = (vec(\hat{D}_n(\theta'_n)), \hat{\xi}_n(\theta'_n))$ and $Y = (vec(D), \pi_1)$.

# Appendix F Further Comparison of Calibrated Projection and BCS-Profiling

We next show that finite sample power can be higher with calibrated projection than with BCS-profiling, and that, due to the slow rate at which $\kappa_n$ diverges, this effect can be large in samples of considerable size. Thus, the approaches are not nested in terms of power in empirically relevant examples. We then provide an example where

all of calibrated projection, BCS-profiling and the method of Pakes, Porter, Ho, and Ishii (2011) fail in a specific instance where Assumption 4.3 is not satisfied.

## F.1 Finite Sample Comparison in a Specific Example

We next analyze a stylized example of one-sided testing when the support set in direction $p$ is a singleton identified as the intersection of $d$ moment inequalities with regular geometry. In this example, calibrated projection has more power (less false coverage) than BCS-profiling, and the numerical difference can be large. The example resembles empirically important cases, namely polyhedral identified sets with large interior, e.g. linear regression with interval outcome data; recall that by Theorem 4.3, the two-sided testing problem reduces to two one-sided ones in these cases. At the same time, we emphasize that other examples will go the other way, especially as the present example utilizes the simplifications from Theorem 4.3 and therefore has no $\rho$-box.

Let $\theta$ be partially identified by moment conditions

$$E_P(z^{j\prime}\theta - X_j) \leqslant 0, j = 1, \ldots, d.$$

Note that to simplify the analysis, we assume exactly $d$ conditions. Assume that $\{z^1, \ldots, z^d\}$ are linearly independent and also that $p$ is in their positive span, so that $\Theta_I$ is bounded in direction $p$ but not $-p$. The confidence intervals will be accordingly one-sided. Since gradients are known, all simplifications from Theorem 4.3 apply. We borrow from algebra in the proof of Theorem 4.4 to observe that, with the simplifications in place, $CI_n$ and $CI_n^{prof}$ invert tests that use the same test statistic but different bootstrap approximations to its distribution as follows:

$$\begin{aligned}
T_n^{DR} &= \max_j \left\{\mathbb{G}_{n,j}^b\right\} \\
T_n^{PR}(s_n) &= \min_{p'\lambda=0} \max_j \left\{ \mathbb{G}_{n,j}^b + \underbrace{\frac{\sqrt{n}}{\kappa_n} \frac{z^{j\prime}\hat{\theta}_{p,s_n}^* - \bar{X}_j}{\hat{\sigma}_{n,j}}}_{>0} + \frac{z^{j\prime}\lambda}{\hat{\sigma}_{n,j}} \right\} \\
T_n^b &= \min_{p'\lambda=0} \max_j \left\{ \mathbb{G}_{n,j}^b + \underbrace{\frac{\sqrt{n}}{\kappa_n} \frac{z^{j\prime}\hat{\theta}_p^* - \bar{X}_j}{\hat{\sigma}_{n,j}}}_{=0} + \frac{z^{j\prime}\lambda}{\hat{\sigma}_{n,j}} \right\} \leqslant \min\{T_n^{DR}, T_n^{PR}(s_n)\},
\end{aligned}$$

where (as in Theorem 4.3) $s_n$ is the value of $p'\theta$ being tested and $\hat{\theta}_{p,s_n}$ minimizes the sample criterion subject to $p'\theta = s_n$. The last inequality is strict unless the problem defining $T_n^b$ is solved by $\lambda = 0$. The assessments of intercept terms in $T_n^{PR}(s_n)$ and $T_n^b$ use that by construction of the example, all sample constraints bind at $\hat{\theta}_p^*$ and are violated at $\hat{\theta}_{p,s_n}^*$ (else, the test statistic would be 0 and the critical value not computed). Equality thus requires knife-edge realizations of $\mathbb{G}_{n,j}^b$, so its probability vanishes as $\mathbb{G}_{n,j}^b$ approaches multivariate normality and is in fact 0 for typical empirical samples. We conclude that the calibrated projection $CI_n$ is deterministically a weak (and essentially always a strict) subset of the BCS-profiling $CI_n^{prof}$ in this example.

We next provide a numerical comparison in a further stripped-down version of the example. Thus, consider one-sided testing with moment conditions

$$\begin{aligned}
-\theta_1 + \theta_2 - E_P(X_1) &\leqslant 0 \\
\theta_1 + \theta_2 - E_P(X_2) &\leqslant 0
\end{aligned}$$

where the data are $(X_1, X_2) \sim N((E_P(X_1), E_P(X_2)), I_2)$ and $E_P(X_1) = E_P(X_2) = 0$. All of these facts other than

$E_P(X_1) = E_P(X_2) = 0$, but including the gradients and variance matrix, are known. This enables closed form arguments. Also, for a researcher knowing this, the natural bootstrap implementation is a parametric bootstrap:

$$\begin{aligned} & \left(X_1^b, X_2^b\right) \sim N((\bar{X}_1, \bar{X}_2), I_2) \\ \implies \quad & \sqrt{n}(\bar{X}_1^b - \bar{X}_1, \bar{X}_2^b - \bar{X}_1) = (Z_1, Z_2) \sim N(0, I_2) \end{aligned}$$

which we will use, i.e. $(Z_1, Z_2)$ will take the role of $(\mathbb{G}_{n,1}^b, \mathbb{G}_{n,2}^b)$. Numerical computations refer to $\alpha = 5\%$.

Let $p = (0, 1)$. We construct one-sided confidence intervals for $s(p, \Theta_I(P))$. All intervals contain $(-\infty, s(p, \hat{\Theta}_I)]$, and simple algebra shows $s(p, \hat{\Theta}_I) = \frac{\bar{X}_1 + \bar{X}_2}{2}$. Also noting that in this example $s(p, \Theta_I(P)) = 0$ and, for $s_n > s(p, \hat{\Theta}_I)$,

$$\begin{aligned} H(p, \hat{\Theta}_I) &= \left\{ \left( \frac{-\bar{X}_1 + \bar{X}_2}{2}, \frac{\bar{X}_1 + \bar{X}_2}{2} \right) \right\} \\ \hat{\Theta}_I(s_n) &\equiv \left\{ \theta \in \Theta : p'\theta = s_n,\, Q_n(\theta) \leqslant \inf_{\theta \in \Theta : p'\theta = s_n} Q_n(\theta) \right\} = \left\{ \left( \frac{-\bar{X}_1 + \bar{X}_2}{2}, s_n \right) \right\} \\ T_n(s_n) &= \sqrt{n} \max \left\{ s_n - \frac{\bar{X}_1 + \bar{X}_2}{2}, 0 \right\}, \end{aligned}$$

where $Q_n(\theta) = \max_{j=1,\ldots,J_1} (\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta))_+$, we compute

$$\begin{aligned} T_n^{DR} &= \min_{\theta \in \hat{\Theta}_I(s_n)} \max \left\{ \sqrt{n}\left(\bar{X}_1^b - \bar{X}_1\right), \sqrt{n}\left(\bar{X}_2^b - \bar{X}_2\right), 0 \right\} \\ &= \max \left\{ \sqrt{n}\left(\bar{X}_1^b - \bar{X}_1\right), \sqrt{n}\left(\bar{X}_2^b - \bar{X}_2\right), 0 \right\} \sim \max\{Z_1, Z_2, 0\} \\ T_n^{PR}(s_n) &= \min_{\theta_1 \in \mathbb{R}} \max \left\{ \sqrt{n}\left(\bar{X}_1^b - \bar{X}_1\right) + \kappa_n^{-1}\sqrt{n}\left(-\theta_1 + s_n - \bar{X}_1\right), \sqrt{n}\left(\bar{X}_2^b - \bar{X}_2\right) + \kappa_n^{-1}\sqrt{n}\left(\theta_1 + s_n - \bar{X}_2\right), 0 \right\}. \end{aligned}$$

Unless its value is 0, the minimization problem defining $T_n^{PR}(s_n)$ is solved by setting two terms equal:

$$\theta_1 = \frac{\sqrt{n}\left(\bar{X}_1^b - \hat{\mu}_1\right) - \sqrt{n}\left(\bar{X}_2^b - \bar{X}_2\right) + \kappa_n^{-1}\sqrt{n}\left(\bar{X}_2 - \bar{X}_1\right)}{2\kappa_n^{-1}\sqrt{n}},$$

leading to

$$\begin{aligned} T_n^{PR}(s_n) &= \max \left\{ \frac{\sqrt{n}\left(\bar{X}_1^b - \bar{X}_1\right) + \sqrt{n}\left(\bar{X}_2^b - \bar{X}_2\right)}{2} + \kappa_n^{-1}\sqrt{n}\left(s_n - \frac{\bar{X}_1 + \bar{X}_2}{2}\right), 0 \right\} \\ &= \max \left\{ \frac{Z_1 + Z_2}{2} + \kappa_n^{-1}\sqrt{n}\left(s_n - \frac{\bar{X}_1 + \bar{X}_2}{2}\right), 0 \right\} = \max \left\{ \frac{Z_1 + Z_2}{2} + \kappa_n^{-1} T_n(s_n), 0 \right\}. \end{aligned}$$

Finally, very similar reasoning to the above gives

$$\begin{aligned} T_n^b &= \min_{\lambda \in \mathbb{R}} \max \left\{ \sqrt{n}\left(\bar{X}_1^b - \bar{X}_1\right) + \kappa_n^{-1}\sqrt{n} \min \left( \frac{\bar{X}_1 - \bar{X}_2}{2} + \frac{\bar{X}_1 + \bar{X}_2}{2} - \bar{X}_1, 0 \right) - \lambda, \right. \\ &\qquad \left. \sqrt{n}\left(\bar{X}_2^b - \bar{X}_2\right) + \kappa_n^{-1}\sqrt{n} \min \left( \frac{-\bar{X}_1 + \bar{X}_2}{2} + \frac{\bar{X}_1 + \bar{X}_2}{2} - \bar{X}_2, 0 \right) + \lambda, 0 \right\} \\ &= \min_{\lambda \in \mathbb{R}} \max \left\{ \sqrt{n}\left(\bar{X}_1^b - \bar{X}_1\right) - \lambda, \sqrt{n}\left(\bar{X}_2^b - \bar{X}_2\right) + \lambda, 0 \right\} \\ &= \max \left\{ \frac{\sqrt{n}\left(\bar{X}_1^b - \bar{X}_1\right) + \sqrt{n}\left(\bar{X}_2^b - \bar{X}_2\right)}{2}, 0 \right\} \\ &= \max \left\{ \frac{Z_1 + Z_2}{2}, 0 \right\}. \end{aligned}$$

Thus calibrated projection yields a critical value of $\hat{c}_n = \Phi^{-1}(1-\alpha)/\sqrt{2} \approx 1.16$, whereas simple projection uses

Table F.1: Finite sample noncoverage rates in a specific example.

| Type of cv | n | Value | Power at $\gamma n^{-1/2}$, $\gamma$ = ... | | | | |
|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 |
| $\hat{c}_n^{proj}$ | any | 1.95 | .003 | .089 | .523 | .930 | .998 |
| $\tilde{c}_n^{prof}$ | $10^3$ | 1.63 | .011 | .188 | .701 | .974 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^5$ | 1.52 | .016 | .231 | .751 | .982 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^7$ | 1.47 | .019 | .254 | .774 | .985 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^9$ | 1.43 | .022 | .271 | .790 | .987 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^{11}$ | 1.40 | .024 | .284 | .800 | .988 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^{13}$ | 1.38 | .025 | .292 | .807 | .989 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^{15}$ | 1.37 | .026 | .299 | .813 | .989 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^{17}$ | 1.36 | .027 | .307 | .819 | .990 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^{19}$ | 1.35 | .028 | .313 | .823 | .990 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^{50}$ | 1.28 | .036 | .348 | .847 | .993 | 1.000 |
| $\tilde{c}_n^{prof}$ | $10^{100}$ | 1.24 | .039 | .366 | .858 | .994 | 1.000 |
| $\hat{c}_n$ | any | 1.16 | .050 | .409 | .882 | .995 | 1.000 |

$\hat{c}_n^{proj} = \Phi^{-1}(\sqrt{1-\alpha}) \approx 1.95$; both are independent of $s_n$ as well as $n$. BCS-profiling uses a critical value $\hat{c}_n^{prof}(s_n)$ that increases in the test statistic (hence, conditional on the data, in $s_n$) because the statistic itself enters $T_n^{PR}$. To facilitate a comparison, one can compute the fixed point at which $T_n(s_n) = \hat{c}_n^{prof}(s_n)$. BCS-profiling is equivalent to comparing $T_n(s_n)$ to that fixed point at all $s_n$, and we will therefore equate it with use of this critical value, labeled $\tilde{c}_n^{prof}$ below. This critical value converges to $\hat{c}_n$ at a rate of $\kappa_n^{-1}$, illustrating asymptotic equivalence of inference methods off the null in this case. However, for the popular choice of $\kappa_n = \sqrt{\log n}$, convergence is so slow that it should not be taken to describe behavior at realistic sample sizes. Table F.1 displays the numerical value of $\tilde{c}_n^{prof}$ and the implied noncoverage probability (or power) at $\gamma/\sqrt{n}$ for $\gamma \in \{0, 1, 2, 3, 4\}$; note that $\gamma = 0$ corresponds to the true support function. By construction, $\tilde{c}_n^{prof}$ interpolates between $\hat{c}_n^{proj}$ and $\hat{c}_n$ in this example, but convergence to $\hat{c}_n$ requires extreme sample sizes. For example, on the boundary edge of the true projection $CI_{.95}^{prof}$ has finite sample coverage of .975, which is effectively halfway between projection and calibrated projection, for $n = 10^{13}$.

## F.2 Example of Methods Failure When Assumption 4.3 Fails

Consider one-sided testing with two inequality constraints in $\mathbb{R}^2$. The constraints are

$$\begin{aligned} \theta_1 + \theta_2 &\leqslant E_P(X_1) \\ \theta_1 - \theta_2 &\leqslant E_P(X_2). \end{aligned}$$

The projection of $\Theta_I(P)$ in direction $p = (1, 0)$ is $(-\infty, (E_P(X_1) + E_P(X_2))/2]$, the support set is $H(p, \Theta_I) = \{((E_P(X_1)+E_P(X_2))/2, (E_P(X_1)-E_P(X_2))/2)\}$, and the support function takes value $\theta_1^* = (E_P(X_1)+E_P(X_2))/2$.

The random variables $(X_1, X_2)'$ have a mixture distribution as follows:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \begin{cases} N\left(0, \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right) & \text{with probability } 1 - 1/n, \\ \delta_{(1,1)} \text{ (degenerate)} & \text{otherwise,} \end{cases}$$

hence $E_P(X_1) = E_P(X_2) = \theta_1^* = 1/n$. Note in particular the implication that

$$\frac{X_1 + X_2}{2} = \begin{cases} 0 & \text{with probability } 1 - 1/n, \\ 1 & \text{otherwise.} \end{cases}$$

The natural estimator of $\theta_1^*$ is $\hat{\theta}_1^* = (\bar{X}_1 + \bar{X}_2)/2$. It is distributed as $Z/n$, where $Z$ is Binomial with parameters $(1/n, n)$. For large $n$, the distribution of $Z$ is well approximated as Poisson with parameter 1. In particular, with probability approximately $e^{-1} \approx 37\%$, every sample realization of $(X_1+X_2)/2$ equals zero. In this case, the following happens: (i) The projection of the sample analog of the identified set is $(-\infty, 0]$, so that a strictly positive critical value or level would be needed to cover the true projection. (ii) Because the empirical distribution of $(X_1 + X_2)/2$ is degenerate at zero, the distribution of $(\bar{X}_1^b + \bar{X}_2^b)/2$ is as well. Hence, all of Pakes, Porter, Ho, and Ishii (2011), Bugni, Canay, and Shi (2017), and calibrated projection (each with either parametric or nonparametric bootstrap) compute critical values or relaxation levels of 0.

This bounds from above the true coverage of all of these methods at $e^{-1} \approx 63\%$. Note that $(m < n)$-subsampling will encounter the same problem. Next we provide some discussion of the example.

**Violation of Assumptions.** The example violates our Assumption 4.3 because $Cov(X_1, X_2) \to 1$. It also violates Assumption 2 in Bugni, Canay, and Shi (2017): Their Assumption A2-(b) should apply, but the profiled test statistic on the true null concentrates at $1/n$. The example satisfies the assumptions explicitly stated in Pakes, Porter, Ho, and Ishii (2011), illustrating an oversight in their Theorem 2. (We here refer to the inference part of their 2011 working paper. We identified corresponding oversights in the proof of their Proposition 6.)

The example satisfies the assumptions of Andrews and Soares (2010) and Andrews and Guggenberger (2009), and both methods work here. The reason is that both focus on the distribution of the criterion function at a fixed $\theta$ and are not affected by the irregularity of $\hat{\theta}_1^*$.

**Relation to Mammen (1992).** In this example, all of Bugni, Canay, and Shi (2017), Pakes, Porter, Ho, and Ishii (2011), and our calibrated projection method reduce to one-sided nonparametric percentile bootstrap confidence intervals for $(E_P(X_1)+E_P(X_2))/2$ estimated by $(\bar{X}_1+\bar{X}_2)/2$. By Mammen (1992, Theorem 1), asymptotic normality of an appropriately standardized estimator, i.e.

$$\exists\{a_n\} : a_n \left((\bar{X}_1 + \bar{X}_2) - (E_P(X_1) + E_P(X_2))\right) \xrightarrow{d} N(0, 1),$$

is *necessary and* sufficient for this interval to be valid. This fails (the true limit is recentered Poisson at rate $a_n = n$), so that validity of any of the aforementioned methods would contradict the Theorem.

# Appendix G Comparison with Projection of AS

In this Appendix we establish that for each $n \in \mathbb{N}$, $CI_n$ is a subset of a confidence interval obtained by projecting an AS confidence set and denoted $CI_n^{proj}$.[50] 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 $\hat{c}_n^{proj}$ denote the critical value obtained applying AS with criterion function $Q_n(\theta) = \max\{\max_{j=1,\ldots,J_1}(\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta))_+, \max_{j=J_1+1,\cdots,J_1+J_2}|\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)|\}$ and with the same choice as for $\hat{c}_n$ of GMS function $\varphi$ and tuning parameter $\kappa_n$.

THEOREM G.1: *Suppose Assumptions 4.1, 4.2, 4.3, 4.4, and 4.5 hold. Let $0 < \alpha < 1/2$. Then for each $n \in \mathbb{N}$*

$$CI_n \subseteq [-s(-p, \mathcal{C}_n(\hat{c}_n^{proj})), s(p, \mathcal{C}_n(\hat{c}_n^{proj}))], \tag{G.1}$$

*where for given function $c$, $\mathcal{C}_n(c)$ is defined in (1.1)*

*Proof.* For given $\theta$, the event

$$\max_{j=1,\ldots,J}\left\{\mathbb{G}_{n,j}^b(\theta) + \varphi_j(\hat{\xi}_{n,j}(\theta))\right\} \leqslant c \tag{G.2}$$

implies the event

$$\max_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \geqslant 0 \geqslant \min_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda, \tag{G.3}$$

with $\Lambda_n^b$ defined in (3.1). This is so because if $\max_{j=1,\ldots,J}\left\{\mathbb{G}_{n,j}^b(\theta) + \varphi_j(\hat{\xi}_{n,j}(\theta))\right\} \leqslant c$, $\lambda = 0$ is feasible in both optimization problems in (G.3), hence the event in (G.3) is implied. In turn this yields that for each $n \in \mathbb{N}$ and $\theta \in \Theta$,

$$c_n^{proj}(\theta) \geqslant \hat{c}_n(\theta), \tag{G.4}$$

and therefore the result follows. □

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

$$\max_{j=1,\ldots,J}\left\{\mathbb{G}_{n,j}^b(\theta) + \varphi_j(\hat{\xi}_{n,j}(\theta))\right\} \leqslant c. \tag{G.5}$$

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 $\hat{c}_n^{proj}(\theta)$ at each $\theta$ that is relevant for projection. For this, restrict $\varphi(\cdot)$ to satisfy $\varphi_j(x) \leqslant 0$ for all $x$, fix $\theta$ and consider the pointwise limit of (G.5):

$$\mathbb{G}_{P,j}(\theta) + \zeta_{P,j}(\theta) \leqslant c, \ j = 1, \cdots, J, \tag{G.6}$$

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, $\hat{c}_n^{proj}(\theta)$ then converges in probability to a critical value $c = c^{proj}(\theta)$ such that (G.6) holds with probability $1 - \alpha$. Similarly, the limiting event that corresponds to our problem (3.4) is

$$\Lambda(\theta, \rho, c) \cap \{p'\lambda = 0\} \neq \varnothing, \tag{G.7}$$

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

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

$$\Lambda(\theta, \rho, c) = \{\lambda \in \rho B^d_{n,\rho} : \mathbb{G}_{P,j}(\theta) + D_{P,j}(\theta)\lambda + \zeta_{P,j}(\theta) \leqslant c, j = 1, \cdots, J\}. \tag{G.8}$$

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 (G.6) implies (G.7) (by arguing that $\lambda = 0$ is in $\Lambda(\theta, \rho, c)$), and for the converse implication, whenever (G.7) holds, there is $\lambda$ such that $\mathbb{G}_{P,j}(\theta) + D_{P,j}(\theta)\lambda + \zeta_{P,j}(\theta) \leqslant c$ and $p'\lambda = 0$. Since $D_{P,j}(\theta)\lambda = 0$ for all $j$ such that $\zeta_{P,j}(\theta) = 0$, one has $\mathbb{G}_{P,j}(\theta) + \zeta_{P,j}(\theta) \leqslant c$ for all $j$.[51] In this special case, the limits of the two critical values coincide asymptotically, but any other case is characterized by projection conservatism. Lemma G.1 below formalizes this insight. Specifically, for fixed $\theta$, the limit of $\hat{c}_n(\theta)$ is strictly less than the limit of $\hat{c}^{proj}_n(\theta)$ if and only if there is a constraint that binds or is violated at $\theta$ and has a gradient that is not a scalar multiple of $p$.[52]

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 $\theta$ (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).

Example G.1 (Linear regression with an interval valued outcome): Consider a linear regression model:

$$E[Y|Z] = Z'\theta, \tag{G.9}$$

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:

$$E_P[Y_L|Z = z_j] \leqslant z_j'\theta \leqslant E_P[Y_U|Z = z_j], \ j = 1, \cdots, K, \tag{G.10}$$

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

$$m_j(X, \theta) = \begin{cases} Y_L 1\{Z = z_j\}/g(z_j) - z_j'\theta, & j = 1, \cdots, K \\ z_{j-K}'\theta - Y_U 1\{Z = z_{j-K}\}/g(z_{j-K}) & j = K+1, \cdots, 2K, \end{cases} \tag{G.11}$$

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 $D_{P,j}(\theta) = -z_j/\sigma_{P,j}, j = 1, \cdots, K$ 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 $\theta$ 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 $\theta$, our critical value is strictly smaller than $c^{proj}_n(\theta)$ 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

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

[52] 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 $\Theta_I$.

Table G.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** | $\infty$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{c}_n$ | 1.64 | 1.16 | 0.95 | 0.82 | 0.74 | 0.67 | 0.62 | 0.58 | 0.55 | 0.52 | 0.16 | 0 |
| $\hat{c}_n^{proj}$ | 1.64 | 1.95 | 2.12 | 2.23 | 2.32 | 2.39 | 2.44 | 2.49 | 2.53 | 2.57 | 3.28 | $\infty$ |
| $1-\alpha^*$ | .95 | .77 | .57 | .40 | .27 | .18 | .11 | .07 | .04 | .03 | $10^{-25}$ | 0 |

the (co)variance of the Gaussian process in our limiting approximation but does not affect any other term.

We now provide a numerical illustration for a further simplified example. 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, $\Theta_I$ is the negative quadrant. Its unboundedness from below is strictly for simplicity.) Also, by Theorem 4.3, one can set $\rho = +\infty$ in this example.

Under these simplifying assumptions (which can, of course, be thought of as asymptotic approximations), it is easy to calculate in closed form that

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

Furthermore, for any $\alpha < 1/2$, one can compute $\alpha^*$ s.t. applying $\hat{c}_n$ with target coverage $(1-\alpha)$ yields the same confidence interval as using $\hat{c}_n^{proj}$ with target coverage $(1-\alpha^*)$.[53] Some numerical values are provided in Table G.1 (with $\alpha = 0.05$).

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 in Section 5 showcase less extreme cases. However, 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.

## G.1 Necessary and Sufficient Condition for $\hat{c}_n(\theta) < \hat{c}_n^{proj}(\theta)$

The following lemma establishes the effect of $\rho$ on $\hat{c}_n(\theta)$. In doing so it establishes a necessary and sufficient condition for $\hat{c}_n(\theta) < \hat{c}_n^{proj}(\theta)$, because the latter can be seen as the former calibrated with $\rho$ set equal to zero. The lemma requires $\varphi_j(x) \leqslant 0$ for all $x$.[54]

LEMMA G.1: *Fix $\theta \in \Theta$, $P \in \mathcal{P}$ and a value $\rho \in \mathbb{R}_+$. Suppose Assumptions 4.1, 4.2, 4.3, 4.4 and 4.5 hold and also that $\varphi_j(x) \leqslant 0$ for all $x$ and $j$. Let $0 < \delta < \rho$. For $n \geqslant N$, let $\hat{c}_n(\theta)$ be calibrated using $\rho$ in place of $\rho$, which*

[53] Equivalently, $(1-\alpha^*)$ is the probability that $\mathcal{C}_n(\hat{c}_n^{proj})$ 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.

[54] To keep the treatment general, we have not imposed this restriction throughout the paper. However, we only recommend functions $\varphi_j$ with this feature anyway: for any $\varphi_j$ 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.

*necessarily yields a larger value for* $\hat{c}_n(\theta)$*. With a modification of notation, explicitly highlight* $\hat{c}_n(\theta)$*'s dependence on* $\rho$ *through the notation* $\hat{c}_n(\theta, \rho)$*. Then*

$$|\hat{c}_n(\theta, \rho) - \hat{c}_n(\theta, \rho - \delta)| \xrightarrow{p} 0 \tag{G.12}$$

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

REMARK G.1: For $\theta$ such that $\mathcal{J}^*(\theta) = \varnothing$, we have $\hat{c}_n(\theta, \rho) \xrightarrow{p} 0$ but also $\hat{c}_n^{proj}(\theta) \xrightarrow{p} 0$. This is consistent with Lemma G.1 because the condition on gradients vacuously holds in this case.

*Proof.* Recall that $\theta$ and $P$ are fixed, i.e. we assume a pointwise perspective. Then

$$\hat{c}_n(\theta, \rho) \xrightarrow{p} \inf\{c \geqslant 0 : P(\{\lambda \in \rho B_{n,\rho}^d : \mathbb{G}_{P,j}(\theta) + D_{P,j}(\theta)\lambda \leqslant c, j \in \mathcal{J}^*(\theta)\} \cap \{p'\lambda = 0\} \neq \varnothing) \geqslant 1 - \alpha\}. \tag{G.13}$$

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

$$\varphi_j(\kappa_n^{-1}\sqrt{n}\bar{m}_j(X_i, \theta)/\sigma_{P,j}(\theta)) \xrightarrow{p} \begin{cases} 0 & \text{if } \ j \in \mathcal{J}^*(\theta) \\ -\infty & \text{otherwise,} \end{cases} \tag{G.14}$$

where the first of those cases uses nonpositivity of $\varphi_j$. It therefore suffices to show that the right hand side of G.13 strictly decreases in $\rho$ 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 $\theta$. 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 $K\lambda \leqslant g$ has a solution if and only if for all $\mu \in \mathbb{R}_+^{J^*+2d+2}$,

$$\mu' K = 0 \Rightarrow \mu' g \geqslant 0. \tag{G.15}$$

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

$$\begin{aligned} & P(\{\lambda \in \rho B_{n,\rho}^d : \mathbb{G}_{P,j} + D_{P,j}\lambda \leqslant c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \varnothing) \\ > \ & P(\{\lambda \in (\rho - \delta)\rho B_{n,\rho}^d : \mathbb{G}_{P,j} + D_{P,j}\lambda \leqslant c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \varnothing) \end{aligned} \tag{G.16}$$

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

$$P(\{\mu' g \geqslant 0 \ \forall \mu \in \mathcal{M}\} \cap \{\mu' g - \delta\tau < 0 \ \exists \mu \in \mathcal{M}\}) > 0. \tag{G.17}$$

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

$$\min_{\mu \in \mathbb{R}_+^{J^*+2d+2}} \left\{ \mu' g : \ \mu' K = 0 \right\} \geqslant 0 \tag{G.18}$$

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

$$\min_{\mu \in \mathbb{R}_+^{J^*+2d+2}} \left\{ \mu' g : \ \mu' K = 0, \ \mu \in \arg \min_{\tilde{\mu} \in \mathbb{R}_+^{J^*+2d+2}} (\tilde{\mu}' g : \ \tilde{\mu}^{[1:J^*]} = \mu^{[1:J^*]}, \tilde{\mu}' K = 0) \right\}, \tag{G.19}$$

That is, we can restrict attention to a concentrated out subset of vectors $\mu$, where the last $(2d+2)$ components of any $\mu$ minimize the objective function among all vectors that agree with $\mu$ in the first $J^*$ components. The inner minimization problem in equation (G.19) can be written as

$$\min_{\tilde{\mu}^{[J^*+1:J^*+2d+2]} \in \mathbb{R}_+^{2d+2}} \rho \sum_{j=J^*+1}^{J^*+2d} \tilde{\mu}_j \quad \text{s.t.} \quad \begin{bmatrix} \tilde{\mu}_{J^*+1} - \tilde{\mu}_{J^*+d+1} + \tilde{\mu}_{J^*+2d+1} - \tilde{\mu}_{J^*+2d+2} \\ \tilde{\mu}_{J^*+2} - \tilde{\mu}_{J^*+d+2} \\ \vdots \\ \tilde{\mu}_{J^*+d} - \tilde{\mu}_{J^*+2d} \end{bmatrix} = -\mu^{[1:J^*]\prime} D^{[1:J^*,:]}. \tag{G.20}$$

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

$$\mu^{[J^*+1:J^*+2d+2]} = \begin{bmatrix} 0 \\ -\left[ D^{\left[1:J^*,2:d\right]\prime} \mu^{[1:J^*]} \wedge 0 \cdot \mathbf{1}_{d-1} \right] \\ 0 \\ D^{\left[1:J^*,2:d\right]\prime} \mu^{[1:J^*]} \vee 0 \cdot \mathbf{1}_{d-1} \\ -\left[ D^{\left[1:J^*,1\right]\prime} \mu^{[1:J^*]} \wedge 0 \right] \\ D^{\left[1:J^*,1\right]\prime} \mu^{[1:J^*]} \vee 0 \end{bmatrix}, \tag{G.21}$$

where $D^{\left[1:J^*,2:d\right]\prime} \mu^{[1:J^*]} \vee 0 \cdot \mathbf{1}_{d-1}$ 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 $\mu^{[1:J^*]\prime} D = (\mu^{[1:J^*]\prime} D^{\left[1:J^*,1\right]}, 0, ..., 0)'$ and therefore all but the last two entries of $\mu^{[J^*+1:J^*+2d+2]}$ equal zero. One can, therefore, restrict attention to vectors $\mu$ 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 (G.17) 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

$$\varpi = \begin{bmatrix} 1 \\ 0 \cdot \mathbf{1}_{J^*-1} \\ 0 \\ -\left[ (D^{[1,2:d]\prime}) \wedge 0 \cdot \mathbf{1}_{d-1} \right] \\ 0 \\ (D^{[1,2:d]\prime}) \vee 0 \cdot \mathbf{1}_{d-1} \\ -\left[ (D^{[1,1]}) \wedge 0 \right] \\ (D^{[1,1]}) \vee 0 \end{bmatrix} \tag{G.22}$$

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

As in the proof of Lemma E.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 (G.21). But then $\varpi$ 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

$$\begin{aligned} \nu^{t\prime} g &\geqslant 0, \text{ all } \nu^t \\ \nu^{t\prime}(g - \delta\tau) &< 0, \text{ some } \nu^t \end{aligned}$$

by observing that if we define

$$\iota_k = \begin{bmatrix} -\rho \cdot \sum_{i=2}^{d} \left|D^{[1,i]}\right| \\ k \cdot \mathbf{1}_{J^*-1} \\ \rho \cdot \mathbf{1}_d \\ \rho \cdot \mathbf{1}_d \\ 0 \\ 0 \end{bmatrix}, \tag{G.23}$$

then we have

$$0 = \varpi'\iota_k = \min_t \nu^{t\prime}\iota_k.$$

Any other spanning vector $\nu^t$ 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 $\iota_k$ if $k$ is chosen large enough.

A realization of $g = \iota_k$ would, therefore, yield

$$\nu^{t\prime} g \geqslant 0 \ \forall \nu^t \in \mathcal{M}, \text{ and } \varpi^{t\prime}(g - \delta\tau) < -\epsilon, \tag{G.24}$$

for some $\epsilon > 0$. Let

$$\Gamma_k = \{\iota : \iota = \iota_k + \epsilon/2b, \ \|b\| \leqslant 1 \text{ and } \varpi' b > 0\}. \tag{G.25}$$

Then

$$\nu^{t\prime}\iota \geqslant 0 \ \forall \nu^t \in \mathcal{M}, \text{ and } \varpi^{t\prime}(\iota - \delta\tau) < -\epsilon/2, \ \forall \iota \in \Gamma_k. \tag{G.26}$$

The probability in equation (G.17) is therefore strictly positive.

**Step 2.** Next, we argue that

$$P(\{\lambda \in \rho B_{n,\rho}^d : \ \mathbb{G}_j + D_j\lambda \leqslant c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \varnothing) \tag{G.27}$$

strictly continuously increases in $c$. The rigorous argument is very similar to the use of Farkas' Lemma in step 1 and in Lemma E.6. We leave it at an intuition: As $c$ increases, the set of vectors $g$ fulfilling the right hand side of (G.15) 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

$$\begin{aligned} & \inf_{c \geqslant 0}\{P(\{\lambda \in \rho B^d_{n,\rho} : \mathbb{G}_j + D_j\lambda \leqslant c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \varnothing) \geqslant 1-\alpha\} \\ > \; & \inf_{c \geqslant 0}\{P(\{\lambda \in (\rho-\delta)\rho B^d_{n,\rho} : \mathbb{G}_j + D_j\lambda \leqslant c, j \in \mathcal{J}^*\} \cap \{p'\lambda = 0\} \neq \varnothing) \geqslant 1-\alpha\} \end{aligned} \tag{G.28}$$

and hence the result. □

# References


Adams, R. A., and J. J. Fournier (2003): *Sobolev spaces*, vol. 140. Academic press.

Andrews, D. W. (1994): "Chapter 37 Empirical process methods in econometrics," *Handbook of Econometrics*, 4, 2247 – 2294.

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.

——— (2010): "Asymptotic Size and a Problem With Subsampling and With the $m$ Out Of $n$ Bootstrap," *Econometric Theory*, 26, 426–468.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

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.

Brent, R. P. (1971): "An algorithm with guaranteed convergence for finding a zero of a function," *The Computer Journal*, 14(4), 422–425.

Bugni, F. A., I. A. Canay, and X. Shi (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185(1), 259–282.

——— (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8(1), 1–38.

Bull, A. D. (2011): "Convergence rates of efficient global optimization algorithms," *Journal of Machine Learning Research*, 12(Oct), 2879–2904.

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.

Davydov, Y. A., M. Lifshitz, and N. Smorodina (1995): *Local properties of distributions of stochastic functionals*. American Mathematical Society.

Dekker, T. (1969): "Finding a zero by means of successive linear interpolation," *Constructive aspects of the fundamental theorem of algebra*, pp. 37–51.

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.

Kaido, H., F. Molinari, and J. Stoye (2017): "Constraint qualifications in projection inference," Work in progress.

Mammen, E. (1992): *When Does Bootstrap Work? Asymptotic Results and Simulations*. Spinger Verlag, New York, NY.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

Molchanov, I. (2005): *Theory of Random Sets*. Springer, London.

Narcowich, F., J. Ward, and H. Wendland (2003): "Refined error estimates for radial basis function interpolation," *Constructive approximation*.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2011): "Moment Inequalities and Their Application," Discussion Paper, Harvard University.

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.

Steinwart, I., and A. Christmann (2008): *Support vector machines*. Springer Science & Business Media.

Stoye, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299–1315.

Tartar, L. (2007): *An introduction to Sobolev spaces and interpolation spaces*, vol. 3. Springer Science & Business Media.

van der Vaart, A., and J. Wellner (2000): *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, Berlin.

van der Vaart, A. W., and J. H. van Zanten (2008): "Reproducing kernel Hilbert spaces of Gaussian priors," in *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pp. 200–222. Institute of Mathematical Statistics.