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 CWP26/19

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§

June 5, 2019

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 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 function of θ, not θ 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, and we establish its rate of convergence. The algorithm is of independent interest for optimization problems with simple objectives and complicated constraints. An empirical application estimating an entry game illustrates the usefulness of the method. Monte Carlo simulations confirm the accuracy of the solution algorithm, the good statistical as well as computational performance of calibrated projection (including in comparison to other methods), and the algorithm's potential to greatly accelerate computation of other confidence intervals.

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 Amsterdam, Bonn, BC/BU joint workshop, Brown, Cambridge, Chicago, Cologne, Columbia, Cornell, CREST, Duke, ECARES, Harvard/MIT, Kiel, Kobe, Luxembourg, Mannheim, Maryland, Michigan, Michigan State, NUS, NYU, Penn, Penn State, Rochester, Royal Holloway, SMU, Syracuse, Toronto, Toulouse, UCL, UCLA, UCSD, Vanderbilt, Vienna, Yale, Western, and Wisconsin as well as CEME, Cornell-Penn State IO/Econometrics 2015 Conference, ES Asia Meeting 2016, ES European Summer Meeting 2017, ES North American Winter Meeting 2015, ES World Congress 2015, Frontiers of Theoretical Econometrics Conference (Konstanz), KEA-KAEA International Conference, Notre Dame Second Econometrics Workshop, Verein für Socialpolitik Ausschuss für Ökonometrie 2017. We are grateful to Undral Byambadalai, 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. We gratefully acknowledge financial support through NSF grants SES-1230071 and SES-1824344 (Kaido), SES-0922330 and SES-1824375 (Molinari), and SES-1260980 and SES-1824375 (Stoye).

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

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

§Department of Economics, Cornell University, stoye@cornell.edu.

1 Introduction

This paper provides novel 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. In addition, we develop a new algorithm for computing these confidence intervals and, more generally, for solving optimization problems with "black box" constraints, and obtain its rate of convergence.

Until recently, the rich literature on inference for moment (in)equalities focused on confidence sets for the entire vector θ, 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 the test statistic $T_n(\theta)$ aggregates violations of the sample analog of the moment (in)equalities and the critical value $c_{1-\alpha}(\theta)$ controls asymptotic coverage, often uniformly over a large class of data generating processes (DGPs). However, applied researchers are frequently interested in a specific component (or function) of θ, 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 inference on the projection $p'\theta$, where p is a known unit vector. To date, it is common to report as confidence set the corresponding projection of $\mathcal{C}_n(c_{1-\alpha})$ or the interval

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

which will miss any "gaps" in a disconnected projection but is much easier to compute. This approach yields asymptotically valid but typically conservative and therefore needlessly large confidence regions. 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, the usual 95% confidence interval for θ_k equals $[\hat{\theta}_{n,k} - 1.96, \hat{\theta}_{n,k} + 1.96]$. Yet the analogy to CI_n^{proj} would be projection of a 95% confidence ellipsoid, which with $d = 10$ yields $[\hat{\theta}_{n,k} - 4.28, \hat{\theta}_{n,k} + 4.28]$ and a true coverage of essentially 1.

Our first contribution is to provide a bootstrap-based *calibrated projection* method to largely anticipate and correct for the conservative effect of projection. The method uses an estimated critical level $\hat{c}_{n,1-\alpha}$ calibrated so that the projection of $\mathcal{C}_n(\hat{c}_{n,1-\alpha})$ covers $p'\theta$ (but not necessarily θ) with probability at least $1 - \alpha$. As a confidence region for the true $p'\theta$, one may report this projection, i.e.

$$\{p'\theta : \theta \in \mathcal{C}_n(\hat{c}_{n,1-\alpha})\}, \tag{1.3}$$

or, for computational simplicity and presentational convenience, the interval

$$CI_n \equiv \left[\inf_{\theta \in \mathcal{C}_n(\hat{c}_{n,1-\alpha})} p'\theta, \sup_{\theta \in \mathcal{C}_n(\hat{c}_{n,1-\alpha})} p'\theta \right]. \tag{1.4}$$

We prove uniform asymptotic validity of both over a large class of DGPs.

Computationally, calibration of $\hat{c}_{n,1-\alpha}$ is relatively attractive: We linearize all constraints around θ, so that coverage of $p'\theta$ can be calibrated by analyzing many linear programs. Nonetheless, computing the above objects is challenging in moderately high dimension. This brings us to our second contribution, namely a general method to accurately and rapidly compute confidence intervals whose construction resembles (1.4). Additional applications within partial identification include projection of confidence regions defined in Chernozhukov, Hong, and Tamer (2007), Andrews and Soares (2010), or Andrews and Shi (2013), as well as (with minor tweaking; see Appendix B) the confidence interval proposed in Bugni, Canay, and Shi (2017, BCS henceforth) and further discussed later. In an application to a point identified setting, Freyberger and Reeves (2017, Supplement Section S.3) use our method to construct uniform confidence bands for an unknown function of interest under (nonparametric) shape restrictions. They benchmark it against gridding and find it to be accurate at considerably improved speed. More generally, the method can be broadly used to compute confidence intervals for optimal values of optimization problems with estimated constraints.

Our algorithm (henceforth called E-A-M for Evaluation-Approximation-Maximization) is based on the response surface method, thus it belongs to the family of *expected improvement algorithms* (see e.g. Jones, 2001; Jones, Schonlau, and Welch, 1998, and references therein). Bull (2011) established convergence of an expected improvement algorithm for unconstrained optimization problems where the objective is a "black box" function. The rate of convergence that he derives depends on the smoothness of the black box objective function. We substantially extend his results to show convergence, at a slightly slower rate, of our similar algorithm for constrained optimization problems in which the constraints are sufficiently smooth "black box" functions. Extensive Monte Carlo experiments (see Appendix C and Section 5 of Kaido, Molinari, and Stoye (2017)) confirm that the E-A-M algorithm is fast and accurate.

Relation to existing literature. The main alternative inference prodedure for projections – introduced in Romano and Shaikh (2008) and significantly advanced in BCS – is based on profiling out a test statistic. The classes of DGPs for which calibrated projection and the profiling-based method of BCS (BCS-profiling henceforth) can be shown to be uniformly valid are non-nested.[1]

Computationally, calibrated projection has the advantage that the bootstrap iterates over linear as opposed to nonlinear programming problems. While the "outer" optimization problems in (1.4) are potentially intricate, our algorithm is geared toward them. Monte Carlo

[1] See Kaido, Molinari, and Stoye (2017, Section 4.2 and Supplemental Appendix F) for a comparison of the statistical properties of calibrated projection and BCS-profiling, summarized here at the end of Section 3.2.

simulations suggest that these two factors give calibrated projection a considerable computational edge over profiling, though profiling can also benefit from the E-A-M algorithm. Indeed, in Appendix C we replicate the Monte Carlo experiment of BCS and find that adapting E-A-M to their method improves computation time by a factor of about 4, while switching to calibrated projection improves it by a further factor of about 17.

In an influential paper, Pakes, Porter, Ho, and Ishii (2011, PPHI henceforth) also use linearization but, subject to this approximation, directly bootstrap the sample projection. This is valid only under stringent conditions.[2] Other related articles that explicitly consider inference on projections include Beresteanu and Molinari (2008), Bontemps, Magnac, and Maurin (2012), Kaido (2016), and Kline and Tamer (2016). None of these establish uniform validity of confidence sets. Chen, Christensen, and Tamer (2018) establish uniform validity of MCMC-based confidence intervals for projections, but aim at covering the projection of the entire identified region $\Theta_I(P)$ (defined later) and not just of the true θ. Gafarov, Meier, and Montiel-Olea (2016) use our insight in the context of set identified spatial VARs.

Regarding computation, previous implementations of projection-based inference (e.g., Ciliberto and Tamer, 2009; Grieco, 2014; Dickstein and Morales, 2018) reported the smallest and largest value of $p'\theta$ among parameter values $\theta \in \mathcal{C}_n(c_{1-\alpha})$ that were discovered using, e.g., grid-search or simulated annealing with no cooling. This becomes computationally cumbersome as d increases because it typically requires a number of evaluation points that grows exponentially with d. In contrast, using a probabilistic model, our method iteratively draws evaluation points from regions that are considered highly relevant for finding the confidence interval's end point. In applications, this tends to substantially reduce the number of evaluation points.

Structure of the paper. Section 2 sets up notation and describes our approach in detail, including computational implementation of the method and choice of tuning parameters. Section 3.1 establishes uniform asymptotic validity of CI_n, and Section 3.2 shows that our algorithm converges at a specific rate which depends on the smoothness of the constraints. Section 4 reports the results of an empirical application that revisits the analysis in Kline and Tamer (2016, Section 8). Section 5 draws conclusions. The proof of convergence of our algorithm is in Appendix A. Appendix B shows that our algorithm can be used to compute BCS-profiling confidence intervals. Appendix C reports the results of Monte Carlo simulations comparing our proposed method with that of BCS. All other proofs, background material for our algorithm, and additional results are in the Online Appendix.[3]

[2] The published version of PPHI, i.e. Pakes, Porter, Ho, and Ishii (2015), does not contain the inference part. Kaido, Molinari, and Stoye (2017, Section 4.2) show that calibrated projection can be much simplified under the conditions imposed by PPHI.

[3] Appendix D provides convergence-related results and background material for our algorithm and describes how to compute $\hat{c}_{n,1-\alpha}(\theta)$. Appendix E presents the assumptions under which we prove uniform asymptotic validity of CI_n. Appendix F verifies, for a number of canonical partial identification problems, the assumptions that we invoke to show validity of our inference procedure and for our algorithm. Appendix G contains the proof of Theorem 3.1. Appendix H collects Lemmas supporting this proof.

2 Detailed Explanation of the Method

2.1 Setup and Definition of CI_n

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 known measurable functions characterizing the model. The true parameter value θ is assumed to satisfy the moment inequality and equality restrictions

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

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

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

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

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

for the sample moments and the analog estimators of the population moment functions' standard deviations $\sigma_{P,j}$. The confidence interval in (1.4) then is

$$CI_n = [-s(-p, \mathcal{C}_n(\hat{c}_{n,1-\alpha})), s(p, \mathcal{C}_n(\hat{c}_{n,1-\alpha}))] \tag{2.5}$$

with

$$s(p, \mathcal{C}_n(\hat{c}_{n,1-\alpha})) \equiv \sup_{\theta \in \Theta} \ p'\theta \text{ s.t. } \sqrt{n}\frac{\bar{m}_{n,j}(\theta)}{\hat{\sigma}_{n,j}(\theta)} \leqslant \hat{c}_{n,1-\alpha}(\theta), \ j = 1, \ldots, J \tag{2.6}$$

and similarly for $(-p)$. Henceforth, to simplify notation, we write $\hat{c}_n$ for $\hat{c}_{n,1-\alpha}$. We also 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 $k = 1, \ldots, J_2$. That is, we treat moment equality constraints as two opposing inequality constraints.

For a class of DGPs $\mathcal{P}$ that we specify below, define the asymptotic size of CI_n by[4]

$$\liminf_{n\to\infty} \inf_{P \in \mathcal{P}} \inf_{\theta \in \Theta_I(P)} P(p'\theta \in CI_n). \tag{2.7}$$

We next explain how to control this size and then how to compute CI_n.

2.2 Calibration of $\hat{c}_n(\theta)$

Calibration of $\hat{c}_n$ requires careful analysis of the moment restrictions' local behavior at each point in the identification region. This is because the extent of projection conservatism

[4]Here we focus on the confidence interval CI_n defined in (1.4). See Appendix G.2.3 for the analysis of the confidence region given by the mathematical projection in (1.3).

depends on (i) the asymptotic behavior of the sample moments entering the inequality restrictions, which can change discontinuously depending on whether they bind at θ or not, and (ii) the local geometry of the identification region at θ, i.e. the shape of the constraint set formed by the moment restrictions. Features (i) and (ii) can be quite different at different points in $\Theta_I(P)$, making uniform inference 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.

To build an intuition, fix $P \in \mathcal{P}$ and $\theta \in \Theta_I(P)$. The projection of θ is covered when

$$\begin{aligned}
&\left\{ \begin{array}{c} \inf_{\vartheta \in \Theta} p'\vartheta \\ \text{s.t. } \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leqslant \hat{c}_n(\vartheta), \forall j \end{array} \right\} \leqslant p'\theta \leqslant \left\{ \begin{array}{c} \sup_{\vartheta \in \Theta} p'\vartheta \\ \text{s.t. } \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leqslant \hat{c}_n(\vartheta), \forall j \end{array} \right\} \\
\iff &\left\{ \begin{array}{c} \inf_{\lambda \in \sqrt{n}(\Theta-\theta)} p'\lambda \\ \text{s.t. } \frac{\sqrt{n}\bar{m}_{n,j}(\theta+\lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta+\lambda/\sqrt{n})} \leqslant \hat{c}_n\left(\theta+\lambda/\sqrt{n}\right), \forall j \end{array} \right\} \leqslant 0 \leqslant \left\{ \begin{array}{c} \sup_{\lambda \in \sqrt{n}(\Theta-\theta)} p'\lambda \\ \text{s.t. } \frac{\sqrt{n}\bar{m}_{n,j}(\theta+\lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta+\lambda/\sqrt{n})} \leqslant \hat{c}_n\left(\theta+\lambda/\sqrt{n}\right), \forall j \end{array} \right\} \\
\impliedby &\left\{ \begin{array}{c} \inf_{\lambda \in \sqrt{n}(\Theta-\theta)\cap \rho B^d} p'\lambda \\ \text{s.t. } \frac{\sqrt{n}\bar{m}_{n,j}(\theta+\lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta+\lambda/\sqrt{n})} \leqslant \hat{c}_n\left(\theta+\lambda/\sqrt{n}\right), \forall j \end{array} \right\} \leqslant 0 \leqslant \left\{ \begin{array}{c} \sup_{\lambda \in \sqrt{n}(\Theta-\theta)\cap \rho B^d} p'\lambda \\ \text{s.t. } \frac{\sqrt{n}\bar{m}_{n,j}(\theta+\lambda/\sqrt{n})}{\hat{\sigma}_{n,j}(\theta+\lambda/\sqrt{n})} \leqslant \hat{c}_n\left(\theta+\lambda/\sqrt{n}\right), \forall j \end{array} \right\}.
\end{aligned} \tag{2.8}$$

Here, we first substituted $\vartheta = \theta + \lambda/\sqrt{n}$ and took λ to be the choice parameter; intuitively, this localizes around θ at rate $1/\sqrt{n}$. We then make the event smaller by adding the constraint $\lambda \in \rho B^d$, with $B^d \equiv [-1,1]^d$ and $\rho \geqslant 0$ a tuning parameter. We motivate this step later.

Our goal is to set the probability of (2.8) equal to $1-\alpha$. To ease computation, we approximate (2.8) by linear expansion in λ of the constraint set. For each j, 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

$$\frac{\sqrt{n}\bar{m}_{n,j}\left(\theta+\lambda/\sqrt{n}\right)}{\hat{\sigma}_{n,j}\left(\theta+\lambda/\sqrt{n}\right)} = \left(\mathbb{G}_{n,j}\left(\theta+\lambda/\sqrt{n}\right) + D_{P,j}(\bar{\theta})\lambda + \sqrt{n}\gamma_{1,P,j}(\theta)\right)\frac{\sigma_{P,j}\left(\theta+\lambda/\sqrt{n}\right)}{\hat{\sigma}_{n,j}\left(\theta+\lambda/\sqrt{n}\right)}. \tag{2.9}$$

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 θ and $\theta+\lambda/\sqrt{n}$.[5]

We formally establish that the probability of the last event in (2.8) can be approximated by the probability that 0 lies between the optimal values of two stochastic linear programs. The components that characterize these programs can be estimated. Specifically, we replace $D_{P,j}(\cdot)$ with a uniformly consistent (on compact sets) estimator, $\hat{D}_{n,j}(\cdot)$,[6] and the process $\mathbb{G}_{n,j}(\cdot)$ with its simple nonparametric bootstrap analog, $\mathbb{G}^b_{n,j}(\cdot) \equiv n^{-1/2}\sum_{i=1}^n (m_j(X_i^b,\cdot) - \bar{m}_{n,j}(\cdot))/\hat{\sigma}_{n,j}(\cdot)$.[7] Estimation of $\gamma_{1,P,j}(\theta)$ is more subtle because it enters (2.9) scaled by $\sqrt{n}$,

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

[6] See Online Appendix F for such estimators in some canonical moment (in)equality examples.

[7] BCS approximate $\mathbb{G}_{n,j}(\cdot)$ by $n^{-1/2}\sum_{i=1}^n[(m_j(X_i,\cdot) - \bar{m}_{n,j}(\cdot))/\hat{\sigma}_{n,j}(\cdot)]\chi_i$ with $\{\chi_i \sim N(0,1)\}_{i=1}^n$ i.i.d. This

so that a sample analog estimator will not do. However, this specific issue is well understood in the moment inequalities literature. Following Andrews and Soares (2010, AS henceforth) and others (Bugni, 2010; Canay, 2010; Stoye, 2009), we shrink this sample analog toward zero, leading to conservative (if any) distortion in the limit. Formally, we estimate $\gamma_{1,P,j}(\theta)$ by $\varphi(\hat{\xi}_{n,j}(\theta))$, where $\varphi : \mathbb{R}^J_{[\pm\infty]} \mapsto \mathbb{R}^J_{[\pm\infty]}$ is one of the Generalized Moment Selection (GMS henceforth) functions proposed by AS,

$$\hat{\xi}_{n,j}(\theta) \equiv \begin{cases} \kappa_n^{-1}\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta) & j = 1, \ldots, J_1 \\ 0 & j = J_1 + 1, \ldots, J, \end{cases} \tag{2.10}$$

and $\kappa_n \to \infty$ is a user-specified thresholding sequence.[8] In sum, we replace the random constraint set in (2.8) with the (bootstrap based) random polyhedral set[9]

$$\Lambda_n^b(\theta, \rho, c) \equiv \big\{\lambda \in \sqrt{n}(\Theta - \theta) \cap \rho B^d : \mathbb{G}_{n,j}^b(\theta) + \hat{D}_{n,j}(\theta)\lambda + \varphi_j(\hat{\xi}_{n,j}(\theta)) \leqslant c, j = 1, \ldots, J\big\}. \tag{2.11}$$

The critical level $\hat{c}_n(\theta)$ to be used in (2.6) then is

$$\hat{c}_n(\theta) \equiv \inf\left\{c \in \mathbb{R}_+ : P^*\left(\min_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda \leqslant 0 \leqslant \max_{\lambda \in \Lambda_n^b(\theta,\rho,c)} p'\lambda\right) \geqslant 1 - \alpha\right\} \tag{2.12}$$

$$= \inf\big\{c \in \mathbb{R}_+ : P^*(\Lambda_n^b(\theta, \rho, c) \cap \{p'\lambda = 0\} \neq \varnothing) \geqslant 1 - \alpha\big\}, \tag{2.13}$$

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. Expression (2.13) uses convexity of $\Lambda_n^b(\theta, \rho, c)$ and reveals that the probability inside curly brackets can be assessed by repeatedly checking feasibility of a linear program.[10] We describe in detail in Online Appendix D.4 how we compute $\hat{c}_n(\theta)$ through a root-finding algorithm.

We conclude by motivating the "ρ-box constraint" in (2.8), which is a major novel contribution of this paper. The constraint induces conservative bias but has two fundamental benefits: First, it ensures that the linear approximation of the feasible set in (2.8) by (2.11) is used only in a neighborhood of θ, and therefore that it is uniformly accurate. More subtly,

approximation is equally valid in our approach, and can be faster as it avoids repeated evaluation of $m_j(X_i^b, \cdot)$.

[8] A common choice of φ is given component-wise by

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

Restrictions on φ and the rate at which κ_n diverges are imposed in Assumption E.2. While for concreteness here we write out the "hard thresholding" GMS function, Theorem 3.1 below applies to all but one of the GMS functions in AS, namely to $\varphi^1 - \varphi^4$, all of which depend on $\kappa_n^{-1}\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)$. We do not consider GMS function φ^5, which depends also on the covariance matrix of the moment functions.

[9] Here, we implicitly assume that Θ is a polyhedral set. If it is instead defined by smooth convex (in)equalities, these can be linearized too.

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

it ensures that coverage induced by a given c depends continuously on estimated parameters even in certain intricate cases. This renders calibrated projection valid in cases that other methods must exclude by assumption.[11]

2.3 Computation of CI_n and of Similar Confidence Intervals

Projection based methods as in (1.2) and (1.4) have nonlinear constraints involving a critical value which in general is an unknown function, with unknown gradient, of θ. Similar considerations often apply to critical values used to build confidence intervals for optimal values of optimization problems with estimated constraints. 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 θ is cheap.

This concern motivates this paper's second main contribution, namely a novel algorithm for constrained optimization problems of the following form:

$$\begin{aligned} p'\theta^* \equiv \sup_{\theta \in \Theta} \; & p'\theta \\ \text{s.t.} \quad & g_j(\theta) \leqslant c(\theta), \; j = 1, ..., J, \end{aligned} \tag{2.14}$$

where θ^* is an optimal solution of the problem and $g_j(\cdot), j = 1, ..., J$ as well as $c(\cdot)$ are fixed functions of θ. In our own application, $g_j(\theta) = \sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)$ and, for calibrated projection, $c(\theta) = \hat{c}_n(\theta)$.[12]

The key issue is that evaluating $c(\cdot)$ is costly.[13] Our algorithm does so at relatively few values of θ. Elsewhere, it approximates $c(\cdot)$ through a probabilistic model that gets updated as more values are computed. We use this model to determine the next evaluation point but report as tentative solution the best value of θ at which $c(\cdot)$ was computed, *not* a value at which it was merely approximated. Under reasonable conditions, the tentative optimal values converge to $p'\theta^*$ at a rate (relative to iterations of the algorithm) that is formally established in Section 3.2.

After drawing an initial set of evaluation points that we set to grow linearly with d, the algorithm has three steps called E, A, and M below.

[11] In (2.11), set $(\mathbb{G}^b_{n,1}(\cdot), \mathbb{G}^b_{n,2}(\cdot)) \sim N(0, I_2)$, $p = \hat{D}_{n,1} = \hat{D}_{n,2} = (0, 1)$, $\varphi_1(\cdot) = \varphi_2(\cdot) = 0$, and $\alpha = .05$. Then simple algebra reveals that (with or without ρ-box) $\hat{c}_n(\cdot) = \Phi^{-1}(\sqrt{.95}) \approx 1.95$. If $\hat{D}_{n,1} = (0, 1-\delta)$ and $\hat{D}_{n,2} = (0, 1-\delta)$, then without ρ-box we have $\hat{c}_n(\cdot) = \Phi^{-1}(.95)/\sqrt{2} \approx 1.16$ for any small $\delta > 0$, and we therefore cannot expect to get $\hat{c}_n(\cdot)$ right if gradients are estimated. With ρ-box, $\hat{c}_n(\cdot) \to 1.95$ as $\delta \to 0$, so the problem goes away. This stylized example is relevant because it resembles polyhedral identified sets where one face is near orthogonal to p. It violates assumptions in BCS and PPHI.

[12] We emphasize that, in analyzing the computational problem, we take the data, including bootstrap data, as given. Thus, while an econometrician would usually think of $\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)$ and $\hat{c}_n(\theta)$ as random variables, for this section's purposes they are indeed just functions of θ.

[13] For simplicity and to mirror our motivating application, we suppose that $g_j(\cdot)$ is easy to compute. The algorithm is easily adapted to the case where it is not. Indeed, in Appendix B, we show how E-A-M can be employed to compute BCS-profiling confidence intervals, where the profiled test statistic itself is costly to compute and is approximated together with the critical value.

Initialization: Draw randomly (uniformly) over Θ a set $(\theta^{(1)}, ..., \theta^{(k)})$ of initial evaluation points. Evaluate $c(\theta^{(\ell)})$ for $\ell = 1, ..., k-1$. Initialize $L = k$.

E-Step: Evaluate $c(\theta^{(L)})$ and record the tentative optimal value

$$p'\theta^{*,L} \equiv \max\{p'\theta^{(\ell)} : \ell \in \{1, ..., L\}, \bar{g}(\theta) \leqslant c(\theta^{(\ell)})\}, \tag{2.15}$$

with $\bar{g}(\theta) = \max_{j=1,...,J} g_j(\theta)$.

A-step: 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 $\zeta(\cdot)$ indexed by θ and with constant variance ς^2 specifies

$$\Upsilon^{(\ell)} = \mu + \zeta(\theta^{(\ell)}), \ \ell = 1, ..., L, \tag{2.16}$$

$$Corr(\zeta(\theta), \zeta(\theta')) = K_\beta(\theta - \theta'), \ \theta, \theta' \in \Theta, \tag{2.17}$$

where $\Upsilon^{(\ell)} = c(\theta^{(\ell)})$ and K_β is a kernel with parameter vector $\beta \in \times_{h=1}^d [\underline{\beta}_h, \overline{\beta}_h] \subset \mathbb{R}^d_{++}$; e.g., $K_\beta(\theta - \theta') = \exp(-\sum_{h=1}^d |\theta_h - \theta'_h|^2/\beta_h)$. The unknown parameters (μ, ς^2) can be estimated by running a GLS regression of $\boldsymbol{\Upsilon} = (\Upsilon^{(1)}, ..., \Upsilon^{(L)})'$ on a constant with the given correlation matrix. The unknown parameters β can be estimated by a (concentrated) MLE.

The (best linear) predictor of the critical value and its gradient at θ are then given by

$$c_L(\theta) = \hat{\mu} + \mathbf{r}_L(\theta)'\mathbf{R}_L^{-1}(\boldsymbol{\Upsilon} - \hat{\mu}\mathbf{1}), \tag{2.18}$$

$$\nabla_\theta c_L(\theta) = \hat{\mu} + \mathbf{Q}_L(\theta)\mathbf{R}_L^{-1}(\boldsymbol{\Upsilon} - \hat{\mu}\mathbf{1}), \tag{2.19}$$

where $\mathbf{r}_L(\theta)$ is a vector whose ℓ-th component is $Corr(\zeta(\theta), \zeta(\theta^{(\ell)}))$ 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 (ℓ, ℓ') entry is $Corr(\zeta(\theta^{(\ell)}), \zeta(\theta^{(\ell')}))$ with estimated parameters. This surrogate model has the property that its predictor satisfies $c_L(\theta^{(\ell)}) = c(\theta^{(\ell)}), \ell = 1, ..., L$. Hence, it provides an analytical interpolation, with analytical gradient, of evaluation points of $c(\cdot)$.[14] The uncertainty left in $c(\cdot)$ is captured by the variance

$$\hat{\varsigma}^2 s_L^2(\theta) = \hat{\varsigma}^2 \left(1 - \mathbf{r}_L(\theta)'\mathbf{R}_L^{-1}\mathbf{r}_L(\theta) + \frac{(1 - \mathbf{1}'\mathbf{R}_L^{-1}\mathbf{r}_L(\theta))^2}{\mathbf{1}'\mathbf{R}_L^{-1}\mathbf{1}}\right). \tag{2.20}$$

M-step: With probability $1 - \epsilon$, obtain the next evaluation point $\theta^{(L+1)}$ as

$$\theta^{(L+1)} \in \arg\max_{\theta \in \Theta} \mathbb{EI}_L(\theta) = \arg\max_{\theta \in \Theta} (p'\theta - p'\theta^{*,L})_+ \left(1 - \Phi\left(\frac{\bar{g}(\theta) - c_L(\theta)}{\hat{\varsigma} s_L(\theta)}\right)\right), \tag{2.21}$$

[14] 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 empirical application and Monte Carlo experiments.

where $\mathbb{EI}_L(\theta)$ is the *expected improvement function*.[15] This step can be implemented by standard nonlinear optimization solvers, e.g. MATLAB's `fmincon` or `KNITRO` (see Appendix D.3 for details). With probability ϵ, draw $\theta^{(L+1)}$ randomly from a uniform distribution over Θ. Set $L \leftarrow L+1$ and return to the E-step.

The algorithm yields an increasing sequence of tentative optimal values $p'\theta^{*,L}, L = k+1, k+2, ...$, with $\theta^{*,L}$ satisfying the *true* constraints in (2.14) but the sequence of evaluation points leading to it obtained by maximization of expected improvement defined with respect to the *approximated* surface. Once a convergence criterion is met, $p'\theta^{*,L}$ is reported as the end point of CI_n. We discuss convergence criteria in Appendix C.

The advantages of E-A-M are as follows. First, we control the number of points at which we evaluate the critical value; recall that this evaluation is the expensive step. Also, the initial k evaluations can easily be parallelized. For any additional E-step, 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, it trades off "exploitation" (i.e. the benefit of drawing a point at which the optimal value is high) against "exploration" (i.e. the benefit of drawing a point in a region in which the approximation error of c is currently large) through maximizing expected improvement.[16] Finally, the algorithm simplifies the M-step by providing constraints and their gradients for program (2.21) in closed form, thus greatly aiding fast and stable numerical optimization. The price is the additional approximation step. In the empirical application in Section 4 and in the numerical exercises of Appendix C, this price turns out to be low.

2.4 Choice of Tuning Parameters

Practical implementation of calibrated projection and the E-A-M algorithm is detailed in Kaido, Molinari, Stoye, and Thirkettle (2017). It involves setting several tuning parameters, which we now discuss.

Calibration of $\hat{c}_n$ in (2.13) must be tuned at two points, namely the use of GMS and the choice of ρ. The trade-offs in setting these tuning parameters are apparent from inspection of (2.11). GMS is parameterized by a shrinkage function φ and a sequence κ_n that controls the rate of shrinkage. In practice, choice of κ_n is more delicate. A smaller κ_n will make Λ_n^b larger, hence increase bootstrap coverage probability for any given c, hence reduce $\hat{c}_n$ and therefore make for shorter confidence intervals – but the uniform asymptotics will be misleading, and finite sample coverage therefore potentially off target, if κ_n is too small. We follow the industry standard set by AS and recommend $\kappa_n = \sqrt{\log n}$.

[15] Heuristically, $\mathbb{EI}_L(\theta)$ is the expected improvement gained from analyzing parameter value θ for a Bayesian whose current beliefs about c are described by the estimated model. Indeed, for each θ, the maximand in (2.21) multiplies improvement from learning that θ is feasible with this Bayesian's probability that it is.

[16] It is also possible to draw multiple points in each iteration (Schonlau, Welch, and Jones, 1998), as we do in our implementation of the method.

The trade-off in choosing ρ is similar but reversed. A larger ρ will expand Λ_n^b and therefore make for shorter confidence intervals, but (our proof of) uniform validity of inference requires $\rho < \infty$. Indeed, calibrated projection with $\rho = 0$ will disregard any projection conservatism and (as is easy to show) exactly recovers projection of the AS confidence set. Intuitively, we then want to choose ρ large but not too large.

To this end, we heuristically calibrate ρ based on how much conservative distortion one is willing to accept in well-behaved cases. This distortion – denote it η, for which we suggest a numerical value of 0.01 – is compared against a bound on conservative distortion that is itself likely to be conservative but data free and trivial to compute. In particular, we set

$$\rho = \Phi^{-1}\left(\tfrac{1}{2} + \tfrac{1}{2}\left(1 - \eta/\tbinom{J_1+J_2}{d}\right)^{1/d}\right). \tag{2.22}$$

The underlying heuristic is as follows: If all basic solutions (i.e., intersections of exactly d constraints) that potentially define vertices of Λ_n^b realize inside the ρ-box, then the ρ-box cannot affect the values in (2.12) and hence not whether coverage obtains in a given bootstrap sample. Conversely, the probability that at least one basic solution realizes outside the ρ-box bounds from above the conservative distortion. This probability is, of course, dependent on unknown parameters. Our data free approximation imputes multivariate standard normal distributions for all basic solutions and Bonferroni adjustment to handle their covariation.[17]

The E-A-M algorithm also has two tuning parameters. One is k, the initial number of evaluation points. The other is ϵ, the probability of drawing $\theta^{(L+1)}$ randomly from a uniform distribution on Θ instead of by maximizing $\mathbb{EI}_L$. In calibrated projection use of the E-A-M algorithm there is a single "black box" function, $\hat{c}_n(\theta)$. We therefore suggest setting $k = 10d + 1$, similarly to the recommendation in Jones, Schonlau, and Welch (1998, p. 473). In our Monte Carlo exercises we experimented with larger values, e.g. $k = 20d + 1$, and found that the increased number had no noticeable effect on the computed CI_n. If a user applies our E-A-M algorithm to a constrained optimization problem with *many* "black box" functions to approximate, we suggest using a larger number of initial points.

The role of ϵ (e.g., Bull, 2011, p. 2889) is to trade off the greediness of the $\mathbb{EI}_L$ maximization criterion with the overarching goal of global optimization. Sutton and Barto (1998, pp. 28-29) explore the effect of setting $\epsilon = 0.1$ and 0.01 on different optimization problems, and find that for sufficiently large L, $\epsilon = 0.01$ performs better. In our own simulations we have found that drawing *both* a uniform point and computing the value of θ for each L (thereby sidestepping the choice of ϵ) is fast and accurate, and that is what we recommend doing.

[17] To reproduce the expression, recall that if $a \equiv \binom{J_1+J_2}{d}$ random variables in $\mathbb{R}^d$ are individually multivariate standard normal, then a Bonferroni upper bound on the probability that *not* all of them realize inside the ρ-box equals $a\big(1 - (1 - 2\Phi(-\rho))^d\big)$. Also, if Bonferroni is replaced with an independence assumption, the expression changes to $\rho = \Phi^{-1}\big(\frac{1}{2} + \frac{1}{2}(1-\eta)^{1/ad}\big)$. The numerical difference is negligible for moderate $J_1 + J_2$.

3 Theoretical Results

3.1 Asymptotic Validity of Inference

In this section we establish that CI_n is uniformly asymptotically valid in the sense of ensuring that (2.7) equals at least $1-\alpha$. The result applies to: (i) Confidence intervals for one projection; (ii) joint confidence regions for several projections, in particular confidence hyperrectangles for subvectors; (iii) confidence intervals for smooth nonlinear functions $f : \Theta \mapsto \mathbb{R}$. Examples of the latter extension include policy analysis and estimation of partially identified counterfactuals as well as demand extrapolation subject to rationality constraints.[18]

THEOREM 3.1: *Suppose Assumptions E.1, E.2, E.3, E.4, and E.5 hold. Let* $0 < \alpha < 1/2$.

(I) Let CI_n *be as defined in* (1.4)*, with* $\hat{c}_n$ *as in* (2.13)*. Then:*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(p'\theta \in CI_n) \geqslant 1-\alpha. \tag{3.1}$$

(II) Let $p^1, \ldots, p^h$ *denote unit vectors in* $\mathbb{R}^d$*,* $h \leqslant d$*. Then:*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(p^{k\prime}\theta \in CI_{n,k}, k = 1, \ldots, h) \geqslant 1-\alpha, \tag{3.2}$$

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

(III) Let CI_n^f *be a confidence interval whose lower and upper points are obtained solving*

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

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 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* $f(\theta)$*.*[19] *Let* $0 < \alpha < 1/2$*. Then:*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(f(\theta) \in CI_n^f) \geqslant 1-\alpha. \tag{3.3}$$

All assumptions can be found in Online Appendix E.1. Assumptions E.1 and E.5 are mild regularity conditions typical in the literature; see, e.g., Definition 4.2 and the corresponding discussion in BCS. Assumption E.2 is based on AS and constrains the GMS function $\varphi(\cdot)$

[18] In Appendix G.2.3, we show that the result actually applies to the mathematical projection in (1.3).

[19] Because the function f is known, these conditions can be easily verified in practice (especially if the first one is strengthened to hold over Θ).

as well as the rate at which κ_n diverges. Assumption E.4 requires normalized population moments to be sufficiently smooth and consistently estimable. Assumption E.3 is our key departure from the related literature. In essence, it requires that the correlation matrix of the moment functions corresponding to close-to-binding moment conditions has eigenvalues uniformly bounded from below.[20] Under this condition, we are able to show that in the limit problem corresponding to (2.8) –where constraints are replaced with their local linearization using population gradients and Gaussian processes– the probability of coverage increases continuously in c. If such continuity is directly assumed (Assumption E.6), Theorem 3.1 remains valid (Online Appendix G.2.2). While the high level Assumption E.6 is similar in spirit to a key condition (Assumption A.2) in BCS, we propose Assumption E.3 due to its familiarity and ease of interpretation; a similar condition is required for uniform validity of standard point identified Generalized Method of Moments inference. In Online Appendix F.2 we verify that our assumptions hold in some of the canonical examples in the partial identification literature: mean with missing data, linear regression and best linear prediction with interval data (and discrete covariates), entry games with multiple equilibria (and discrete covariates), and semi-parametric binary regression models with discrete or interval valued covariates (as in Magnac and Maurin, 2008).

Assumptions E.1-E.5 define the class of DGPs over which our proposed method yields uniformly asymptotically valid coverage. This class is non-nested with the class of DGPs over which the profiling-based methods of Romano and Shaikh (2008) and BCS are uniformly asymptotically valid. Kaido, Molinari, and Stoye (2017, Section 4.2 and Supplemental Appendix F) show that in well behaved cases, calibrated projection and BCS-profiling are asymptotically equivalent. They also provide conditions under which calibrated projection has lower probability of false coverage in finite sample, thereby establishing that the two methods' finite sample power properties are non-ranked.

3.2 Convergence of the E-A-M Algorithm

We next provide formal conditions under which the sequence $p'\theta^{*,L}$ generated by the E-A-M algorithm converges to the true end point of CI_n as $L \to \infty$ at a rate that we obtain. Although $p'\theta^{*,L} = \max\{p'\theta^{(\ell)} : \ell \in \{1, ..., L\}, \bar{g}(\theta) \leqslant c(\theta^{(\ell)})\}$, so that $\theta^{*,L}$ satisfies the *true* constraints for each L, the sequence of evaluation points $\theta^{(\ell)}$ is mostly obtained through expected improvement maximization (M-Step) with respect to the *approximating* surface $c_L(\cdot)$. Because of this, a requirement for convergence is that the function $c(\cdot)$ is sufficiently smooth, so that the approximation error in $|c(\theta) - c_L(\theta)|$ vanishes uniformly in θ as $L \to \infty$.[21] We furthermore assume that the constraint set in (2.14) satisfies a degeneracy condition

[20] Assumption E.3 allows for high correlation among moment inequalities that cannot cross. This covers equality constraints but also entry games as the ones studied in Ciliberto and Tamer (2009).

[21] As in Bull (2011), our convergence result accounts for the fact that the parameters of the Gaussian process prior in (2.16) are re-estimated for each iteration of the A-step using the "training data" $\{\theta^{\ell}, c(\theta^{\ell})\}_{\ell=1}^{L}$.

introduced to the partial identification literature by Chernozhukov, Hong, and Tamer (2007, Condition C.3).[22] In our application, the condition requires that $\mathcal{C}_n(\hat{c}_n)$ has an interior and that the inequalities in (2.6), when evaluated at points in a (small) τ-contraction of $\mathcal{C}_n(\hat{c}_n)$, are satisfied with a slack that is proportional to τ. Theorem 3.2 below establishes that these conditions jointly ensure convergence of the E-A-M algorithm at a specific rate. This is a novel contribution to the literature on response surface methods for constrained optimization.

In the formal statement below, the expectation $E_{\mathbb{Q}}$ is taken with respect to the law of $(\theta^{(1)}, ..., \theta^{(L)})$ determined by the Initialization step and the M-step but conditioning on the sample. We refer to Appendix A for a precise definition of $E_{\mathbb{Q}}$ and a proof of the theorem.

THEOREM 3.2: *Suppose $\Theta \subset \mathbb{R}^d$ is a compact hyperrectangle with nonempty interior, that $\|p\| = 1$, and that Assumptions A.1, A.2, and A.3 hold. Let the evaluation points $(\theta^{(1)}, \cdots, \theta^{(L)})$ be drawn according to the Initialization and M-steps. Then*

$$\|p'\theta^* - p'\theta^{*,L}\|_{L^1_{\mathbb{Q}}} = O\Big(\Big(\frac{L}{\ln L}\Big)^{-\nu/d} (\ln L)^{\delta}\Big), \tag{3.4}$$

where $\|\cdot\|_{L^1_{\mathbb{Q}}}$ is the L^1-norm under $\mathbb{Q}$, $\delta \geqslant 1 + \chi$, and the constants $0 < \nu \leqslant \infty$ and $0 < \chi < \infty$ are defined in Assumption A.1. If $\nu = \infty$, the statement in (3.4) holds for any $\nu < \infty$.

The requirement that Θ is a compact hyperrectangle with nonempty interior can be replaced by a requirement that Θ belongs to the interior of a closed hyperrectangle in $\mathbb{R}^d$. Assumption A.1 specifies the types of kernel to be used to define the correlation functional in (2.17). Assumption A.2 collects requirements on differentiability of $g_j(\theta), j = 1, \ldots, J$, and smoothness of $c(\theta)$. Assumption A.3 is the degeneracy condition discussed above.

To apply Theorem 3.2 to calibrated projection, we provide low level conditions (Assumption D.1 in Online Appendix D.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 $\hat{c}_{n,\tau_n}$ in equation (D.1), where $\tau_n = o(n^{-1/2})$.[23] Theorem D.1 in the Online Appendix shows that $\hat{c}_n$ and $\hat{c}_{n,\tau_n}$ can be made uniformly arbitrarily close, and that $\hat{c}_{n,\tau_n}$ yields valid inference as in (3.1). In practice, we directly apply the E-A-M steps to $\hat{c}_n$.

The key condition imposed in Theorem D.1 is Assumption D.1. It requires that the GMS function used is Lipschitz in its argument,[24] and that the standardized moment functions are Lipschitz in θ. In Online Appendix F.1 we establish that the latter condition is satisfied by some canonical examples in the moment (in)equality literature: mean with missing data, linear regression and best linear prediction with interval data (and discrete covariates), entry games with multiple equilibria (and discrete covariates), and semi-parametric binary regres-

[22] Chernozhukov, Hong, and Tamer (2007, eq. (4.6)) impose the condition on the population identified set.
[23] For a discussion of mollification, see e.g. Rockafellar and Wets (2005, Example 7.19).
[24] This requirement rules out the GMS function in footnote 8, but it is satisfied by other GMS functions proposed by AS.

sion models with discrete or interval valued covariates (as in Magnac and Maurin, 2008).[25]

The E-A-M algorithm is proposed as a method to implement our statistical procedure, not as part of the statistical procedure itself. As such, its approximation error is not taken into account in Theorem 3.1. Our comparisons of the confidence intervals obtained through the use of E-A-M as opposed to directly solving problems (2.6) through the use of MATLAB's `fmincon` in our empirical application in the next section suggest that such error is minimal.

4 Empirical Illustration: Estimating a Binary Game

We employ our method to revisit the study in Kline and Tamer (2016, Section 8) of "what explains the decision of an airline to provide service between two airports." We use their data and model specification.[26] Here we briefly summarize the set-up and refer to Kline and Tamer (2016) for a richer discussion.

The study examines entry decisions of two types of firms, namely Low Cost Carriers (LCC) versus Other Airlines (OA). A market is defined as a trip between two airports, irrespective of intermediate stops. The entry decision $Y_{\ell,i}$ of player $\ell \in \{LCC, OA\}$ in market i is recorded as a 1 if a firm of type ℓ serves market i and 0 otherwise. Firm ℓ's payoff equals $Y_{\ell,i}(Z'_{\ell,i}\vartheta_\ell + \delta_i Y_{-\ell,i} + u_{\ell,i})$, where $Y_{-\ell,i}$ is the opponent's entry decision. Each firm enters if doing so generates non-negative payoffs. The observable covariates in the vector $Z_{\ell,i}$ include the constant and the variables W_i^{size} and $W_{\ell,i}^{pres}$. The former is market size, a market-specific variable common to all airlines in that market and defined as the population at the endpoints of the trip. The latter is a firm-and-market-specific variable measuring the market presence of firms of type ℓ in market i (see Kline and Tamer, 2016, p. 356 for its exact definition). While W_i^{size} enters the payoff function of both firms, $W_{LCC,i}^{pres}$ (respectively, $W_{OA,i}^{pres}$) is excluded from the payoff of firm OA (respectively, LCC). Each of market size and of the two market presence variables are transformed into binary variables based on whether they realized above or below their respective median. This leads to a total of 8 market types, hence $J_1 = 16$ moment inequalities and $J_2 = 16$ moment equalities. The unobserved payoff shifters $u_{\ell,i}$ are assumed to be i.i.d. across i and to have a bivariate normal distribution with $E(u_{\ell,i}) = 0$, $Var(u_{\ell,i}) = 1$, and $Corr(u_{LCC,i}, u_{OA,i}) = r$ for each i and $\ell \in \{LCC, OA\}$, where the correlation r is to be estimated. Following Kline and Tamer (2016), we assume that the strategic interaction parameters δ_{LCC} and δ_{OA} are negative, that $r \geqslant 0$, and that the researcher imposes these sign restrictions. To ensure that Assumption E.4 is satisfied,[27] we furthermore assume that $r \leqslant 0.85$ and use this value as its upper bound in the definition

[25] For these same examples we verify the differentiability requirement in Assumption A.2 on $g_j(\theta)$.

[26] The data, which pertains to the second quarter of the year 2010, is downloaded from http://qeconomics.org/ojs/index.php/qe/article/downloadSuppFile/371/1173.

[27] This assumption, common in the literature on projection inference, requires that $D_{P,j}(\theta)$ are Lipschitz in θ and have bounded norm. But $\partial(\{E_P[m_j(X,\cdot)]/\sigma_{P,j}(\cdot)\})/\partial r$ includes a denominator equal to $(1-r^2)^2$. As $r \to 1$, this leads to a violation of the assumption and to numerical instability.

of the parameter space.

The results of the analysis are reported in Table 1, which displays 95% nominal confidence intervals (our CI_n as defined in equations (2.5)-(2.6)) for each parameter. The output of the E-A-M algorithm is displayed in the accordingly labeled column. The next column shows a robustness check, namely the output of MATLAB's `fmincon` function, henceforth labelled "direct search," that was started at each of a widely spaced set of feasible points that were previously discovered by the E-A-M algorithm. We emphasize that this is a robustness or accuracy check, not a horse race: Direct search mechanically improves on E-A-M because it starts (among other points) at the point reported by E-A-M as optimal feasible. Using the standard `MultiStart` function in MATLAB instead of the points discovered by E-A-M produces unreliable and extremely slow results. In 10 out of 18 optimization problems that we solved, the E-A-M algorithm's solution came within its set tolerance (0.005) from the direct search solution. The other optimization problems were solved by E-A-M with a minimal error of less than 5%.

Table 1 also reports computational time of the E-A-M algorithm, of the subsequent direct search, and the total time used to compute the confidence intervals. The direct search greatly increases computation time with small or negligible benefit. Also, computational time varied substantially across components. We suspect this might be due to the shape of the level sets of $\max_{j=1,\ldots,J} \sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)$: By manually searching around the optimal values of the program, we verified that the level sets in specific directions can be extremely thin, rendering search more challenging.

Comparing our findings with those in Kline and Tamer (2016), we see that the results qualitatively agree. The confidence intervals for the interaction effects (δ_{LCC} and δ_{OA}) and for the effect of market size on payoffs (ϑ^{size}_{LCC} and ϑ^{size}_{OA}) are similar to each other across the two types of firms. The payoffs of LCC firms seem to be impacted more than those of OA firms by market presence. On the other hand, monopoly payoffs for LCC firms seem to be smaller than for OA firms.[28] The confidence interval on the correlation coefficient is quite large and includes our upper bound of 0.85.[29]

For most components, our confidence intervals are narrower than the corresponding 95% credible sets reported in Kline and Tamer (2016).[30] However, the intervals are not comparable for at least two reasons: We impose a stricter upper bound on r and we aim to cover the projections of the true parameter value as opposed to the identified set.

Overall, our results suggest that in a reasonably sized, empirically interesting problem, calibrated projection yields informative confidence intervals. Furthermore, the E-A-M algo-

[28] Monopoly payoffs are those associated with a market with below-median size and below-median market presence (i.e., the constant terms).

[29] Being on the boundary of the parameter space is not a problem for calibrated projection; indeed, it is accounted for in the calibration of $\hat{c}_n$ in equations (2.11)-(2.13).

[30] For the interaction parameters δ, Kline and Tamer's upper confidence points are lower than ours; for the correlation coefficient r, their lower confidence point is higher than ours.

rithm appears to accurately and quickly approximate solutions to complex smooth nonlinear optimization problems.

5 Conclusion

This paper proposes a 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 θ 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 furthermore attractive thanks to an application of the response surface method for global optimization; this is a novel contribution of independent interest. Indeed, one key result is a convergence rate for this algorithm when applied to constrained optimization problems in which the objective function is easy to evaluate but the constraints are "black box" functions. The result is applicable to any instance when the researcher wants to compute confidence intervals for optimal values of constrained optimization problems. Our empirical application and Monte Carlo analysis show that, in the DGPs that we considered, calibrated projection is fast and accurate, and also that the E-A-M algorithm can greatly improve computation of other confidence intervals.

References

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.

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. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78(2), 735–753.

Bugni, F. A., I. A. Canay, and X. Shi (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.

Canay, I. (2010): "EL inference for partially identified models: large deviations optimality and bootstrap validity," *Journal of Econometrics*, 156(2), 408–425.

Chen, X., T. M. Christensen, and E. Tamer (2018): "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86(6), 1965–2018.

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. J., and E. Morales (2018): "What do Exporters Know?," *The Quarterly Journal of Economics*, 133(4), 1753–1801.

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

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

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): "Confidence Intervals for Projections of Partially Identified Parameters," CeMMAP Working Paper CWP 49/17, available at https://www.cemmap.ac.uk/publication/id/10139.

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.

Kline, B., and E. Tamer (2016): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, 7(2), 329–366.

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.

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, 315–334.

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.

Sutton, R. S., and A. G. Barto (1998): *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA.

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 2.3. Below, let $(\Omega, \mathcal{F})$ be a measurable space and ω a generic element of Ω. Let $L \in \mathbb{N}$ and let $(\theta^{(1)}, ..., \theta^{(L)})$ be a measurable map on $(\Omega, \mathcal{F})$ whose law is specified below. The value of the function c in (2.14) is unknown ex ante. Once the evaluation points $\theta^{(\ell)}, \ell = 1, ..., L$ realize, the corresponding values of c, i.e. $\Upsilon^{(\ell)} \equiv c(\theta^{(\ell)}), \ell = 1, ..., L$, are known. We may therefore define the information set

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

Let $\mathcal{C}_L \equiv \{\theta^{(\ell)} : \ell \in \{1, \cdots, L\}, g_j(\theta^{(\ell)}) \leqslant c(\theta^{(\ell)}), j = 1, \cdots, J\}$ be the set of feasible evaluation points. Then $\text{argmax}_{\theta \in \mathcal{C}_L} p'\theta$ is measurable with respect to $\mathcal{F}_L$ and we take a measurable selection $\theta^{*,L}$ from it.

Our algorithm iteratively determines evaluation points based on the *expected improvement* criterion (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[31]

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

where $\beta = (\beta_1, ..., \beta_d) \in \mathbb{R}^d$ 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 β_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 $k = 1, ..., d$. We let $\bar{\beta} = (\bar{\beta}_1, ..., \bar{\beta}_d)' \in \mathbb{R}^d$. Specific suggestions on the forms of K_β are given in Appendix D.2.

For a given (μ, ς, β), 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):

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

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

Given this, the expected improvement function can be written as

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

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

[31] 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.

in Section 2.3).

ALGORITHM A.1: *Let $k \in \mathbb{N}$.*
Step 1: Initial evaluation points $\theta^{(1)}, ..., \theta^{(k)}$ are drawn uniformly over Θ 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 ϵ, draw $\theta^{(L+1)}$ uniformly at random from Θ.

Below, we use $\mathbb{Q}$ to denote the law of $(\theta^{(1)}, ..., \theta^{(L)})$ 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 $\mathbb{Q}$. We let

$$\mathcal{C} \equiv \{\theta \in \Theta : \bar{g}(\theta) - c(\theta) \leqslant 0\}. \tag{A.6}$$

We require that the kernel used to define the correlation functional for the Gaussian process in (2.17) satisfies some basic regularity conditions. For this, let $\hat{K}_\beta = \int e^{-2\pi i x'\xi} K_\beta(x)dx$ denote the Fourier transform of K_β. Note also that, for real valued functions f, g, $f(y) = \Theta(g(y))$ means $f(y) = O(g(y))$ as $y \to \infty$ and $\liminf_{y\to\infty} f(y)/g(y) > 0$.

ASSUMPTION A.1 (Kernel Function): *(i) K_β is continuous and integrable; (ii) $\hat{K}_\beta = \hat{k}_\beta(\|x\|)$ for some nonincreasing function $\hat{k}_\beta : \mathbb{R}_+ \to \mathbb{R}_+$; (iii) As $x \to \infty$ either $\hat{K}_\beta(x) = \Theta(\|x\|^{-2\nu-d})$ for some $\nu > 0$ or $\hat{K}_\beta(x) = O(\|x\|^{-2\nu-d})$ for all $\nu > 0$; (iv) K_β is k-times continuously differentiable for $k = \lfloor 2\nu \rfloor$, and at the origin K has k-th order Taylor approximation P_k satisfying $|K(x) - P_k(x)| = O(\|x\|^{2\nu}(-\ln\|x\|)^{2\chi})$ as $x \to 0$, for some $\chi > 0$.*

Assumption A.1 is essentially the same as Assumptions 1-4 in Bull (2011). When a kernel satisfies the second condition of Assumption A.1 (iii), i.e. $\hat{K}_\beta(x) = O(\|x\|^{-2\nu-d}), \forall \nu > 0$, we say $\nu = \infty$. Assumption A.1 is satisfied by popular kernels such as the Matérn kernel (with $0 < \nu < \infty$ and $\chi = 1/2$) and the Gaussian kernel ($\nu = \infty$ and $\chi = 0$). These kernels are discussed in Appendix D.2.

Finally, we require that the functions g_j are differentiable with continuous Lipschitz gradient,[32] that the function c is smooth, and we impose on the constraint set $\mathcal{C}$ (which is a confidence set in our application) a degeneracy condition inspired by Chernozhukov, Hong, and Tamer (2007, Condition C.3).[33] 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 (2.17). The norm on this space is $\|\cdot\|_{\mathcal{H}_\beta}$; see Online Appendix D.2 for details.

ASSUMPTION A.2 (Continuity and Smoothness): *(i) For each $j = 1, \ldots, J$, the function $g_j(\theta)$ is differentiable in θ with Lipschitz continuous gradient. (ii) The function $c : \Theta \mapsto \mathbb{R}$ satisfies $\|c\|_{\mathcal{H}_{\bar{\beta}}} \leqslant R$ for some $R > 0$, where $\bar{\beta} = (\bar{\beta}_1, \cdots, \bar{\beta}_d)'$.*

ASSUMPTION A.3 (Degeneracy): *There exist constants (C_1, M, τ_1) such that for all $\varpi \in [0, \tau_1]$,*

$$\max_j g_j(\theta) - c(\theta) \leqslant -C_1\varpi, \text{ for all } \theta \in \mathcal{C}^{-\varpi},$$
$$d_H(\mathcal{C}^{-\varpi}, \mathcal{C}) \leqslant M\varpi,$$

[32] This requirement holds in the canonical partial identification examples discussed in Online Appendix F, using the same arguments as in Online Appendix F.1, provided $\hat{\sigma}_{n,j}(\theta) > 0$.

[33] Chernozhukov, Hong, and Tamer (2007) impose the degeneracy condition on the population identified set.

where $\mathcal{C}^{-\varpi} \equiv \{\theta \in \mathcal{C} : d(\theta, \Theta \backslash \mathcal{C}) \geqslant \varpi\}$.

Assumptions A.2-A.3 jointly imply a linear minorant property on $\max_j (g_j(\theta) - c(\theta))_+$:

$$\exists C_2 > 0, \tau_2 > 0 : \quad \max_j (g_j(\theta) - c(\theta))_+ \geqslant C_2 \min\{d(\theta, \mathcal{C}), \tau_2\}. \tag{A.7}$$

To see this, define $f_j(\theta) \equiv g_j(\theta) - c(\theta)$, so that the l.h.s. of the above inequality is $\max_j f_j(\theta)$. By Assumptions A.2-A.3 and compactness of Θ, $f_j(\cdot)$ is differentiable with Lipschitz continuous gradient. Let $\tilde{D}_j(\cdot)$ denote its gradient and let $\tilde{M}$ denote the corresponding Lipschitz constant. Let $\varepsilon = C_1/(M\tilde{M}J)$, where (C_1, M) are from Assumption A.3. We will show that, for constants (C_2, τ_2) to be determined, (i) $d(\theta, \mathcal{C}) \leqslant \varepsilon \Rightarrow \max_j f_j(\theta) \geqslant C_2 d(\theta, \mathcal{C})$ and (ii) $d(\theta, \mathcal{C}) \geqslant \varepsilon \Rightarrow \max_j f_j(\theta) \geqslant C_2 \tau_2$, so that the minimum between these bounds applies to any θ.

To see (i), write $\theta = \theta^* + r$, where θ^* is the projection of θ onto $\mathcal{C}$. Fix a sequence $\varpi_m \to 0$. By assumption A.3, there exists a corresponding sequence $\theta^*_m \to \theta^*$ with (for m large enough) $\|\theta^*_m - \theta^*\| \leqslant M\varpi_m$ but also $\max_j f_j(\theta^*_m) \leqslant -C_1\varpi_m$. Let $t_m \equiv (\theta^*_m - \theta^*)/\|\theta^*_m - \theta^*\|$ be the sequence of corresponding directions. Then for any accumulation point t of t_m and any active constraint j (i.e., $f_j(\theta^*) = 0$; such j necessarily exists due to continuity of $f_j(\cdot)$), one has $\tilde{D}_j(\theta^*)t \leqslant -C_1/M$. We note for future reference that this finding implies $\|\tilde{D}_j(\theta^*)\| \geqslant C_1/M$. It also implies that the Mangasarian-Fromowitz constraint qualification holds at θ^*, hence r (being in the normal cone of $\mathcal{C}$ at θ^*) is in the positive span of the active constraints' gradients. Thus j can be chosen such that $f_j(\theta^*) = 0$ and $\tilde{D}_j(\theta^*)r \geqslant \|\tilde{D}_j(\theta^*)\|\|r\|/J$. For any such j, write

$$\begin{aligned}
f_j(\theta) &= f_j(\theta^*) + \int_0^1 \frac{df_j(\theta^* + kr)}{dk} dk \\
&= 0 + \int_0^1 \tilde{D}_j(\theta^* + kr) r dk \\
&= \int_0^1 \left(\tilde{D}_j(\theta^*)r + \left(\tilde{D}_j(\theta^* + kr) - \tilde{D}_j(\theta^*)\right)r \right) dk \\
&\geqslant \|\tilde{D}_j(\theta^*)\|\|r\|/J + \int_0^1 (-\tilde{M}k\|r\|)\|r\| dk \\
&\geqslant \tfrac{C_1}{MJ}\|r\| - \tilde{M}\|r\|^2/2 \\
&\geqslant \tfrac{C_1}{2MJ}\|r\|.
\end{aligned}$$

In the inequality steps, we successively substituted bounds stated before the display, evaluated the integral in k, and (in the last step) used $\|r\| \leqslant \varepsilon$. This establishes (i), where $C_2 = C_1/(2MJ)$. Next, by continuity of $\max_j f_j(\cdot)$ and compactness of the constraint set, $\tau \equiv \min_\theta\{\max_j f_j(\theta) : d(\theta, \mathcal{C}) \geqslant \varepsilon\}$ is well-defined and strictly positive. This establishes (ii) with $\tau_2 = \tau/C_2$.

A.1 Proof of Theorem 3.2

For each $L \in \mathbb{N}$, let

$$r_L \equiv \left(\frac{L}{\ln L}\right)^{-\nu/d} (\ln L)^\chi. \tag{A.8}$$

Proof of Theorem 3.2. First, note that

$$\|p'\theta^* - p'\theta^{*,L}\|_{L^1_{\mathbb{Q}}} = E_{\mathbb{Q}}\big[\,|p'\theta^* - p'\theta^{*,L}|\,\big] = E_{\mathbb{Q}}\big[p'\theta^* - p'\theta^{*,L}\big], \tag{A.9}$$

where the last equality follows form $p'\theta^* - p'\theta^{*,L+1} \geqslant 0, \mathbb{Q}-a.s.$ Hence, it suffices to show

$$E_{\mathbb{Q}}\big[p'\theta^* - p'\theta^{*,L}\big] = O\Big(\Big(\frac{L}{\ln L}\Big)^{-\nu/d}(\ln L)^{\delta}\Big). \tag{A.10}$$

Let $(\Omega, \mathcal{F})$ be a measurable space. 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 Θ. 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:

$$h_L \equiv \sup_{\theta \in \Theta} \min_{\ell = 1, \cdots L} \|\theta - \theta^{(\ell)}\|. \tag{A.11}$$

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

$$D_L \equiv A_L \cap B_L \cap C_L. \tag{A.12}$$

For each $\omega \in D_L$, let

$$\ell(\omega, L) \equiv \inf\{\tilde{\ell} \in \mathbb{N} : L \leqslant \tilde{\ell} \leqslant 2L, \theta^{(\tilde{\ell})} \in \arg\max_{\theta \in \Theta} \mathbb{EI}_{\tilde{\ell}-1}(\theta)\}. \tag{A.13}$$

This is a (random) index that is associated with the first maximizer of the expected improvement between L and $2L$.

Let $\varepsilon_L = (L/\ln L)^{-\nu/d}(\ln L)^{\delta}$ for $\delta \geqslant 1 + \chi$ and note that ε_L is a positive sequence such that $\varepsilon_L \to 0$ and $r_L = o(\varepsilon_L)$. We further define the following events:

$$E_{1L} \equiv \{\omega \in \Omega : 0 < \bar{g}(\theta^{(\ell(\omega,L))}) - c(\theta^{(\ell(\omega,L))}) \leqslant \varepsilon_{\ell(\omega,L)}\} \tag{A.14}$$

$$E_{2L} \equiv \{\omega \in \Omega : -\varepsilon_{\ell(\omega,L)} \leqslant \bar{g}(\theta^{(\ell(\omega,L))}) - c(\theta^{(\ell(\omega,L))}) < 0\} \tag{A.15}$$

$$E_{3L} \equiv \{\omega \in \Omega : |\bar{g}(\theta^{(\ell(\omega,L))}) - c(\theta^{(\ell(\omega,L))})| > \varepsilon_{\ell(\omega,L)}\}. \tag{A.16}$$

Note that D_L can be partitioned into $D_L \cap E_{1L}$, $D_L \cap E_{2L}$, and $D_L \cap E_{3L}$. By Lemmas A.2, A.3, and A.4, there exists a constant $M > 0$ such that, respectively,

$$\sup_{\omega \in D_L \cap E_{1L}} |p'\theta^* - p'\theta^{*,\ell(\omega,L)}|/\varepsilon_{\ell(\omega,L)} \leqslant M \tag{A.17}$$

$$\sup_{\omega \in D_L \cap E_{2L}} |p'\theta^* - p'\theta^{*,\ell(\omega,L)}|/\varepsilon_{\ell(\omega,L)} \leqslant M \tag{A.18}$$

$$\sup_{\omega \in D_L \cap E_{3L}} |p'\theta^* - p'\theta^{*,\ell(\omega,L)}|/\exp(-M\eta_{\ell(\omega,L)}) \leqslant M, \tag{A.19}$$

where $\eta_L \equiv \varepsilon_L / r_L$. Note that

$$\eta_L = \varepsilon_L / r_L = (\ln L)^{\delta - \chi}. \tag{A.20}$$

Hence, by taking M sufficiently large so that $M > \nu/d$,

$$\exp(-M\eta_L) = \exp\left(-M(\ln L)^{\delta-\chi}\right) \leqslant \exp\left(-M\ln L\right) = L^{-M} = O(L^{-\nu/d}) = O(\varepsilon_L), \tag{A.21}$$

where the inequality follows from $M(\ln L)^{\delta-\chi} \geqslant M\ln L$ by $\delta \geqslant 1+\chi$. By (A.17)-(A.21),

$$\sup_{\omega\in D_L} |p'\theta^* - p'\theta^{*,\ell(\omega,L)}|/\varepsilon_{\ell(\omega,L)} \leqslant M, \tag{A.22}$$

for some constant $M > 0$ for all L sufficiently large. Since $L \leqslant \ell(\omega, L) \leqslant 2L$, $p'\theta^{*,L}$ is non-decreasing in L, and ε_L is non-increasing in L, we have

$$p'\theta^* - p'\theta^{*,2L} \leqslant M(L/\ln L)^{-\nu/d}(\ln L)^{\delta} \leqslant M(2L/\ln 2L)^{-\nu/d}(\ln 2L)^{\delta} \tag{A.23}$$

where the last equality follows from $L^{-\nu/d} = 2^{\nu/d}(2L)^{-\nu/d}$ and $\ln L \leqslant \ln 2L$.

Now consider the case $\omega \notin D_L$. By (A.12),

$$\mathbb{Q}(D_L^c) \leqslant \mathbb{Q}(A_L^c) + \mathbb{Q}(B_L^c) + \mathbb{Q}(C_L^c). \tag{A.24}$$

Let Z_ℓ 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),

$$\mathbb{Q}(A_L^c) = \mathbb{Q}\Big(\sum_{\ell=k+1}^{L} Z_\ell < \lfloor \eta L \rfloor\Big) \leqslant \exp(-(L-k+1)\epsilon(\epsilon-\eta)^2/2). \tag{A.25}$$

Further, by the definition of B_L,

$$\mathbb{Q}(B_L^c) = \epsilon^L, \tag{A.26}$$

and finally by taking $\bar{M}$ large upon defining the event C_L and applying Lemma 12 in Bull (2011), one has

$$\mathbb{Q}(C_L^c) = O(L^{-\gamma}), \tag{A.27}$$

for any $\gamma > 0$. Combining (A.24)-(A.27), for any $\gamma > 0$,

$$\mathbb{Q}(D_L^c) = O(L^{-\gamma}). \tag{A.28}$$

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

$$\begin{aligned} E_{\mathbb{Q}}\left[p'\theta^* - p'\theta^{*,2L}\right] &= \int_{D_L} p'\theta^* - p'\theta^{*,2L} d\mathbb{Q} + \int_{D_L^c} p'\theta^* - p'\theta^{*,2L} d\mathbb{Q} \\ &= O((2L/\ln 2L)^{-\nu/d}(\ln 2L)^{\delta}) + O(2L^{-\gamma}), \end{aligned} \tag{A.29}$$

where the second equality follows from (A.23) and (A.28). Since $\gamma > 0$ can be made aribitrarily large, one may let the second term on the right hand side of (A.29) converge to 0 faster than the first term.

Therefore

$$E_{\mathbb{Q}}\big[p'\theta^* - p'\theta^{*,2L}\big] = O((2L/\ln 2L)^{-\nu/d}(\ln 2L)^{\delta}), \tag{A.30}$$

which establishes the claim of the theorem for $0 < \nu < \infty$. When the second condition of Assumption A.1 (iii) holds (i.e., $\nu = \infty$), the argument above holds for any $0 < \nu < \infty$. □

A.2 Auxiliary Lemmas for the Proof of Theorem 3.2

Let D_L be defined as in (A.12). The following lemma shows that on $D_L \cap E_{1L}$, $p'\theta^*$ and $p'\theta^{(\ell(\omega,L))}$ are close to each other, where we recall that $\theta^{(\ell(\omega,L))}$ is the expected improvement maximizer (but does not belong to $\mathcal{C}$ for $\omega \in E_{1L}$).

LEMMA A.1: *Suppose Assumptions A.1, A.2, and A.3 hold. Let ε_L be a positive sequence such that $\varepsilon_L \to 0$ and $r_L = o(\varepsilon_L)$. Then, there exists a constant $M > 0$ such that $\sup_{\omega \in D_L \cap E_{1L}} |p'\theta^* - p'\theta^{(\ell(\omega,L))}|/\varepsilon_{\ell(\omega,L)} \leqslant M$ for all L sufficiently large.*

Proof. We show the result by contradiction. Let $\{\omega_L\} \subset \Omega$ be a sequence such that $\omega_L \in D_L \cap E_{1L}$ for all L. First, assume that, for any $M > 0$, there is a subsequence such that $|p'\theta^* - p'\theta^{(\ell(\omega_L,L))}| > M\varepsilon_{\ell(\omega_L,L)}$ for all L. This occurs if it contains a further subsequence along which, for all L, (i) $p'\theta^{(\ell(\omega_L,L))} - p'\theta^* > M\varepsilon_{\ell(\omega_L,L)}$ or (ii) $p'\theta^* - p'\theta^{(\ell(\omega_L,L))} > M\varepsilon_{\ell(\omega_L,L)}$.

Case (i): $p'\theta^{(\ell(\omega_L,L))} - p'\theta^* > M\varepsilon_{\ell(\omega_L,L)}$ for all L for some subsequence.

To simplify notation, we select a further subsequence $\{a_L\}$ of $\{L\}$ such that for any $a_L < a_{L'}$, $\ell(\omega_{a_L}, a_L) < \ell(\omega_{a_{L'}}, a_{L'})$. This then induces a sequence $\{\theta^{(\ell)}\}$ of expected improvement maximizers such that $p'\theta^{(\ell)} - p'\theta^* > M\varepsilon_\ell$ for all ℓ, where each ℓ equals $\ell(\omega_{a_L}, a_L)$ for some $a_L \in \mathbb{N}$. In what follows, we therefore omit the arguments of ℓ, but this sequence's dependence on (w_{a_L}, a_L) should be implicitly understood.

Recall that $\mathcal{C}$ defined in equation (A.6) is a compact set and that $\Pi_{\mathcal{C}}\theta^{(\ell)} = \arg\min_{\theta\in\mathcal{C}} \|\theta^{(\ell)} - \theta\|$ denotes the projection of $\theta^{(\ell)}$ on $\mathcal{C}$. Then

$$\begin{aligned} p'\theta^{(\ell)} - p'\theta^* &= (p'\theta^{(\ell)} - p'\Pi_{\mathcal{C}}\theta^{(\ell)}) + (p'\Pi_{\mathcal{C}}\theta^{(\ell)} - p'\theta^*) \\ &\leqslant \|p\|\|\theta^{(\ell)} - \Pi_{\mathcal{C}}\theta^{(\ell)}\| + (p'\Pi_{\mathcal{C}}\theta^{(\ell)} - p'\theta^*) \leqslant d(\theta^{(\ell)}, \mathcal{C}), \end{aligned} \tag{A.31}$$

where the first inequality follows from the Cauchy-Schwarz inequality, and the second inequality follows from $p'\Pi_{\mathcal{C}}\theta^{(\ell)} - p'\theta^* \leqslant 0$ due to $\Pi_{\mathcal{C}}\theta^{(\ell)} \in \mathcal{C}$. Therefore, by equation (A.7), for any $M > 0$

$$\bar{g}(\theta^{(\ell)}) - c(\theta^{(\ell)})_+ \geqslant C_2 d(\theta^{(\ell)}, \mathcal{C}) > C_2 M\varepsilon_\ell, \tag{A.32}$$

for all ℓ sufficiently large, where the last inequality follows from $p'\theta^{(\ell)} - p'\theta^* > M\varepsilon_\ell$. Take M such that $C_2 M > 1$. Then $(\bar{g}(\theta^{(\ell)}) - c(\theta^{(\ell)}))/\varepsilon_\ell > C_2 M > 1$ for all ℓ sufficiently large, contradicting $\omega_L \in E_{1L}$.

Case (ii): Similar to Case (i), we work with a further subsequence along which $p'\theta^* - p'\theta^{(\ell)} > M\varepsilon_\ell$ for all ℓ. Recall that along this subsequence, $\theta^{(\ell)} \notin \mathcal{C}$ because $0 < \bar{g}(\theta^{(\ell)}) - c(\theta^{(\ell)}) \leqslant \varepsilon_\ell$. We will construct $\tilde{\theta}^{(\ell)} \in \mathcal{C}^{-\varepsilon_\ell}$ s.t. $\mathbb{E}\mathbb{I}_{\ell-1}(\tilde{\theta}^{(\ell)}) > \mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)})$, contradicting the definition of $\theta^{(\ell)}$.

By Assumption A.3,

$$d_H(\mathcal{C}^{-\varepsilon_\ell}, \mathcal{C}) \leqslant M\varepsilon_\ell, \tag{A.33}$$

for all ℓ such that $\varepsilon_\ell \leqslant \tau_1$. By the Cauchy-Schwarz inequality, for any $\tilde{\theta}$,

$$p'\theta^* - p'\tilde{\theta} \leqslant \|p\|\|\theta^* - \tilde{\theta}\|. \tag{A.34}$$

Therefore, minimizing both sides with respect to $\tilde{\theta} \in \mathcal{C}^{-\varepsilon_\ell}$ and noting that $\|p\| = 1$, we obtain

$$p'\theta^* - \sup_{\tilde{\theta}\in\mathcal{C}^{-\varepsilon_\ell}} p'\tilde{\theta} \leqslant \inf_{\tilde{\theta}\in\mathcal{C}^{-\varepsilon_\ell}} \|\theta^* - \tilde{\theta}\|. \tag{A.35}$$

Further, noting that $\theta^* \in \mathcal{C}$,

$$\inf_{\tilde{\theta}\in\mathcal{C}^{-\varepsilon_\ell}} \|\theta^* - \tilde{\theta}\| \leqslant \sup_{\theta\in\mathcal{C}} \inf_{\tilde{\theta}\in\mathcal{C}^{-\varepsilon_\ell}} \|\theta - \tilde{\theta}\| \leqslant d_H(\mathcal{C}^{-\varepsilon_\ell}, \mathcal{C}). \tag{A.36}$$

By (A.33)-(A.36),

$$p'\theta^* - \sup_{\theta\in\mathcal{C}^{-\varepsilon_\ell}} p'\theta \leqslant M\varepsilon_\ell, \tag{A.37}$$

for all ℓ sufficiently large. Therefore, for all ℓ sufficiently large, one has

$$p'\theta^* - \sup_{\theta\in\mathcal{C}^{-\varepsilon_\ell}} p'\theta < p'\theta^* - p'\theta^{(\ell)}, \tag{A.38}$$

implying existence of $\tilde{\theta}^{(\ell)} \in \mathcal{C}^{-\varepsilon_\ell}$ s.t.

$$p'\tilde{\theta}^{(\ell)} > p'\theta^{(\ell)}. \tag{A.39}$$

By Lemma A.6, for $t(\theta) \equiv (\bar{g}(\theta) - c(\theta))/s_\ell(\theta)$, one can write

$$\mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)}) \leqslant (p'\theta^{(\ell)} - p'\theta^{*,\ell-1})_+ \Big(1 - \Phi\Big(\frac{t(\theta^{(\ell)}) - R}{\varsigma}\Big)\Big) \tag{A.40}$$

$$\leqslant (p'\theta^{(\ell)} - p'\theta^{*,\ell-1})_+ (1 - \Phi(-R/\varsigma)), \tag{A.41}$$

where the last inequality uses $t(\theta^{(\ell)}) > 0$. Lemma A.6 also yields

$$\begin{aligned}\mathbb{E}\mathbb{I}_{\ell-1}(\tilde{\theta}^{(\ell)}) &\geqslant (p'\tilde{\theta}^{(\ell)} - p'\theta^{*,\ell-1})_+ \Big(1 - \Phi\Big(\frac{t(\tilde{\theta}^{(\ell)}) + R}{\varsigma}\Big)\Big) \\ &> (p'\theta^{(\ell)} - p'\theta^{*,\ell-1})_+ \Big(1 - \Phi\Big(\frac{t(\tilde{\theta}^{(\ell)}) + R}{\varsigma}\Big)\Big)\end{aligned} \tag{A.42}$$

for all ℓ sufficiently large, where the second inequality follows from (A.39). Next, by Assumption A.3,

$$t(\tilde{\theta}^{(\ell)}) = \frac{\bar{g}(\tilde{\theta}^{(\ell)}) - c(\tilde{\theta}^{(\ell)})}{s_\ell(\tilde{\theta}^{(\ell)})} \leqslant \frac{-C_1\varepsilon_\ell}{s_\ell(\tilde{\theta}^{(\ell)})} \tag{A.43}$$

for all ℓ sufficiently large. Note that $s_\ell(\tilde{\theta}^{(\ell)}) = O(r_\ell)$ by (A.62) and $r_\ell = o(\varepsilon_\ell)$ by assumption. Hence,

$t(\tilde{\theta}^{(\ell)}) \to -\infty$. This in turn implies

$$\mathbb{E}\mathbb{I}_{\ell-1}(\tilde{\theta}^{(\ell)}) > (p'\theta^{(\ell)} - p'\theta^{*,\ell-1})_+(1-\Phi(-R/\varsigma)) \tag{A.44}$$

for all ℓ sufficiently large. (A.41) and (A.44) jointly establish the desired contradiction. □

The next lemma shows that on $D_L \cap E_{1L}$, $p'\theta^*$ and $p'\theta^{*,(\ell(\omega,L))}$ are close to each other, where we recall that $\theta^{*,(\ell(\omega,L))}$ is the optimum value among the available feasible points (it belongs to $\mathcal{C}$).

LEMMA A.2: *Suppose Assumptions A.1, A.2, and A.3 hold. Let ε_L be a positive sequence such that $\varepsilon_L \to 0$ and $r_L = o(\varepsilon_L)$. Then, there exists a constant $M > 0$ such that $\sup_{\omega\in D_L\cap E_{1L}} |p'\theta^* - p'\theta^{*,\ell(\omega,L)}|/\varepsilon_{\ell(\omega,L)} \leqslant M$ for all L sufficiently large.*

Proof. We show below $p'\theta^* - p'\theta^{*,\ell(\omega,L)-1} = O(\varepsilon_{\ell(\omega,L)})$ uniformly over $D_L \cap E_{1L}$ for some decreasing sequence ε_ℓ satisfying the assumptions of the lemma. The claim then follows by re-labeling ε_ℓ.

Suppose by contradiction that, for any $M > 0$, there is a subsequence $\{\omega_{a_L}\} \subset \Omega$ along which $\omega_{a_L} \in D_{a_L}$ and $|p'\theta^* - p'\theta^{*,\ell(\omega_{a_L},a_L)-1}| > M\varepsilon_{\ell(\omega_{a_L},a_L)}$ for all L sufficiently large. To simplify notation, we select a subsequence $\{a_L\}$ of $\{L\}$ such that for any $a_L < a_{L'}$, $\ell(\omega_{a_L}, a_L) < \ell(\omega_{a_{L'}}, a_{L'})$. This then induces a sequence such that $|p'\theta^* - p'\theta^{*,\ell-1}| > M\varepsilon_\ell$ for all ℓ, where each ℓ equals $\ell(\omega_{a_L}, a_L)$ for some $a_L \in \mathbb{N}$. Similar to the proof of Lemma A.1, we omit the arguments of ℓ below and construct a sequence of points $\tilde{\theta}^{(\ell)} \in \mathcal{C}^{-\varepsilon_\ell}$ such that $\mathbb{E}\mathbb{I}_{\ell-1}(\tilde{\theta}^{(\ell)}) > \mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)})$.

Arguing as in (A.33)-(A.36), one may find a sequence of points $\tilde{\theta}^{(\ell)} \in \mathcal{C}^{-\varepsilon_\ell}$ such that

$$p'\theta^* - p'\tilde{\theta}^{(\ell)} \leqslant M_1\varepsilon_\ell, \tag{A.45}$$

for some $M_1 > 0$ and for all ℓ sufficiently large. Furthermore, by Lemma A.1,

$$|p'\theta^* - p'\theta^{(\ell)}| \leqslant M_2\varepsilon_\ell, \tag{A.46}$$

for some $M_2 > 0$ and for all ℓ sufficiently large. Arguing as in (A.41),

$$\begin{aligned}
\mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)}) &\leqslant (p'\theta^{(\ell)} - p'\theta^{*,\ell-1})_+\big(1-\Phi(-R/\varsigma)\big) \\
&= (p'\theta^* - p'\theta^{*,\ell-1} - (p'\theta^* - p'\theta^{(\ell)}))_+\big(1-\Phi(-R/\varsigma)\big) \\
&\leqslant (p'\theta^* - p'\theta^{*,\ell-1})\big(1-\Phi(-R/\varsigma)\big) + |p'\theta^* - p'\theta^{(\ell)}|,
\end{aligned} \tag{A.47}$$

where the last inequality follows from the triangle inequality, $p'\theta^* - p'\theta^{*,\ell-1} \geqslant 0$, and $1-\Phi(\frac{-R}{\varsigma}) \leqslant 1$. Similarly, by Lemma A.6,

$$\begin{aligned}
\mathbb{E}\mathbb{I}_{\ell-1}(\tilde{\theta}^{(\ell)}) &\geqslant (p'\tilde{\theta}^{(\ell)} - p'\theta^{*,\ell-1})_+\Big(1-\Phi\Big(\frac{t(\tilde{\theta}^{(\ell)})+R}{\varsigma}\Big)\Big) \\
&= (p'\theta^* - p'\theta^{*,\ell-1} - (p'\theta^* - p'\tilde{\theta}^{(\ell)}))_+\Big(1-\Phi\Big(\frac{t(\tilde{\theta}^{(\ell)})+R}{\varsigma}\Big)\Big) \\
&\geqslant (p'\theta^* - p'\theta^{*,\ell-1})\Big(1-\Phi\Big(\frac{t(\tilde{\theta}^{(\ell)})+R}{\varsigma}\Big)\Big) - (p'\theta^* - p'\tilde{\theta}^{(\ell)}),
\end{aligned} \tag{A.48}$$

where the last inequality holds for all ℓ sufficiently large because $p'\theta^* - p'\tilde{\theta}^{(\ell)} \in (0, M_2\varepsilon_\ell]$ and one can find a subsequence $p'\theta^* - p'\theta^{*,\ell-1} > M_2\varepsilon_\ell$ so that $p'\theta^* - p'\theta^{*,\ell-1} - (p'\theta^* - p'\tilde{\theta}^{(\ell)}) > 0$ for all ℓ

sufficiently large.

Subtracting (A.47) from (A.48) yields

$$\begin{aligned}
&\mathbb{E}\mathbb{I}_{\ell-1}(\tilde{\theta}^{(\ell)}) - \mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)}) \\
&\geqslant (p'\theta^* - p'\theta^{*,\ell-1})\Big(\Phi\Big(\frac{-R}{\varsigma}\Big) - \Phi\Big(\frac{t(\tilde{\theta}^{(\ell)}) + R}{\varsigma}\Big)\Big) - (p'\theta^* - p'\tilde{\theta}^{(\ell)}) - |p'\theta^* - p'\theta^{(\ell)}| \\
&\geqslant (p'\theta^* - p'\theta^{*,\ell-1})\Big(\Phi\Big(\frac{-R}{\varsigma}\Big) - \Phi\Big(\frac{t(\tilde{\theta}^{(\ell)}) + R}{\varsigma}\Big)\Big) - (M_1 + M_2)\varepsilon_\ell, \qquad \text{(A.49)}
\end{aligned}$$

where the last inequality follows from (A.45) and (A.46). Note that there is a constant $\zeta > 0$ s.t.

$$\Phi\Big(\frac{-R}{\varsigma}\Big) - \Phi\Big(\frac{t(\tilde{\theta}^{(\ell)}) + R}{\varsigma}\Big) > \zeta, \qquad \text{(A.50)}$$

due to $t(\tilde{\theta}^{(\ell)}) \to -\infty$ by (A.43), (A.62), and $r_\ell = o(\varepsilon_\ell)$. Therefore, for all ℓ sufficiently large,

$$\mathbb{E}\mathbb{I}_{\ell-1}(\tilde{\theta}^{(\ell)}) - \mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)}) > M\zeta\varepsilon_\ell - (M_1 + M_2)\varepsilon_\ell. \qquad \text{(A.51)}$$

One may take M large enough so that, for some positive constant γ, $M\zeta\varepsilon_\ell - (M_1 + M_2)\varepsilon_\ell > \gamma\varepsilon_\ell$ for all ℓ sufficiently large, which implies $\mathbb{E}\mathbb{I}_{\ell-1}(\tilde{\theta}^{(\ell)}) - \mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)}) > 0$ for all ℓ sufficiently large. However, this contradicts the assumption that $\theta^{(\ell)} \notin \mathcal{C}^{-\varepsilon_\ell}$ is the expected improvement maximizer. □

The next lemma shows that on $D_L \cap E_{2L}$, $p'\theta^*$ and $p'\theta^{*,(\ell(\omega,L))}$ are close to each other.

LEMMA A.3: *Suppose Assumptions A.1, A.2, and A.3 hold. Let $\{\varepsilon_L\}$ be a positive sequence such that $\varepsilon_L \to 0$ and $r_L = o(\varepsilon_L)$. Then, there exists a constant $M > 0$ such that $\sup_{\omega\in D_L\cap E_{2L}} |p'\theta^* - p'\theta^{*,\ell(\omega,L)}|/\varepsilon_{\ell(\omega,L)} \leqslant M$ for all L sufficiently large.*

Proof. Note that, for any $L \in \mathbb{N}$, $\omega \in D_L \cap E_{2L}$, and $\ell = \ell(\omega, L)$, $\theta^{(\ell)}$ satisfies $\bar{g}(\theta^{(\ell)}) - c(\theta^{(\ell)}) \leqslant 0$, hence $p'\theta^{*,\ell} \geqslant p'\theta^{(\ell)}$, which in turn implies

$$0 \leqslant p'\theta^* - p'\theta^{*,\ell} \leqslant p'\theta^* - p'\theta^{(\ell)}. \qquad \text{(A.52)}$$

Therefore, it suffices to show the existence of $M > 0$ that ensures $(p'\theta^* - p'\theta^{(\ell(\omega,L))})_+ \leqslant M\varepsilon_{\ell(\omega,L)}$ uniformly over $D_L \cap E_{2L}$ for all L. Suppose by contradiction that, for any $M > 0$, there is a subsequence $\{\omega_{a_L}\} \subset \Omega$ along which $\omega_{a_L} \in D_{a_L} \cap E_{2a_L}$ and $p'\theta^* - p'\theta^{(\ell(\omega_{a_L},a_L))} > M\varepsilon_{\ell(\omega_{a_L},a_L)}$ for all L sufficiently large. Again, we select a subsequence $\{a_L\}$ of $\{L\}$ such that for any $a_L < a_{L'}$, $\ell(\omega_{a_L}, a_L) < \ell(\omega_{a_{L'}}, a_{L'})$. This then induces a sequence $\{\theta^{(\ell)}\}$ of expected improvement maximizers such that $(p'\theta^* - p'\theta^{(\ell)})_+ > M\varepsilon_\ell$ for all ℓ, where each ℓ equals $\ell(\omega_{a_L}, a_L)$ for some $a_L \in \mathbb{N}$.

Similar to the proof of Lemma A.1, we omit the arguments of ℓ below and prove the claim by contradiction. Below, we assume that, for any $M > 0$, there is a further subsequence along which $p'\theta^* - p'\theta^{(\ell)} > M\varepsilon_\ell$ for all ℓ sufficiently large.

Now let $\varepsilon'_\ell = \tilde{C}\varepsilon_\ell$ with $\tilde{C} > 0$ specified below. By Assumption A.3, for all $\tilde{\theta} \in \mathcal{C}^{-\varepsilon'_\ell}$, it holds that

$$\bar{g}(\tilde{\theta}) - c(\tilde{\theta}) \leqslant -\tilde{C}C_1\varepsilon_\ell, \qquad \text{(A.53)}$$

for all ℓ sufficiently large. Noting that $-\varepsilon_\ell \leqslant \bar{g}(\theta^{(\ell)}) - c(\theta^{(\ell)})$ and taking $\tilde{C}$ such that $\tilde{C}C_1 > 1$, it

follows that $\theta^{(\ell)} \notin \mathcal{C}^{-\varepsilon'_\ell}$ for all ℓ sufficiently large.

Arguing as in (A.33)-(A.36), one may find a sequence of points $\tilde{\theta}^{(\ell)} \in \mathcal{C}^{-\varepsilon'_\ell}$ such that

$$p'\theta^* - p'\tilde{\theta}^{(\ell)} \leqslant M_1 \varepsilon'_\ell = M_1 \tilde{C} \varepsilon_\ell, \tag{A.54}$$

This and the assumption that one can find a subsequence such that $p'\theta^* - p'\theta^{(\ell)} > M_1 \tilde{C} \varepsilon_\ell$ for all ℓ imply

$$p'\theta^* - p'\tilde{\theta}^{(\ell)} < p'\theta^* - p'\theta^{(\ell)}, \tag{A.55}$$

for all ℓ sufficiently large. Now mimic the argument along (A.41)-(A.44) to deduce

$$\mathbb{E}\mathbb{I}_{\ell-1}(\tilde{\theta}^{(\ell)}) > \mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)}) \tag{A.56}$$

for all ℓ sufficiently large. However, this contradicts the assumption that $\theta^{(\ell)} \notin \mathcal{C}^{-\varepsilon'_\ell}$ is the expected improvement maximizer. □

The next lemma shows that on $D_L \cap E_{3L}$, $p'\theta^*$ and $p'\theta^{*,(\ell(\omega,L))}$ are close to each other.

LEMMA A.4: *Suppose Assumptions A.1, A.2, and A.3 hold. Let* $\varepsilon_L = (L/\ln L)^{-\nu/d}(\ln L)^\delta$ *for* $\delta \geqslant 1+\chi$. *Let* $\eta_L = \varepsilon_L / r_L = (\ln L)^{\delta-\chi}$. *Then there exists a constant* $M > 0$ *such that* $\sup_{\omega \in D_L \cap E_{3L}} |p'\theta^* - p'\theta^{*,\ell(\omega,L)}| / \exp(-M\eta_{\ell(\omega,L)}) \leqslant M$ *for all* L *sufficiently large.*

Proof. Let $\{\omega_L\} \subset \Omega$ be a sequence such that $\omega_L \in D_L$ for all L. Since $\omega_L \in B_L$, there is $\ell = \ell(\omega_L, L)$ such that $L \leqslant \ell \leqslant 2L$ and $\theta^{(\ell)}$ is chosen by maximizing the expected improvement. For later use, we note that, for any $\tilde{M} > 0$, it can be shown that $\exp(-\tilde{M}\eta_{L-1}) / \exp(-\tilde{M}\eta_L) \to 1$, which in turn implies that there exists a constant $C > 1$ such that

$$\exp(-\tilde{M}\eta_{L-1}) \leqslant C \exp(-\tilde{M}\eta_L), \tag{A.57}$$

for all L sufficiently large.

For $\theta \in \Theta$ and $L \in \mathbb{N}$, let $\mathbb{I}_L(\theta) \equiv (p'\theta - p'\theta^{*,L})_+ 1\{\bar{g}(\theta) \leqslant c(\theta)\}$. Recall that θ^* is an optimal

solution to (2.14). Then, for all L sufficiently large,

$$\begin{aligned}
p'\theta^* - p'\theta^{*,\ell-1} &\overset{(1)}{=} \mathbb{I}_{\ell-1}(\theta^*) \overset{(2)}{\leqslant} \mathbb{E}\mathbb{I}_{\ell-1}(\theta^*)\big(1-\Phi(R/\varsigma)\big)^{-1} \overset{(3)}{\leqslant} \mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell)})\big(1-\Phi(R/\varsigma)\big)^{-1} \\
&\overset{(4)}{\leqslant} \Big(\mathbb{I}_{\ell-1}(\theta^{(\ell)}) + M_1\exp(-\tilde{M}\eta_{\ell-1})\Big)\big(1-\Phi(R/\varsigma)\big)^{-1} \\
&\overset{(5)}{\leqslant} \Big(\mathbb{I}_{\ell-1}(\theta^{(\ell)}) + M_2\exp(-\tilde{M}\eta_{\ell})\Big)\big(1-\Phi(R/\varsigma)\big)^{-1} \\
&\overset{(6)}{\leqslant} \Big(\mathbb{I}_{\ell-1}(\theta^{*,\ell}) + M_2\exp(-\tilde{M}\eta_{\ell})\Big)\big(1-\Phi(R/\varsigma)\big)^{-1} \\
&\overset{(7)}{\leqslant} \Big(\mathbb{E}\mathbb{I}_{\ell-1}(\theta^{*,\ell}) + 2M_2\exp(-\tilde{M}\eta_{\ell})\Big)\big(1-\Phi(R/\varsigma)\big)^{-1} \\
&\overset{(8)}{\leqslant} \Big(\mathbb{E}\mathbb{I}_{\ell-1}(\theta^{(\ell-1)}) + 2M_2\exp(-\tilde{M}\eta_{\ell})\Big)\big(1-\Phi(R/\varsigma)\big)^{-1} \\
&\overset{(9)}{\leqslant} \Big(\mathbb{I}_{\ell-1}(\theta^{(\ell-1)}) + 3M_2\exp(-\tilde{M}\eta_{\ell})\Big)\big(1-\Phi(R/\varsigma)\big)^{-1} \\
&\overset{(10)}{\leqslant} 3M_2\exp(-\tilde{M}\eta_{\ell})\big(1-\Phi(R/\varsigma)\big)^{-1},
\end{aligned}$$

where (1) follows by construction, (2) follows from Lemma A.6 (ii), (3) follows from $\theta^{(\ell)}$ being the maximizer of the expected improvement, (4) follows from Lemma A.5, (5) follows from (A.57) with $M_2 = CM_1$, (6) follows from $\theta^{*,\ell} = \operatorname{argmax}_{\theta\in\mathcal{C}_\ell} p'\theta$, (7) follows from Lemma A.5, (8) follows from $\theta^{(\ell-1)}$ being the expected improvement maximizer, (9) follows from Lemma A.5, and (10) follows from $\mathbb{I}_{\ell-1}(\theta^{(\ell-1)}) = 0$ due to the definition of $\theta^{*,\ell-1}$. This establishes the claim. □

For evaluation points θ_L such that $|\bar{g}(\theta_L) - c(\theta_L)| > \varepsilon_L$, 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 θ.

LEMMA A.5: *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 let Assumptions A.1 and A.2-(ii) hold. For $\theta \in \Theta$ and $L \in \mathbb{N}$, let $\mathbb{I}_L(\theta) \equiv (p'\theta - p'\theta^{*,L})_+ 1\{\bar{g}(\theta) \leqslant c(\theta)\}$. Let $\{\varepsilon_L\}$ be a positive sequence such that $\varepsilon_L \to 0$ and $r_L = o(\varepsilon_L)$. Let $\eta_L \equiv \varepsilon_L / r_L$. Then, for any sequence $\{\theta_L\} \subset \Theta$ such that $|\bar{g}(\theta_L) - c(\theta_L)| > \varepsilon_L$,*

$$\mathbb{I}_L(\theta_L) - \gamma_L \leqslant \mathbb{E}\mathbb{I}_L(\theta_L) \leqslant \mathbb{I}_L(\theta_L) + \gamma_L, \tag{A.58}$$

where $\gamma_L = O(\exp(-M\eta_L))$.

Proof of Lemma A.5. If $s_L(\theta_L) = 0$, then the posterior variance of $c(\theta_L)$ is zero. Hence, $\mathbb{E}\mathbb{I}_L(\theta_L) = \mathbb{I}_L(\theta_L)$, and the claim of the lemma holds.

Suppose $s_L(\theta_L) > 0$. We first show the upper bound. Let $u \equiv (\bar{g}(\theta_L) - c_L(\theta_L))/s_L(\theta_L)$ and $t \equiv (\bar{g}(\theta_L) - c(\theta_L))/s_L(\theta_L)$. By Lemma 6 in Bull (2011), we have $|u - t| \leqslant R$. Starting from Lemma

A.6(i), we can write

$$\begin{aligned}\mathbb{E}\mathbb{I}_L(\theta_L) &\leqslant (p'\theta_L - p'\theta^{*,L})_+\Big(1-\Phi\Big(\frac{t-R}{\varsigma}\Big)\Big)\\ &= (p'\theta_L - p'\theta^{*,L})_+(1\{\bar{g}(\theta_L)\leqslant c(\theta_L)\}+1\{\bar{g}(\theta_L)>c(\theta_L)\})\Big(1-\Phi\Big(\frac{t-R}{\varsigma}\Big)\Big)\\ &\leqslant \mathbb{I}_L(\theta_L)+(p'\theta_L - p'\theta^{*,L})_+1\{\bar{g}(\theta_L)>c(\theta_L)\}\Big(1-\Phi\Big(\frac{t-R}{\varsigma}\Big)\Big),\end{aligned} \tag{A.59}$$

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

$$1\{\bar{g}(\theta_L)>c(\theta_L)\}\Big(1-\Phi\Big(\frac{t-R}{\varsigma}\Big)\Big)=1\{\bar{g}(\theta_L)>c(\theta_L)\}\Big(1-\Phi\Big(\frac{\bar{g}(\theta_L)-c(\theta_L)-s_L(\theta_L)R}{\varsigma s_L(\theta_L)}\Big)\Big). \tag{A.60}$$

To be clear about the hyperparameter value at which we evaluate s_L, we will write $s_L(\theta_L;\beta)$. By the hypothesis that $\|c\|_{\mathcal{H}_{\bar{\beta}}}\leqslant R$ and Lemma 4 in Bull (2011), we have

$$\|c\|_{\mathcal{H}_{\beta_L}}\leqslant R^2\prod_{k=1}^{d}(\overline{\beta}_k/\underline{\beta}_k)\equiv S. \tag{A.61}$$

Note that there are $\lfloor \eta L\rfloor$ uniformly sampled points, and K_β is associated with index $\nu\in(0,\infty)$. As shown in the proof of Theorem 5 in Bull (2011), this ensures that

$$\sup_{\beta\in\prod_{k=1}^{d}[\underline{\beta}_k,\overline{\beta}_k]} s_L(\theta_L;\beta)=O(h_L^{\nu}(\ln L)^{\chi})=O(r_L). \tag{A.62}$$

Below, we simply write this result $s_L(\theta_L)=O(r_L)$. This, together with $|\bar{g}(\theta_L)-c(\theta_L)|>\varepsilon_L$ and the fact that $1-\Phi(\cdot)$ is decreasing, yields

$$\begin{aligned}1\{\bar{g}(\theta_L)>c(\theta_L)\}\Big(1-\Phi\Big(\frac{\bar{g}(\theta_L)-c(\theta_L)-s_L(\theta_L)R}{\varsigma s_L(\theta_L)}\Big)\Big)&\leqslant 1-\Phi\Big(\frac{\varepsilon_L}{\varsigma s_L(\theta_L)}-\frac{R}{\varsigma}\Big)\\ &\leqslant 1-\Phi(M_1\eta_L-M_2),\end{aligned} \tag{A.63}$$

for some $M_1>0$ and where $M_2=R/\varsigma$. Note that, by the triangle inequality,

$$1-\Phi(M_1\eta_L-M_2)\leqslant 1-\Phi(M_1\eta_L)+|(1-\Phi(M_1\eta_L-M_2))-(1-\Phi(M_1\eta_L))|, \tag{A.64}$$

and

$$1-\Phi(M_1\eta_L)\leqslant\frac{1}{M_1\eta_L}\phi(M_1\eta_L)=O(\exp(-M\eta_L)), \tag{A.65}$$

for some $M>0$, where ϕ is the density of the standard normal distribution, and the inequality follows from $1-\Phi(x)\leqslant\phi(x)/x$. The second term on the right hand side of (A.64) can be bounded as

$$|(1-\Phi(M_1\eta_L-M_2))-(1-\Phi(M_1\eta_L))|\leqslant\phi(\tilde{\eta}_L)M_2=O(\exp(-M\eta_L)) \tag{A.66}$$

by the mean value theorem, where $\tilde{\eta}_L$ is a point between $M_1\eta_L$ and $M_1\eta_L-M_2$. The claim of the lemma then follows from (A.59), (A.63)-(A.66), and $(p'\theta_L-p'\theta_L^{*,L})$ being bounded because Θ is bounded.

Similarly, for the lower bound, we have

$$\begin{aligned}\mathbb{E}\mathbb{I}_L(\theta_L) &\geqslant (p'\theta_L - p'\theta_L^*)_+\Big(1-\Phi\Big(\frac{t+R}{\varsigma}\Big)\Big)\\ &\geqslant (p'\theta_L - p'\theta_L^*)_+1\{\bar{g}(\theta_L)\leqslant c(\theta_L)\}\Big(1-\Phi\Big(\frac{t+R}{\varsigma}\Big)\Big)\\ &\geqslant \mathbb{I}_L(\theta_L) - (p'\theta_L - p'\theta_L^*)_+1\{\bar{g}(\theta_L)\leqslant c(\theta_L)\}\Phi\Big(\frac{t+R}{\varsigma}\Big). \end{aligned} \tag{A.67}$$

Note that we may write

$$1\{\bar{g}(\theta_L)\leqslant c(\theta_L)\}\Phi\Big(\frac{t+R}{\varsigma}\Big) = 1\{\bar{g}(\theta_L)< c(\theta_L)\}\Phi\Big(\frac{\bar{g}(\theta_L)-c(\theta_L)+s_L(\theta_L)R}{\varsigma s_L(\theta_L)}\Big), \tag{A.68}$$

by $|\bar{g}(\theta_L)-c(\theta_L)|>\varepsilon_L$. Arguing as in (A.67) and noting that Φ is increasing, one has

$$\begin{aligned}1\{\bar{g}(\theta_L)< c(\theta_L)\}\Phi\Big(\frac{\bar{g}(\theta_L)-c(\theta_L)+s_L(\theta_L)R}{\varsigma s_L(\theta_L)}\Big) &\leqslant \Phi\Big(\frac{-\varepsilon_L}{\varsigma s_L(\theta_L)}+M_2\Big)\\ &\leqslant \Phi(-M_1\eta_L+M_2), \end{aligned} \tag{A.69}$$

for some $M_1>0$ and $M_2>0$. By the triangle inequality,

$$\Phi(-M_1\eta_L+M_2)\leqslant \Phi(-M_1\eta_L)+|\Phi(-M_1\eta_L+M_2)-\Phi(-M_1\eta_L)|, \tag{A.70}$$

where arguing as in (A.65),

$$\Phi(-M_1\eta_L) = 1-\Phi(M_1\eta_L) = O(\exp(-M\eta_L)). \tag{A.71}$$

The second term on the right hand side of (A.70) can be bounded as

$$\begin{aligned}&|\Phi(-M_1\eta_L+M_2)-\Phi(-M_1\eta_L)|\\ &\qquad = |(1-\Phi(M_1\eta_L-M_2))-(1-\Phi(M_1\eta_L))|\leqslant \phi(\tilde{\eta}_L)M_2 = O(\exp(-M\eta_L)), \end{aligned} \tag{A.72}$$

by the mean value theorem, where $\tilde{\eta}_L$ is a point between $M_1\eta_L$ and $M_1\eta_L-M_2$. The claim of the lemma then follows from (A.67)-(A.72), and $(p'\theta_L - p'\theta_L^{*,L})$ being bounded because Θ is bounded. □

LEMMA A.6: *Suppose $\Theta\subset\mathbb{R}^d$ is bounded and $p\in\mathbb{S}^{d-1}$ and let Assumptions A.1 and A.2-(ii) hold. Let $t(\theta)\equiv(\bar{g}(\theta)-c(\theta))/s_L(\theta)$. For $\theta\in\Theta$ and $L\in\mathbb{N}$, let $\mathbb{I}_L(\theta)\equiv(p'\theta-p'\theta^{*,L})_+1\{\bar{g}(\theta)\leqslant c(\theta)\}$. Then, (i) for any $L\in\mathbb{N}$ and $\theta\in\Theta$,*

$$(p'\theta-p'\theta^{*,L})_+\Big(1-\Phi\Big(\frac{t(\theta)+R}{\varsigma}\Big)\Big)\leqslant \mathbb{E}\mathbb{I}_L(\theta)\leqslant (p'\theta-p'\theta^{*,L})_+\Big(1-\Phi\Big(\frac{t(\theta)-R}{\varsigma}\Big)\Big). \tag{A.73}$$

Further, (ii) for any $L\in\mathbb{N}$ and $\theta\in\Theta$ such that $s_L(\theta)>0$,

$$\mathbb{I}_L(\theta)\leqslant \mathbb{E}\mathbb{I}_L(\theta)\Big(1-\Phi\Big(\frac{R}{\varsigma}\Big)\Big)^{-1}. \tag{A.74}$$

Proof. (i) Let $u(\theta)\equiv(\bar{g}(\theta)-c_L(\theta))/s_L(\theta)$ and $t(\theta)\equiv(\bar{g}(\theta)-c(\theta))/s_L(\theta)$. By Lemma 6 in Bull (2011),

we have $|u(\theta) - t(\theta)| \leqslant R$. Since $1 - \Phi(\cdot)$ is decreasing, we have

$$\mathbb{E}\mathbb{I}_L(\theta) = (p'\theta - p'\theta^{*,L})_+\Big(1 - \Phi\Big(\frac{u(\theta)}{\varsigma}\Big)\Big) \leqslant (p'\theta - p'\theta^{*,L})_+\Big(1 - \Phi\Big(\frac{t(\theta) - R}{\varsigma}\Big)\Big). \tag{A.75}$$

Similarly,

$$\mathbb{E}\mathbb{I}_L(\theta) = (p'\theta - p'\theta^{*,L})_+\Big(1 - \Phi\Big(\frac{u(\theta)}{\varsigma}\Big)\Big) \geqslant (p'\theta - p'\theta^{*,L})_+\Big(1 - \Phi\Big(\frac{t(\theta) + R}{\varsigma}\Big)\Big). \tag{A.76}$$

(ii) For the lower bound in (A.74), we have

$$\begin{aligned}
\mathbb{E}\mathbb{I}_L(\theta) &\geqslant (p'\theta - p'\theta^{*,L})_+\Big(1 - \Phi\Big(\frac{t(\theta) + R}{\varsigma}\Big)\Big) \\
&\geqslant (p'\theta - p'\theta^{*,L})_+ 1\{\bar{g}(\theta) \leqslant c(\theta)\}\Big(1 - \Phi\Big(\frac{t(\theta) + R}{\varsigma}\Big)\Big) \\
&\geqslant \mathbb{I}_L(\theta)\big(1 - \Phi(R/\varsigma)\big),
\end{aligned} \tag{A.77}$$

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

B Applying the E-A-M Algorithm to Profiling

We describe below how to use the E-A-M procedure to compute BCS-profiling based confidence intervals. Let $\mathcal{T} \subset \mathbb{R}$ denote the parameter space for $\tau = p'\theta$. The (one-dimensional) profiling confidence region is

$$\Big\{\tau \in \mathcal{T} : \inf_{\theta : p'\theta = \tau} T_n(\theta) \leqslant c_n^{MR}(\tau)\Big\}, \tag{B.1}$$

where c_n^{MR} is the critical value proposed in Bugni, Canay, and Shi (2017) and T_n is any test statistic that they allow for. The E-A-M algorithm can be used to compute the endpoints of this set so that the researcher may report an interval.

For ease of exposition, we discuss below the computation of the right end point of the confidence interval, which is the optimal value of the following problem:[34]

$$\begin{aligned}
&\max_{\tau \in \mathcal{T}} \tau \\
&\text{s.t.} \quad \inf_{\theta \in \Theta : p'\theta = \tau} T_n(\theta) \leqslant c_n^{MR}(\tau).
\end{aligned} \tag{B.2}$$

We then take $c(\tau) \equiv -\inf_{\theta \in \Theta : p'\theta = \tau} T_n(\theta) + c_n^{MR}(\tau)$ as a black-box function and apply the E-A-M algorithm.[35] We include the profiled statistic in the black-box function because it involves a non-linear optimization problem, which is also relatively expensive. The modified procedure is as follows.

Initialization: Draw randomly (uniformly) over $\mathcal{T} \subset \mathbb{R}$ a set $(\tau^{(1)}, \ldots, \tau^{(k)})$ of initial evaluation points and evaluate $c(\tau^{(\ell)})$ for $\ell = 1, \ldots, k-1$. Initialize $L = k$.

[34] The left end point is the optimal value of a program that replaces max with min.

[35] One may view (B.2) as a special case of (2.14) with a scalar control variable and a single constraint $g_1(\tau) \leqslant c(\tau)$ with $g_1(\tau) = 0$.

E-Step: Evaluate $c(\tau^{(L)})$ and record the tentative optimal value

$$\tau^{*,L} \equiv \max\{\tau^{\ell} : \ell \in \{1,\dots,L\}, c(\tau^{(\ell)}) \geqslant 0\}.$$

A-step: (Approximation) Approximate $\tau \mapsto c(\tau)$ by a flexible auxiliary model. We again use the kriging approximation, which for a mean-zero Gaussian process $\zeta(\cdot)$ indexed by τ and with constant variance ς^2 specifies

$$\Upsilon^{(\ell)} = \mu + \zeta(\tau^{(\ell)}),\ \ell = 1,\dots,L \tag{B.3}$$

$$Corr(\zeta(\tau),\zeta(\tau')) = K_\beta(\tau-\tau'),\ \tau,\tau' \in \mathbb{R}, \tag{B.4}$$

where K_β is a kernel with a scalar parameter $\beta \in [\underline{\beta},\overline{\beta}] \subset \mathbb{R}_{++}$. The parameters are estimated in the same way as before.

The (best linear) predictor of c and its derivative are then given by

$$c_L(\tau) = \hat{\mu} + \mathbf{r}_L(\tau)'\mathbf{R}_L^{-1}(\boldsymbol{\Upsilon} - \hat{\mu}\mathbf{1}), \tag{B.5}$$

$$\nabla_\tau c_L(\tau) = \hat{\mu} + \mathbf{Q}_L(\tau)\mathbf{R}_L^{-1}(\boldsymbol{\Upsilon} - \hat{\mu}\mathbf{1}), \tag{B.6}$$

where $\mathbf{r}_L(\tau)$ is a vector whose ℓ-th component is $Corr(\zeta(\tau),\zeta(\tau^{(\ell)}))$ as given above with estimated parameters, $\mathbf{Q}_L(\tau) = \nabla_\tau \mathbf{r}_L(\tau)'$, and $\mathbf{R}_L$ is an L-by-L matrix whose (ℓ,ℓ') entry is $Corr(\zeta(\tau^{(\ell)}),\zeta(\tau^{(\ell')}))$ with estimated parameters. The amount of uncertainty left in $c(\tau)$ is captured by the following variance:

$$\hat{\varsigma}^2 s_L^2(\tau) = \hat{\varsigma}^2\Big(1 - \mathbf{r}_L(\tau)'\mathbf{R}_L^{-1}\mathbf{r}_L(\tau) + \frac{(1-\mathbf{1}'\mathbf{R}_L^{-1}\mathbf{r}_L(\tau))^2}{\mathbf{1}'\mathbf{R}_L^{-1}\mathbf{1}}\Big). \tag{B.7}$$

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

$$\tau^{(L+1)} \equiv \arg\max_{\tau\in\mathcal{T}} \mathbb{EI}_L(\tau) = \arg\max_{\tau\in\mathcal{T}} (\tau - \tau^{*,L})_+\Big(1 - \Phi\Big(\frac{-c_L(\tau)}{\hat{\varsigma} s_L(\tau)}\Big)\Big). \tag{B.8}$$

With probability ϵ, draw $\tau^{(L+1)}$ randomly from a uniform distribution over $\mathcal{T}$.

As before, $\tau^{*,L}$ is reported as end point of CI_n upon convergence. In order for Theorem 3.2 to apply to this algorithm, the profiled statistic $\inf_{\theta\in\Theta:p'\theta=\tau} T_n(\theta)$ and the critical value $\hat{c}_n^{MR}$ need to be sufficiently smooth. We leave derivation of sufficient conditions for this to be the case to future research.

C An Entry Game Model and Some Monte Carlo Simulations

We evaluate the statistical and numerical performance of calibrated projection and E-A-M in comparison with BCS-profiling in a Monte Carlo experiment 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. The experiment simulates a two-player entry game in the Monte Carlo exercise of BCS, using their

code to implement their method.[36]

C.1 The General Entry Game Model

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

	$Y_2 = 0$	$Y_2 = 1$
$Y_1 = 0$	$0, 0$	$0, Z_2'\vartheta_1 + u_2$
$Y_1 = 1$	$Z_1'\vartheta_1 + u_1, 0$	$Z_1'(\vartheta_1 + \Delta_1) + u_1, Z_2'(\vartheta_2 + \Delta_2) + u_2$

Here, Y_ℓ, Z_ℓ, and u_ℓ denote player ℓ'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 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 $\theta \equiv (\vartheta_1', \vartheta_2', \Delta_1', \Delta_2', r)'$. 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:

$$E[1\{Y = (0,0)\}1\{Z = z\}] - G_r((-\infty, -z_1'\vartheta_1) \times (-\infty, -z_2'\vartheta_2))p_z = 0 \tag{C.1}$$

$$E[1\{Y = (1,1)\}1\{Z = z\}] - G_r([-z_1'(\vartheta_1 + \Delta_1), +\infty) \times [-z_2'(\vartheta_2 + \Delta_2), +\infty))p_z = 0 \tag{C.2}$$

$$E[1\{Y = (0,1)\}1\{Z = z\}] - G_r((-\infty, -z_1'(\vartheta_1 + \Delta_1)) \times [-z_2'\vartheta_2, +\infty))p_z \leqslant 0 \tag{C.3}$$

$$\begin{aligned} -E[1\{Y = (0,1)\}1\{Z = z\}] + \Big[& G_r((-\infty, -z_1'(\vartheta_1 + \Delta_1)) \times [-z_2'\vartheta_2, +\infty) \\ & - G_r([-z_1'\vartheta_1, -z_1'(\vartheta_1 + \Delta_1)) \times [-z_2'\vartheta_2, -z_2'(\vartheta_2 + \Delta_2))\Big] p_z \leqslant 0. \end{aligned} \tag{C.4}$$

We show in Online Appendix F that this model satisfies Assumptions D.1 and E.3-2.[37] Throughout, we analytically compute the moments' gradients and studentize them using sample analogs of their standard deviations.

C.2 A Comparison to 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 (C.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 indi-

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

[37] 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.

cators, $\Delta_\ell = [\delta_\ell; 0_{d_W}]$ for $\ell = 1, 2$, and $\vartheta_1 = \vartheta_2 = \vartheta = [0; \{\vartheta^{[k]}\}_{k=0}^{d_W}]$.[38] The parameter vector is $\theta = [\delta_1; \delta_2; \vartheta]$ with parameter space $\Theta = \{\theta \in \mathbb{R}^{2+d_W} : (\delta_1, \delta_2) \in [0,1]^2,\ \vartheta_k \in [0, \min\{\delta_1, \delta_2\}],\ k = 1, \ldots, d_W\}$. 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 $\delta_1 \in [0.3872, 0.4239]$, $\delta_2 \in [0.5834, 0.6084]$, $\vartheta^{[1]} \in [0.0996, 0.1006]$, $\vartheta^{[2]} \in [0.1994, 0.2010]$, and $\vartheta^{[3]} \in [0.2992, 0.3014]$.

The BCS-profiling confidence interval CI_n^{prof} inverts a test of $H_0 : p'\theta = \tau$ over a grid for τ. We do not in practice exhaust the grid but search inward from the extreme points of Θ in directions $\pm p$. At each τ that is visited, we use BCS code to compute a profiled test statistic and the corresponding critical value $\hat{c}_n^{MR}(\tau)$. The latter 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.[39] It computes the extreme points of CI_n by solving the nonlinear program (2.6) 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.[40] We run all methods with $B = 301$ bootstrap draws, and calibrated and "uncalibrated" (i.e., based on Andrews and Soares (2010)) projection also with $B = 1001$.[41] 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 2 displays results for (δ_1, δ_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. The most striking contrast is in computational effort. Here, uncalibrated projection is fastest – indeed, in contrast to received

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

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

[40] This is only one of several individually necessary stopping criteria. Others include that the current optimum $\theta^{*,L}$ and the expected improvement maximizer θ^{L+1} (see equation (2.21)) satisfy $|p'(\theta^{L+1} - \theta^{*,L})| \leqslant 0.005$. See Kaido, Molinari, Stoye, and Thirkettle (2017) for the full list of convergence requirements.

[41] Based on some trial runs of BCS-profiling for δ_1, we estimate that running it with $B = 1001$ throughout would take 3.14-times longer than the computation times reported in Table 2. By comparison, calibrated projection takes only 1.75-times longer when implemented with $B = 1001$ instead of $B = 301$.

wisdom, this procedure is computationally somewhat easy. This is due to our use of the E-A-M algorithm and therefore part of this paper's contribution. Next, our implementation of calibrated projection beats BCS-profiling with gridding by a factor of about 70. This can be disentangled into the gain from using calibrated projection, with its advantage of bootstrapping linear programs, and the gain afforded by the E-A-M algorithm. It turns out that implementing BCS-profiling with the adapted E-A-M algorithm (see Appendix B) improves computation by a factor of about 4; switching to calibrated projection leads to a further improvement by a factor of about 17. Finally, Table 3 extends the analysis to all components of θ and to 1000 Monte Carlo repetitions. We were unable to compute this for BCS-profiling.

In sum, the Monte Carlo experiment on the same DGP used in BCS yields three interesting findings: (i) The E-A-M algorithm accelerates projection of the Andrews and Soares (2010) confidence region to the point that this method becomes reasonably cheap; (ii) it also substantially accelerates computation of profiling intervals, and (iii) for this DGP, calibrated projection combined with the E-A-M algorithm has the most accurate size control while also being computationally attractive.

Tables

Table 1: Results for empirical application, with $\alpha = 0.05$, $\rho = 6.6055$, $n = 7882$, $\kappa_n = \sqrt{\ln n}$. "Direct search" refers to `fmincon` performed after E-A-M and starting from feasible points discovered by E-A-M, including the E-A-M optimum.

	CI_n		Computational Time		
	E-A-M	Direct Search	E-A-M	Direct Search	Total
ϑ_{LCC}^{cons}	$[-2.0603, -0.8510]$	$[-2.0827, -0.8492]$	24.73	32.46	57.51
ϑ_{LCC}^{size}	$[0.1880, 0.4029]$	$[0.1878, 0.4163]$	16.18	230.28	246.49
ϑ_{LCC}^{pres}	$[1.7510, 1.9550]$	$[1.7426, 1.9687]$	16.07	115.20	131.30
ϑ_{OA}^{cons}	$[0.3957, 0.5898]$	$[0.3942, 0.6132]$	27.61	107.33	137.66
ϑ_{OA}^{size}	$[0.3378, 0.5654]$	$[0.3316, 0.5661]$	11.90	141.73	153.66
ϑ_{OA}^{pres}	$[0.3974, 0.5808]$	$[0.3923, 0.5850]$	13.53	148.20	161.75
δ_{LCC}	$[-1.4423, -0.1884]$	$[-1.4433, -0.1786]$	15.65	119.50	135.17
δ_{OA}	$[-1.4701, -0.7658]$	$[-1.4742, -0.7477]$	13.06	114.14	127.23
r	$[0.1855, 0.85]$	$[0.1855, 0.85]$	5.37	42.38	47.78

Table 2: Results for Set 1 with $n = 4000$, $MCs = 300$, $B = 301$, $\rho = 5.04$, $\kappa_n = \sqrt{\ln n}$.

		Median CI			
	$1-\alpha$	CI_n^{prof}		CI_n	CI_n^{proj}
Implementation		Grid	E-A-M	E-A-M	E-A-M
$\delta_1 = 0.4$	0.95	[0.330,0.495]	[0.331,0.495]	[0.336,0.482]	[0.290,0.558]
	0.90	[0.340,0.485]	[0.340,0.485]	[0.343,0.474]	[0.298,0.543]
	0.85	[0.345,0.475]	[0.346,0.479]	[0.348,0.466]	[0.303,0.537]
$\delta_2 = 0.6$	0.95	[0.515,0.655]	[0.514,0.655]	[0.519,0.650]	[0.461,0.682]
	0.90	[0.525,0.647]	[0.525,0.648]	[0.531,0.643]	[0.473,0.675]
	0.85	[0.530,0.640]	[0.531,0.642]	[0.539,0.639]	[0.481,0.671]

		Coverage							
	$1-\alpha$	CI_n^{prof}				CI_n		CI_n^{proj}	
Implementation		Grid		E-A-M		E-A-M		E-A-M	
		Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
$\delta_1 = 0.4$	0.95	0.997	0.990	1.000	0.993	0.993	0.977	1.000	1.000
	0.90	0.990	0.980	0.993	0.977	0.987	0.960	1.000	1.000
	0.85	0.970	0.970	0.973	0.960	0.957	0.930	1.000	1.000
$\delta_2 = 0.6$	0.95	0.987	0.993	0.990	0.993	0.973	0.987	1.000	1.000
	0.90	0.977	0.973	0.980	0.977	0.940	0.953	1.000	1.000
	0.85	0.967	0.957	0.963	0.960	0.943	0.927	1.000	1.000

		Average Time			
	$1-\alpha$	CI_n^{prof}		CI_n	CI_n^{proj}
Implementation		Grid	E-A-M	E-A-M	E-A-M
$\delta_1 = 0.4$	0.95	1858.42	425.49	26.40	18.22
	0.90	1873.23	424.11	25.71	18.55
	0.85	1907.84	444.45	25.67	18.18
$\delta_2 = 0.6$	0.95	1753.54	461.30	26.61	22.49
	0.90	1782.91	472.55	25.79	21.38
	0.85	1809.65	458.58	25.00	21.00

Notes: (1) Projections of Θ_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. (5) "Implementation" refers to the method used to compute the extreme points of the confidence interval.

Table 3: Results for Set 1 with $n = 4000$, $MCs = 1000$, $B = 999$, $\rho = 5.04$, $\kappa_n = \sqrt{\ln n}$.

	$1-\alpha$	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	[0.333,0.478]	[0.288,0.555]	0.988	0.982	1	1	42.41	22.23
	0.90	[0.341,0.470]	[0.296,0.542]	0.976	0.957	1	1	41.56	22.11
	0.85	[0.346,0.464]	[0.302,0.534]	0.957	0.937	1	1	40.47	19.79
$\delta_2 = 0.6$	0.95	[0.525,0.653]	[0.466,0.683]	0.969	0.983	1	1	42.11	24.39
	0.90	[0.538,0.646]	[0.478,0.677]	0.947	0.960	1	1	40.15	28.13
	0.85	[0.545,0.642]	[0.485,0.672]	0.925	0.941	1	1	41.38	26.44
$\zeta^{[1]} = 0.1$	0.95	[0.054,0.142]	[0.020,0.180]	0.956	0.958	1	1	40.31	22.53
	0.90	[0.060,0.136]	[0.028,0.172]	0.911	0.911	1	1	36.80	24.15
	0.85	[0.064,0.132]	[0.032,0.167]	0.861	0.860	0.999	0.999	39.10	21.81
$\zeta^{[2]} = 0.2$	0.95	[0.156,0.245]	[0.121,0.281]	0.952	0.952	1	1	39.23	24.66
	0.90	[0.162,0.238]	[0.128,0.273]	0.914	0.910	0.998	0.998	41.53	21.66
	0.85	[0.165,0.234]	[0.133,0.268]	0.876	0.872	0.996	0.996	39.44	22.83
$\zeta^{[3]} = 0.3$	0.95	[0.257,0.344]	[0.222,0.379]	0.946	0.946	1	1	41.45	22.91
	0.90	[0.263,0.338]	[0.230,0.371]	0.910	0.909	0.997	0.999	42.09	22.83
	0.85	[0.267,0.334]	[0.235,0.366]	0.882	0.870	0.994	0.993	42.19	23.69

Notes: Same DGP and conventions as in Table 2.

Online Appendix: Confidence Intervals for Projections of Partially Identified Parameters

Contents

Structure of the Appendix

Section D states and proofs Theorem D.1, which establishes convergence-related results for our E-A-M algorithm. It also 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 E.1 presents the assumptions under which we prove asymptotic uniform validity of coverage of our procedure. Section F verifies some of our main assumptions for moment (in)equality models that have received much attention in the literature. Section G summarizes the notation we use and the structure of the

proof of Theorem 3.1,[42] and provides a proof of Theorems 3.1 (both under our main assumptions and under a high level assumption replacing Assumption E.3 and dropping the ρ-box constraints). Section H contains the statements and proofs of the lemmas used to establish Theorems 3.1 and D.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 3.1.

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

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

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

D.1.1 Assumption D.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 (D.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 F.1 we verify the condition for the canonical examples in the moment (in)equality literature.

ASSUMPTION D.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) φ_j *is Lipschitz continuous in* $x \in \mathbb{R}$ *for all* $j = 1, \ldots, J$.

D.1.2 Statement and Proof of Theorem D.1

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

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

where the family of functions ϕ_τ 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 D.1: *Suppose Assumptions E.1, E.2, E.4, E.5 and D.1 hold. Let τ_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,*

[42] Section G.1 provides in Table G.1 a summary of the notation used throughout, and in Figure G.1 and Table G.2 a flow diagram and heuristic explanation of how each lemma contributes to the proof of Theorem 3.1.

1.

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

2. Let $\hat{c}_{n,\tau_n}$ be defined as in (D.1) with τ_n replacing τ. Then there exists $C>0$ such that

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}} P\Big(\|\hat{c}_n-\hat{c}_{n,\tau_n}\|_\infty\leqslant C\varepsilon_n\Big)=1; \tag{D.3}$$

3. Let Assumption E.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

$$\hat{c}_{n,\rho,\tau}(\theta)\equiv\inf_{\lambda\in B^d_{n,\rho}}\hat{c}_{n,\tau}(\theta+\frac{\lambda\rho}{\sqrt{n}}). \tag{D.4}$$

For $c\geqslant 0$, let $U_n(\theta_n,c)$ be defined as in (G.25). Then,

$$\liminf_{n\to\infty} P_n\left(U_n(\theta_n,\hat{c}_{n,\rho,\tau_n})\neq\varnothing\right)\geqslant 1-\alpha. \tag{D.5}$$

4. Fix $P\in\mathcal{P}$ and n. There exists $R>0$ such that $\|\hat{c}_{n,\tau_n}\|_{\mathcal{H}_\beta}\leqslant R$.

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

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

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 ∞. By Assumption D.1-(i) and arguing as in the proof of Theorem 2 in Andrews (1994), condition (H.220) in Lemma H.11 is satisfied with $v=d$. Also, by Lemma H.13, (H.221) in Lemma H.11 holds with $\gamma=1$. This therefore ensures the conditions of Lemma H.11.

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

$$\begin{aligned}\left|\frac{m_j^2(x,\theta)}{\sigma_{P,j}^2(\theta)}-\frac{m_j^2(x,\theta)}{\sigma_{P,j}^2(\theta)}\right|&\leqslant\left|\frac{m_j(x,\theta)}{\sigma_{P,j}(\theta)}+\frac{m_j(x,\theta')}{\sigma_{P,j}(\theta')}\right|\left|\frac{m_j(x,\theta)}{\sigma_{P,j}(\theta)}-\frac{m_j(x,\theta')}{\sigma_{P,j}(\theta')}\right| &\text{(D.7)}\\ &\leqslant 2F(x)\bar{M}(x)\|\theta-\theta'\|. &\text{(D.8)}\end{aligned}$$

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

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

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

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

Further, for any $C>0$

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

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 H.12 (i). Note also that Assumption D.1-(i) ensures the conditions of Lemma H.12 (ii). Therefore, by Lemmas H.11-H.12 and Assumption E.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

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

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 $\mathbb{G}^b_{n,j}(\theta) - \mathfrak{G}^b_{n,j}(\theta) = \mathfrak{G}^b_{n,j}(\theta)\eta_{n,j}(\theta)$,

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

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

$$F_n \equiv \Big\{m_n \in Q : \sup_{\|\theta-\theta'\|\leqslant\tau_n}\|\mathfrak{G}^b_n(\theta) - \mathfrak{G}^b_n(\theta')\| \leqslant (\ln n)^2\tau_n,\ \sup_{\theta\in\Theta}\|\mathfrak{G}^b_n(\theta)\| \leqslant C\Big\}. \quad \text{(D.14)}$$

By Lemma H.11 (ii) and the asymptotic tightness of $\mathfrak{G}^b_n$, 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 (D.13) is bounded by $(\ln n)^2\tau_n + C/\sqrt{n}$ for some $C > 0$.

Next, by the triangle inequality and Assumption E.4,

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

Finally, note that by the Lipschitzness of φ_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

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

Hence,

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

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

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

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

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

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

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

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

Since η is arbitrary, we have

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

Reversing the roles of θ and θ' 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 (D.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 Part 2 and the definition of $\hat{c}_{n,\rho,\tau}$ in (D.4), it follows that

$$\begin{aligned}\hat{c}_{n,\rho,\tau_n}(\theta_n) &\geqslant \hat{c}_{n,\rho}(\theta_n) - e_n \\ &\geqslant c_{n,\rho}^I(\theta_n) - e_n,\end{aligned} \quad \text{(D.21)}$$

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

Part 4. 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 Θ, and hence

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

where υ 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 $\|\hat{c}_{n,\tau_n}\|_{\mathcal{H}_\beta} \leqslant C\|\hat{c}_{n,\tau_n}\|_{H^{\nu+d/2}} \leqslant R$.

□

D.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

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

where $0 < \underline{\beta}_k < \overline{\beta}_k < \infty$ for all k. The second one is the class of Matérn kernels (see, e.g., Rasmussen and Williams, 2005, Chapter 4) 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_ν is the modified Bessel function of the second kind.[43] The index ν controls the smoothness of K_β. 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 $K(\theta, \theta') = K_\beta(\theta - \theta')$ 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

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

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

[43] 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 2ν-Hölder continuity of K_β at the origin without a log factor (Bull, 2011, Assumption 4).

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

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

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 3.2 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 D.1 by introducing a mollified version of $\hat{c}_n$.

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

In (2.21), $\theta^{(L+1)}$ is defined as the maximizer of the following maximization problem

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

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

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

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

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

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

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

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.

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

This section explains in detail how $\hat{c}_n(\theta)$ in equation (2.13) 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 (2.11)), 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

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

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

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

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

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

(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.[44] Second, we exploit monotonicity in $\psi_b(\cdot)$, reducing the number of linear programs needed to be solved.

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

Appendix E Assumptions for Asymptotic Coverage Validity

E.1 Main 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 Appendix G.1 for the formal definitions. Below, ϵ, ε, δ, ω, $\underline{\sigma}$, 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 E.1: *(a) $\Theta \subset \mathbb{R}^d$ is a compact hyperrectangle with nonempty interior.*
(b) All distributions $P \in \mathcal{P}$ satisfy the following:

(i) $E_P[m_j(X_i,\theta)] \leqslant 0,\ j = 1,\ldots,J_1$ *and* $E_P[m_j(X_i,\theta)] = 0,\ j = J_1+1,\ldots,J_1+J_2$ *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* $j = 1,\ldots,J$ *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 E.2: *The function φ_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 E.3-2 is imposed, $\kappa_n = o(n^{1/4})$.*

Assumption E.1-(a) requires that Θ is a hyperrectangle, but can be replaced with the assumption that θ is defined through a finite number of nonstochastic inequality constraints smooth in θ and such that Θ is convex. Compactness is a standard assumption on Θ 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 θ; see (2.9). Assumption E.1-(b) defines our moment (in)equalities model. Assumption E.2 constrains the GMS function and the rate at which its tuning parameter diverges. Both E.1-(b) and E.2 are based on Andrews and Soares (2010) and are standard in the literature,[45] 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 E.3: *All distributions $P \in \mathcal{P}$ satisfy* ***one*** *of the following two conditions for some constants $\omega > 0, \underline{\sigma} > 0, \epsilon > 0, \varepsilon > 0, M < \infty$:*

1. *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*

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

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

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

2. *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* $j \in \mathcal{R}_1$,

$$m_{j+R_1}(X_i,\theta) = -m_j(X_i,\theta) - t_j(X_i,\theta). \tag{E.1}$$

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*

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

one has

$$\inf_{\theta\in\Theta_I(P)} \text{eig}(\tilde{\Omega}_P(\theta)) \geqslant \omega. \tag{E.2}$$

Finally,

$$\inf_{\theta\in\Theta_I(P)} \sigma_{P,j}(\theta) > \underline{\sigma} \text{ for } j = 1,\ldots,R_1. \tag{E.3}$$

Assumption E.3-1 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;[46] (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 E.3-1 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 C, is the Ciliberto and Tamer (2009) entry game model when the solution concept is pure strategy Nash equilibrium. We show in Appendix F.2 that these examples satisfy Assumption E.3-2.

Remark E.1: Assumption E.3-2 weakens E.3-1 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 $j = 1,\ldots,J$,

$$E_P[m_j(X_i,\theta)] = E_P[h_j(X_i)] - v_j(\theta), \tag{E.4}$$

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 E.3-2 for them in Appendix F.2. The argument can be generalized to other separable models.

In Appendix F.2, we also verify Assumption E.3-2 for some models that are not separable in the sense of equation (E.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).

[46] 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 what follows, we refer to pairs of inequality constraints indexed by $\{j, j+R_1\}$ and satisfying (E.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 (2.11). We explain it here for the case that the GMS function φ_j is the hard-thresholding one in footnote 8 of the main paper, and refer to Appendix H equations (H.12)-(H.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 (2.11). 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 3.1, we show that this tightens the stochastic program.[47] The rest of the procedure is unchanged.

Instead of Assumption E.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 Appendix G.2.2, we show that we can replace Assumption E.3 with a weaker high level condition (Assumption E.6) 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 E.3 implies Assumption E.6 is suggestive of how difficult these high-level conditions can be to verify.[48] Moreover, in Appendix E.3 we provide a simple example that violates Assumption E.3 and in which all of calibrated projection, BCS-profiling, and the bootstrap 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.[49]

Together with imposition of the ρ-box constraints, Assumption E.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)), 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 E.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 $\theta, \tilde{\theta} \in \Theta^\epsilon$ $\max_{j=1,\ldots,J} \|D_{P,j}(\theta) - D_{P,j}(\tilde{\theta})\| \leqslant M\|\theta - \tilde{\theta}\|$ and $\max_{j=1,\ldots,J} \sup_{\theta\in\Theta_I(P)} \|D_{P,j}(\theta)\| \leqslant \bar{M}$.

Assumption E.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 θ and P.[50] We require

[47] When paired inequalities are present, in equation (2.6) instead of $\hat{\sigma}_{n,j}$ we use the estimator $\hat{\sigma}^M_{n,j}$ specified in (H.192) in Lemma H.10 p.52 of the Appendix for $\sigma_{P,j}, j = 1,\ldots,2R_1$ (with $R_1 \leqslant J_1/2$ defined in the assumption). In equation (2.10) 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.

[48] Assumption E.3 is used exclusively to obtain the conclusions of Lemma H.6, H.7 and H.8, hence any alternative assumption that delivers such results can be used.

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

[50] The requirements are imposed on Θ^ϵ. Under Assumption E.3-1 it suffices they hold on Θ.

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 ϱ_P by

$$\varrho_P(\theta,\tilde{\theta}) \equiv \left\| \left\{ \left[Var_P\left(\sigma_{P,j}^{-1}(\theta) m_j(X,\theta) - \sigma_{P,j}^{-1}(\tilde{\theta}) m_j(X,\tilde{\theta})\right)\right]^{1/2} \right\}_{j=1}^{J} \right\|. \tag{E.5}$$

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 E.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 ϱ_P-equicontinuous. That is, for any $\epsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} P\left(\sup_{\varrho_P(\theta,\tilde{\theta})<\delta} \|\mathbb{G}_n(\theta) - \mathbb{G}_n(\tilde{\theta})\| > \epsilon \right) = 0. \tag{E.6}$$

(iii) Q_P satisfies

$$\lim_{\delta \downarrow 0} \sup_{\|(\theta_1,\tilde{\theta}_1)-(\theta_2,\tilde{\theta}_2)\|<\delta} \sup_{P\in\mathcal{P}} \|Q_P(\theta_1,\tilde{\theta}_1) - Q_P(\theta_2,\tilde{\theta}_2)\| = 0. \tag{E.7}$$

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

E.2 High Level Conditions Replacing Assumption E.3 and the ρ-Box Constraints

Next, we consider two high level assumptions. The first one aims at informally mimicking Assumption A.2 in Bugni, Canay, and Shi (2017) and replaces Assumption E.3. The second one replaces the use of the ρ-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 E.6: *Consider any sequence $\{P_n, \theta_n\} \in \{(P,\theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ such that*

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

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 Ω. Let

$$\mathfrak{w}_j(\lambda) \equiv \mathbb{Z}_j + \rho D_j \lambda + \pi_{1,j}^*. \tag{E.8}$$

Let

$$\mathfrak{W}(c) \equiv \left\{\lambda \in \mathfrak{B}_\rho^d : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c, \ \forall j = 1,\ldots,J\right\}, \tag{E.9}$$

$$c_{\pi *} \equiv \inf\{c \in \mathbb{R}_+ : \Pr(\mathfrak{W}(c) \neq \varnothing) \geqslant 1-\alpha\}. \tag{E.10}$$

Then:

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

2. *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* (G.25).

ASSUMPTION E.7: *Consider any sequence* $\{P_n, \theta_n\} \in \{(P, \theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ *as in Assumption E.6. Let*

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

which differs from (E.9) *by not constraining* λ *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.* $\Pr(\|\bar{\mathfrak{W}}(\bar{c})\|_H > M_\eta) \leqslant \eta$.

E.3 Example of Methods Failure When Assumption E.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 θ_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 E.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

θ 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 F Verification of Assumptions for the Canonical Partial Identification Examples

In this section we verify: (i) Assumption D.1 which is the crucial condition in Theorem D.1, and (ii) Assumption E.3-2, for the canonical examples in the partial identification 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

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

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

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

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

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

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

$$\Theta_I(P) = \{\theta \in \Theta \subset \mathbb{R}^d : \text{ equations (C.1), (C.2), (C.3), (C.4) hold for all } Z = z^r, r = 1, \ldots, k\}. \tag{F.4}$$

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

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

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

as it is done in, e.g., Ciliberto and Tamer (2009). We assume that for each of the four canonical examples under consideration, Assumption E.1 as well as one of the assumptions below hold.

ASSUMPTION F.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 F.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 F.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 F.1, while Beresteanu and Molinari (2008) assume F.1 itself.

F.1 Verification of Assumptions D.1 and A.2-(i)

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 $n^{-1/2}$. The same arguments, with small modification, deliver verification of Assumption A.2-(i) provided $\hat{\sigma}_{n,j}(\theta) > 0$.

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

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

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

   Assumption F.1 then guarantees that Assumption D.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 F.1-F.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 D.1 is satisfied for these rescaled functions by Assumptions F.1-F.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 H.12. Theorem D.1 then holds observing that $|P(Z=z^r)/(\sum_{i=1}^n \mathbf{1}(Z_i=z^r)/n)-1|=O_{\mathcal{P}}(n^{-1/2})$ and treating this random element similarly to how we treat $\eta_{n,j}(\cdot)$ in the proof of Theorem D.1.

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

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

hence is Lipschitz in θ with constant Z_iZ_i'. Under Assumptions F.1-F.2, $Var_P(m_r(X_i,\theta))$ is uniformly bounded from below, and Lipschitz in θ with a constant that depends on Z_i^4. Hence $\frac{m_r(X_i,\theta)}{\sigma_{P,r}(\theta)}$ is Lipschitz in θ with a constant that depends on powers of Z. Because Z has bounded support, Assumption D.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 E.1 and F.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.

F.2 Verification of Assumption E.3-2

Here we verify Assumption E.3-2 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 (E.1)-(E.2) are satisfied by construction, equation (E.3) is directly assumed.

2. **Linear regression with interval outcome data and discrete regressors.** Equation (E.1) is satisfied by construction. Given the estimator that we use for the population moment conditions, we verify equation (E.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 (E.3) follows from Assumption F.1. Concerning equation (E.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 (E.2) holds because the correlation matrix is diagonal.

3. **Best linear prediction with interval outcome data and discrete regressors.** Equation (E.1) is again satisfied by construction. Equation (E.2) holds under Assumptions F.1-F.2. Equation (E.3) is verified to hold under Assumption F.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 (C.3) and (C.4) are paired, but the corresponding moment functions differ by the model implied probability of the region of multiplicity, hence equation (E.1) is satisfied by construction. Given the estimator that we use for the population moment conditions, we verify equations (E.2) and (E.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 (E.2) holds provided that $|Corr(Y_{i1}(1 - Y_{i2}), Y_{i1}Y_{i2})| < 1 - \epsilon$ for some $\epsilon > 0$ and Assumption F.3 holds.[51] To see that equation (E.3) also holds, note that Assumption F.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.

[51] 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 G Proof of Theorem 3.1

G.1 Notation and Structure of the Proof of Theorem 3.1

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

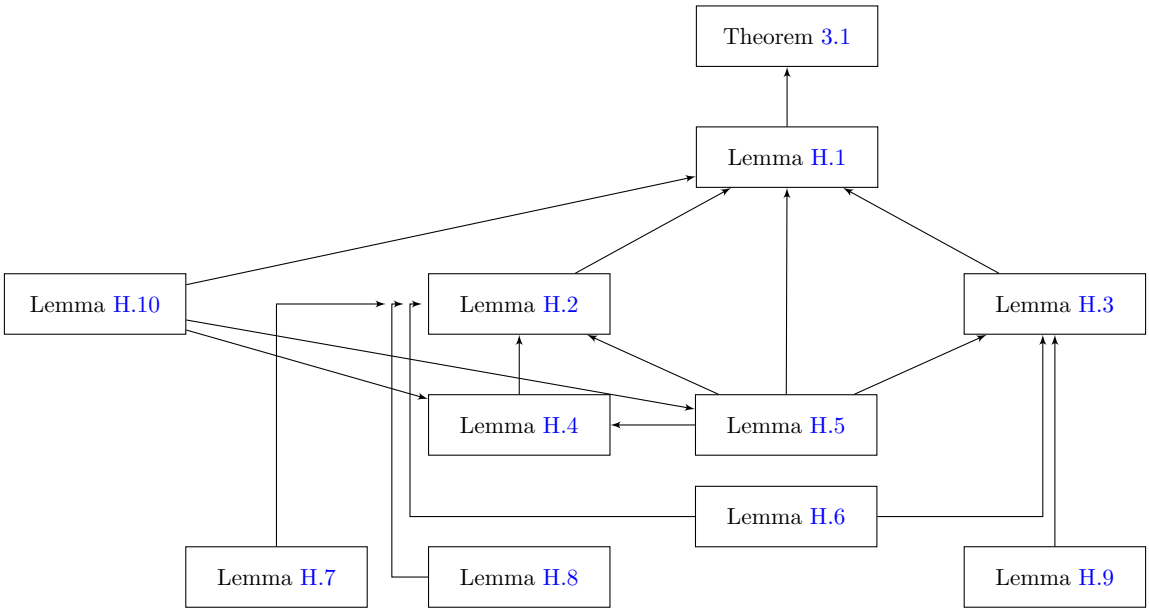
Table G.1: Important notation. Here $(P_n, \theta_n) \in \{(P,\theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ is a subsequence as defined in (G.3)-(G.4) below, $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$, $B^d = \{x \in \mathbb{R}^d : |x_i| \leqslant 1, i = 1, \ldots, d\}$, $B^d_{n,\rho} \equiv \frac{\sqrt{n}}{\rho}(\Theta - \theta_n) \cap B^d$, $\mathfrak{B}^d_\rho = \lim_{n\to\infty} B^d_{n,\rho}$, 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. θ, with estimator $\hat{D}_{n,j}(\cdot)$.
$\gamma_{1,P_n,j}(\cdot)$	=	$\frac{E_{P_n}(m_j(X_i,\cdot))}{\sigma_{P_n,j}(\cdot)},\ j = 1, \ldots, J$	Studentized population moments.
$\pi_{1,j}$	=	$\lim_{n\to\infty} \kappa_n^{-1}\sqrt{n}\gamma_{1,P_n,j}(\theta_n')$	Limit of rescaled population moments, constant $\forall \theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$ by Lemma H.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 φ_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 θ_n and $\theta_n + \frac{\lambda\rho}{\sqrt{n}}$.
$U_n(\theta_n, c)$	=	$\left\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap u_{n,j,\theta_n}(\lambda) \leqslant c,\ \forall j = 1, \ldots, J\right\}$	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.
$\mathfrak{W}(c)$	=	$\left\{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c,\ \forall j = 1, \ldots, J\right\}$	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.
$V^b_n(\theta_n', c)$	=	$\left\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap v^b_{n,j,\theta_n'}(\lambda) \leqslant c,\ \forall j = 1, \ldots, J\right\}$	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.
$V^I_n(\theta_n', c)$	=	$\left\{\lambda \in B^d_{n,\rho} : p'\lambda = 0 \cap v^I_{n,j,\theta_n'}(\lambda) \leqslant c,\ \forall j = 1, \ldots, J\right\}$	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 θ.
$\hat{\sigma}^M_{n,j}(\theta)$	=	$\hat{\mu}_{n,j}(\theta)\hat{\sigma}_{n,j}(\theta) + (1 - \hat{\mu}_{n,j}(\theta))\hat{\sigma}_{n,j+R_1}(\theta)$	Weighted sum of the estimators of the standard deviations of paired inequalities

Table G.2: Heuristics for the role of each Lemma in the proof of Theorem 3.1. Notes: (i) Uniformity in Theorem 3.1 is enforced arguing along subsequences; (ii) When needed, random variables are realized on the same probability space as shown in Lemma H.1 and Lemma H.17 (see Appendix H.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 (G.3)-(G.4) below; (iv) All results hold for any $\theta_n' \in (\theta_n + \rho/\sqrt{n}B^d) \cap \Theta$.

Theorem 3.1	$P_n(p'\theta_n \in CI) \geqslant P_n\left(U_n(\theta_n, \hat{c}_{n,\rho}(\theta_n)) \neq \varnothing\right)$. Coverage is conservatively estimated by the probability that U_n is nonempty.
Lemma H.1	$\liminf P_n\left(U_n(\theta_n, \hat{c}_{n,\rho}(\theta_n)) \neq \varnothing\right) \geqslant 1 - \alpha$.
Lemma H.2	$P_n(U(\theta_n, c_n^I(\theta_n)) \neq \varnothing, \mathfrak{W}(c_{\pi*}) = \varnothing) + P_n(U(\theta_n, c_n^I(\theta_n)) = \varnothing, \mathfrak{W}(c_{\pi*}) \neq \varnothing) = o_{\mathcal{P}}(1)$. Argued by comparing U_n and its limit $\mathfrak{W}$ (after coupling).
Lemma H.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 $c_{\pi*} > 0$. The bootstrap critical value that uses the less conservative GMS yileds a convergent critical value.
Lemma H.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 $v_{n,j,\theta_n'}^I$. The criterion functions entering U_n and $\mathfrak{W}$ converge to each other.
Lemma H.5	Local-to-binding constraints are selected by GMS uniformly over the ρ-box (intuition: $\rho n^{-1/2} = o_{\mathcal{P}}(\kappa_n^{-1})$), and $\lVert\hat{\xi}_n(\theta_n') - \kappa_n^{-1}\sqrt{n}\sigma_{P_n,j}^{-1}(\theta_n')E_{P_n}[m_j(X_i, \theta_n')]\rVert = o_{\mathcal{P}}(1)$.
Lemma H.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. It is unlikely that these sets are nonempty but become empty upon slightly tightening stochastic constraints.
Lemma H.7	Intersections of constraints whose gradients are almost linearly dependent are unlikely to realize inside $\mathfrak{W}$. Hence, we can ignore irregularities that occur as linear dependence is approached.
Lemma H.8	If there are weakly more equality constraints than parameters, then c is uniformly bounded away from zero. This simplifies some arguments.
Lemma H.9	If two paired inequalities are local to binding, then they are also asymptotically identical up to sign. This justifies "merging" them.
Lemma H.10	$\eta_{n,j}(\cdot)$ converges to zero uniformly in P and θ.

Figure G.1: Structure of Lemmas used in the proof of Theorem 3.1-(I).



G.2 Proof of Theorem 3.1

G.2.1 Main Proofs

Proof of Theorem 3.1-(I).

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

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

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

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

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

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

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

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

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

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

The projection of θ_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)) \\
\iff &\left\{ \begin{array}{c} \inf p'\vartheta \\ \text{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 \\ \text{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\} \\
\iff &\left\{ \begin{array}{c} \inf_\lambda p'\lambda \\ \text{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 \\ \text{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{(G.6)} \\
\iff &\left\{ \begin{array}{c} \inf_\lambda p'\lambda \\ \text{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 \\ \text{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{(G.7)}
\end{aligned}$$

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

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

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

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

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

Step 2. This step is used only when Assumption E.3-2 is invoked. When this assumption is invoked, recall that in equation (2.6) we use the estimator specified in Lemma H.10 equation (H.192) 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 (2.11) 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 (2.6) 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

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

where π^*_1 is defined in (G.5), let

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

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

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

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

with the following weighted sum of the paired inequalities

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

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

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

with

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

It then follows from Assumption E.3-2 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 (G.14) implies (G.13) and (G.16) implies (G.15).

Step 3. Next, we make the following comparisons:

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

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

For any constraint j for which $\pi^*_{1,j} = 0$, (G.17) yields that replacing $\sqrt{n}\gamma_{1,P_n,j}(\theta_n)$ in (G.9) with $\pi^*_{1,j}$ introduces a conservative distortion. Under Assumption E.3-2, for any j such that (G.10) holds, the substitutions in (G.14) and (G.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 $\pi^*_{1,j} = -\infty$, (G.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 E.4 (i)-(ii), and $\lambda \in B^d_{n,\rho}$, it follows that

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

where $\bar{M}$ and M are as defined in Assumption E.4-(i) and (ii) respectively, and we used that $\bar{\theta}_n$ lies component-wise between θ_n and $\theta_n + \frac{\lambda\rho}{\sqrt{n}}$. Using that $\mathbb{G}_{n,j}$ is asymptotically tight by Assumption E.5, we have that for any $\tau > 0$, there exists a $T > 0$ and $N_1 \in \mathbb{N}$ such that for all $n \geqslant N_1$,

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

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

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

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

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

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

Since the choice of τ is arbitrary, the limit of the term in (G.21) is not smaller than the limit of the first term in (G.22). Hence, we continue arguing for the event whose probability is evaluated in (G.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 (G.22) is implied by the event

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

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

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

where under Assumption E.3-2 when $\pi^*_{1,j} = 0$ and $\pi^*_{1,j+R_1} = 0$ the substitutions of equation (G.13) with equation (G.14) and of equation (G.15) with equation (G.16) have been performed. Let

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

and define

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

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

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

By passing to a further subsequence, we may assume that

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

for some $J \times d$ matrix D such that $\|D\| \leqslant M$ and $\Omega_{P_n} \xrightarrow{u} \Omega$ for some correlation matrix Ω. By Lemma 2 in Andrews and Guggenberger (2009) and Assumption E.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 Ω. By Lemma H.1,

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

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

Proof of Theorem 3.1-(II).

The result follows immediately from the same steps as in the proof of Theorem 3.1-(I). □

Proof of Theorem 3.1-(III)

The argument of proof is the same as for Theorem 3.1-(I), with the following modification. Take (P_n, θ_n) as defined following equation (G.4). Then $f(\theta_n)$ is covered when

$$\begin{aligned}
&\left\{\begin{array}{l} \inf f(\vartheta) \\ \text{s.t. } \vartheta \in \Theta, \quad \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leqslant \hat{c}^f_n(\vartheta), \forall j \end{array}\right\} \leqslant f(\theta_n) \leqslant \left\{\begin{array}{l} \sup f(\vartheta) \\ \text{s.t. } \vartheta \in \Theta, \quad \frac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leqslant \hat{c}^f_n(\vartheta), \forall j \end{array}\right\} \\
\iff &\left\{\begin{array}{l} \inf_\lambda \nabla f(\tilde{\theta}_n)\lambda \\ \text{s.t. } \lambda \in \frac{\sqrt{n}}{\rho}(\Theta - \theta_n), \quad \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} \leqslant \hat{c}^f_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \forall j \end{array}\right\} \leqslant 0 \\
&\qquad \leqslant \left\{\begin{array}{l} \sup_\lambda \nabla f(\tilde{\theta}_n)\lambda \\ \text{s.t. } \lambda \in \frac{\sqrt{n}}{\rho}(\Theta - \theta_n), \quad \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} \leqslant \hat{c}^f_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}), \forall j \end{array}\right\},
\end{aligned}$$

where we took a mean value expansion yielding

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

for $\tilde{\theta}_n$ a mean value that lies componentwise between θ_n and $\theta_n + \frac{\lambda\rho}{\sqrt{n}}$, and we used that the sign of the last term

in (G.30) is the same as the sign of $\nabla f(\tilde{\theta}_n)\lambda$. With the objective function in (G.30) so redefined, all expression in the proof of Theorem 3.1-(I) up to (G.24) continue to be valid. We can then redefine the set $U_n(\theta_n, c)$ in (G.25) as

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

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 (H.17) is unchanged. This yields the result, noting that by the assumption $\|\nabla f(\tilde{\theta}_n) - \nabla f(\theta'_n)\| = O_{\mathcal{P}}(\rho/\sqrt{n})$ □

G.2.2 Proof of Theorem 3.1-(I) with High Level Assumption E.6 Replacing Assumption E.3, and Dropping the ρ-Box Constraints Under Assumption E.7

LEMMA G.1: *Suppose that Assumption E.1, E.2, E.4 and E.5 hold.*

(I) Let also Assumption E.6 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 E.7 and either Assumption E.3 or E.6 hold. Let $\hat{c}_n = \inf\{c \in \mathbb{R}_+ : P^*(\{\Lambda^b_n(\theta, +\infty, c) \cap \{p'\lambda = 0\}\} \neq \varnothing) \geqslant 1 - \alpha\}$, where Λ^b_n is defined in equation (2.11) 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 (2.6). 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 E.3 with Assumption E.6. Hence we establish the result by showing that all claims that were made under Assumption E.3 remain valid under Assumption E.6. We proceed in steps.

Step 1. Revisiting the proof of Lemma H.6, equation (H.137).

Let $\mathcal{J}^*$ be as defined in (H.29). If $\mathcal{J}^* = \varnothing$ we immediately have that Lemma H.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

$$\mathfrak{W}(c - \delta) \equiv \left\{\lambda \in \mathfrak{B}^d_\rho : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c - \delta,\ \forall j = 1, \ldots, J\right\}, \tag{G.31}$$

where we emphasize that the set $\mathfrak{W}(c-\delta)$ is obtained by a δ-contraction of all constraints, including those indexed by $j = J_1 + 1, \ldots, J$. By Assumption E.6, for any $\eta > 0$ there exists a δ 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 H.2.

Case 1 of Lemma H.2 is unaltered. Case 2 of Lemma H.2 follows from the same argument as used in Case 1 of Lemma H.2, because under Assumption E.6 as shown in step 1 of this proof all inequalities are tightened. In Case 3 of Lemma H.2 the result in (G.29) holds automatically by Assumption E.6-(ii). (As a remark, Lemmas H.7-H.8 are no longer needed to establish Lemma H.2.)

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

In summary, the desired result follows by applying Lemma H.1 in the proof of Theorem 3.1-(I) as Lemmas H.2, H.3 and H.6 remain valid, Lemmas H.4, H.5, H.10 and the Lemmas in Appendix H.3 are unaffected, and Lemmas H.7, H.8, H.9 are no longer needed.

Part (II). This is established by adapting the proof of Theorem 3.1-(I) 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 ρ. 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 $\lambda \in B_{n,\rho}^d$ in equation (G.9). Indeed, (G.9) is replaced by:

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

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

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 ρ in the main proof use that $\|\lambda\| \leqslant \sqrt{d}\rho = O_{\mathcal{P}}(1)$. For example, consider the argument following equation (G.19) or the argument just preceding equation (G.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 H.1. The next major adaptation is that in (H.27) and (H.28): we again drop ρ 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 H.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 H.7. This requires straightforward adaptation of algebra as $\bar{\mathfrak{W}}(\bar{c})$ is only stochastically and not deterministically bounded.

Finally, in Lemma H.3 we remove the ρ-constraint from V_n^b and V_n^I 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 $\tilde{V}_n^b$ and $\tilde{V}_n^I$ with $\widetilde{\bar{\mathfrak{W}}}$. To ensure that λ is uniformly stochastically bounded in expressions like (H.98), we verify that the modified $\tilde{V}_n^b$ and $\tilde{V}_n^I$ inherit the property in Assumption E.7. 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 $\lambda \in \tilde{V}_n^b(\theta_n', c)$ or for $\lambda \in \tilde{V}_n^I(\theta_n', c)$, $0 < c \leqslant \bar{c}$, satisfies this

condition. By Assumption E.7 and the Cauchy-Schwarz inequality, $\max_{\lambda \in \widetilde{\mathfrak{W}}(c)} t'\lambda = O(1)$ 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, $\mathbb{Z}_i^C$ be the vector of realizations $\mathbb{Z}_i^j$ corresponding to $j \in C$, and $K^C(\theta_n')$ the matrix that stacks the corresponding gradients; then $(K^C(\theta_n'))^{-1}(\bar{c}\mathbf{1} - \mathbb{Z}_i^C) = O(1)$. By Lemma H.7 and the fact that $\hat{D}_n(\theta_n') \xrightarrow{P} D$ by Assumption E.4, we then also have that $(\hat{K}^C(\theta_n'))^{-1}(\bar{c}\mathbf{1} - \mathbb{G}_{n,j}^b) = O_{\mathcal{P}}(1)$, 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 $V_n^I(\theta_n', c)$ and $\widetilde{\mathfrak{W}}(c)$ in Lemma H.3 goes through. □

G.2.3 An Extension of Theorem 3.1

In this subsection, we establish that, under the assumptions of Theorem 3.1, we actually have

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} P(p'\theta \in \{p'\vartheta : \vartheta \in \mathcal{C}_n(\hat{c}_n)\}) \geqslant 1-\alpha. \tag{G.32}$$

In words, the mathematical projection of $\mathcal{C}_n(\hat{c}_n)$, which will asymptotically pick up gaps in the projection of Θ_I, is a uniformly asymptotically valid confidence region. This strengthens Theorem 3.1 because $\{p'\vartheta : \vartheta \in \mathcal{C}_n(\hat{c}_n)\} \subseteq CI_n$.

To prove this extension, we modify the proof of Theorem 3.1 after (G.5) as follows: The projection of θ_n is covered when

$$\exists \vartheta \in \Theta : p'\vartheta = p'\theta_n, \tfrac{\sqrt{n}\bar{m}_{n,j}(\vartheta)}{\hat{\sigma}_{n,j}(\vartheta)} \leqslant \hat{c}_n(\vartheta), \forall j \tag{G.33}$$

$$\iff \exists \lambda \in \tfrac{\sqrt{n}}{\rho}(\Theta - \theta_n) : p'\lambda = 0, \tfrac{\sqrt{n}\bar{m}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})}{\hat{\sigma}_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}})} \leqslant \hat{c}_n(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}), \forall j \tag{G.34}$$

$$\iff \exists \lambda \in \tfrac{\sqrt{n}}{\rho}(\Theta - \theta_n) : \tag{G.35}$$

$$p'\lambda = 0, \big(\mathbb{G}_{n,j}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda + \sqrt{n}\gamma_{1,P_n,j}(\theta_n)\big)\big(1 + \eta_{n,j}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}})\big) \leqslant \hat{c}_n(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}), \forall j \tag{G.36}$$

where the last line corresponds to (G.7) and intermediate steps that are exactly analogous to the previous proof were skipped. Subsequent proof steps go through as before until, comparing (G.25) to (G.36), we find (compare to (G.27), noting the change from inequality to equality)

$$P_n\big(p'\theta_n \in \{p'\vartheta : \vartheta \in \mathcal{C}_n(\hat{c}_n)\}\big) = P_n\big(U_n(\theta_n, \hat{c}_{n,\rho}) \neq \varnothing\big). \tag{G.37}$$

The proof then continues as before.

Appendix H Auxiliary Lemmas

H.1 Lemmas Used to Prove Theorem 3.1

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 3.1-(I). That is, along (P_n, θ_n), one has

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

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

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

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

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

where $\pi^*_{1,j}$ is defined in (G.5) and we used Lemma H.5. Under Assumption E.3-2 if

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

we replace the constraints

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

with

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

where

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

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

When Assumption E.3-2 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 (2.11) as described on p.11 of the paper; when it is invoked with a GMS function φ that is smooth in its argument, replace them, respectively, with

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

where

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

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

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

Due to the substitutions in equations (H.6)-(H.9), the paired inequalities (i.e., inequalities for which (H.5) holds under Assumption E.3-2) 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

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

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

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

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

LEMMA H.1: *Let Assumptions E.1, E.2, E.3, E.4, and E.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,*

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

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

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

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 H.3. By Lemma H.3 (iii), this critical value sequence satisfies (H.20) with probability approaching 1. Further, by Lemma H.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

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

Since the $o_{P_n}(1)$ term in (H.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 (H.20) and simple addition and subtraction,

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

As previously argued, $\mathbb{G}_n(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) \xrightarrow{d} \mathbb{Z}$. Moreover, by Lemma H.10, $\sup_{\theta\in\Theta} \|\eta_n(\theta)\| \xrightarrow{p} 0$ uniformly in $\mathcal{P}$, and by Lemma H.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

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

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 $\mathbb{Z}^* \stackrel{d}{=} \mathbb{Z}$.

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

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

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

where we used that by Lemma H.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 (H.6)-(H.9) for the case that $\pi_{1,j}^* = 0 = \pi_{1,j+R_1}^*$. Similarly, we define analogs to the sets in (G.25) and (H.16) as

$$U_n^*(\theta_n, c_n^*) \equiv \{\lambda \in B_{n,\rho}^d : p'\lambda = 0 \cap u_{n,j,\theta_n}^*(\lambda) \leqslant c_n^*,\ \forall j = 1, \ldots, J\}, \tag{H.26}$$

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

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

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

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 (H.28) tends to 0 as $n \to \infty$.

Define

$$\mathcal{J}^* \equiv \{j = 1, \cdots, J : \pi_{1,j}^* = 0\}. \tag{H.29}$$

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

$$U_n^*(\theta_n, c_n^*) = \{\lambda \in B_{n,\rho}^d : p'\lambda = 0\}, \quad \mathfrak{W}^*(c_{\pi*}) = \{\lambda \in \mathfrak{B}_\rho^d : p'\lambda = 0\}, \tag{H.30}$$

with probability 1, and hence

$$\mathbf{P}\Big(\{U_n^*(\theta_n, c_n^*) \neq \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi*}) \neq \varnothing\}\Big) = 1. \tag{H.31}$$

This in turn implies that

$$\Big|\mathbf{P}\Big(U_n^*(\theta_n, c_n^*) \neq \varnothing\}\Big) - \mathbf{P}\Big(\mathfrak{W}^*(c_{\pi*}) \neq \varnothing\}\Big)\Big| = 0, \tag{H.32}$$

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 (H.28) is 0.

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

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

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 (H.33) converges to zero. But this is guaranteed by Lemma H.2. Therefore, the conclusion of the lemma follows. □

LEMMA H.2: *Let Assumptions E.1, E.2, E.3, E.4, and E.5 hold. Let (P_n, θ_n) have the almost sure representations given in Lemma H.1, and let $\mathcal{J}^*$ be defined as in (H.29). Assume that $\mathcal{J}^* \neq \varnothing$. Then for any $\eta > 0$, there*

exists $N \in \mathbb{N}$ *such that*

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

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

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

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

$$\Big(\mathbb{G}_{n,j}^*(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda + \pi_{1,j}^*\Big)(1 + \eta_{n,j}^*) \leqslant c_n^*,$$
$$\mathbb{Z}_j^* + \rho D_j \lambda + \pi_{1,j}^* \leqslant c_{\pi *}$$

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

$$U_n^*(\theta_n, c_n^*) \equiv \big\{\lambda \in B_{n,\rho}^d : p'\lambda = 0 \cap u_{n,j,\theta_n}^*(\lambda) \leqslant c_n^*,\ \forall j \in \mathcal{J}^*\big\}, \tag{H.36}$$

$$\mathfrak{W}^*(c_{\pi *}) \equiv \big\{\lambda \in \mathfrak{B}_\rho^d : p'\lambda = 0 \cap \mathfrak{w}_j^*(\lambda) \leqslant c_{\pi *},\ \forall j \in \mathcal{J}^*\big\}. \tag{H.37}$$

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

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

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

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

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

$$U_n^{*,+\delta}(\theta_n, c) \equiv \{\lambda \in B_{n,\rho}^d : p'\lambda = 0 \cap u_{n,j,\theta_n}^*(\lambda) \leqslant c + \delta,\ j \in \mathcal{J}^*\}, \tag{H.40}$$

$$\mathfrak{W}^{*,+\delta}(c) \equiv \{\lambda \in \mathfrak{B}_\rho^d : p'\lambda = 0 \cap \mathfrak{w}_j^*(\lambda) \leqslant c + \delta,\ j \in \mathcal{J}^*\}. \tag{H.41}$$

Compared to the set in equation (H.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 $A_n^c \subseteq L_n$.

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

$$\mathbf{P}\Big(\{U_n^*(\theta_n, c_n^*) \neq \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi *}) = \varnothing\}\Big) \leqslant \mathbf{P}\Big(\{U_n^*(\theta_n, c_n^*) \neq \varnothing\} \cap \{\mathfrak{W}^{*,+\delta}(c_{\pi *}) = \varnothing\}\Big)$$
$$+ \mathbf{P}\Big(\{\mathfrak{W}^{*,+\delta}(c_{\pi *}) \neq \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi *}) = \varnothing\}\Big), \tag{H.42}$$

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 (H.42) can further be bounded as

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

where the penultimate inequality follows from $A_n^c \subseteq L_n$ as argued above, and the last inequality follows from (H.39).

For the second term on the left hand side of (H.42), by Lemma H.6, there exists $N' \in \mathbb{N}$ such that

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

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

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

Case 1. Suppose first that $J_2 = 0$ (recalling that under Assumption E.3-2 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 (H.36) and (H.37), one may write

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

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

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

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

and note that $A^c_n \subseteq R_{2n}$. The result in equation (H.35) then follows by Lemma H.6 using again similar steps to (H.42)-(H.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 δ the inequality constraints as well as each of the two opposing inequalities obtained from the equality constraints. That is,

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

that is, the same set as in (H.137) with $\mathfrak{w}^*_j(\lambda)$ replacing $\mathfrak{w}_j(\lambda)$ and defining the set using only inequalities in $\mathcal{J}^*$. Note that, by Lemma H.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 H.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 (H.47) now defined using the tightened constraint set $\mathfrak{W}^{*,-\delta}(c_{\pi*})$ being defined as in (H.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 $\mathbf{P}$),

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

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

$$\mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) = \varnothing\} \cap \{\mathfrak{W}^*(c_{\pi*}) \neq \varnothing\}\Big) \tag{H.50}$$

$$\leqslant \mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) = \varnothing\} \cap \{\mathfrak{W}^{*,-\delta}(0) \neq \varnothing\}\Big) + \mathbf{P}\Big(\{\mathfrak{W}^{*,-\delta}(0) = \varnothing\} \cap \{\mathfrak{W}^*(0) \neq \varnothing\}\Big), \tag{H.51}$$

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

there exists $\delta > 0$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\Big(\{\mathfrak{W}^{*,-\delta}(0) = \varnothing\} \cap \{\mathfrak{W}^{*}(0) \neq \varnothing\}\Big) < \eta/3 \text{ for all } n \geqslant N. \tag{H.52}$$

Therefore, the second term on the right hand side of (H.51) can be made arbitrarily small.

We now consider the first term on the right hand side of (H.51). Let g be a $J + 2d + 2$ vector with

$$g_j = \begin{cases} -\mathbb{Z}_j, & j \in \mathcal{J}^*, \\ 0, & j \in \{1, \cdots, J\} \backslash \mathcal{J}^*, \\ 1, & j = J+1, \ldots, J+2d, \\ 0, & j = J+2d+1, J+2d+2, \end{cases} \tag{H.53}$$

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 (H.36)-(H.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

$$K \equiv \begin{bmatrix} [\rho D_j]_{j=1}^{J_1+J_2} \\ [-\rho D_{j-J_2}]_{j=J_1+J_2+1}^{J} \\ I_d \\ -I_d \\ p' \\ -p' \end{bmatrix}. \tag{H.54}$$

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 H.7 establishes that for any $\eta > 0$, there exist $\varepsilon_\eta > 0$ and N s.t. $n \geqslant N$ implies

$$\mathbf{P}\left(\mathfrak{W}^{*,-\delta}(0) \neq \varnothing, \left|\det K^{\mathcal{C}}\right| \leqslant \varepsilon_\eta\right) \leqslant \eta, \tag{H.55}$$

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

$$\mathbf{P}\Big(\left|\det\left(K^{\mathcal{C}}\right)^{-1}\right| \leqslant M\Big) \geqslant 1 - \eta, \ \forall n \geqslant N. \tag{H.56}$$

Let g_n be a $J+2d+2$ vector with

$$g_{n,j}(\theta+\lambda/\sqrt{n}) \equiv \begin{cases} c_n^*/(1+\eta_{n,j}^*) - \mathbb{G}_{n,j}^*(\theta+\frac{\lambda\rho}{\sqrt{n}}) & \text{if } j \in \mathcal{J}^*, \\ 0, & \text{if } j \in \{1,\cdots,J\}\backslash\mathcal{J}^*, \\ 1, & \text{if } j = J+1,\ldots,J+2d, \\ 0, & \text{if } j = J+2d+1, J+2d+2, \end{cases} \tag{H.57}$$

using again that $\pi_{1,j}^* = 0$ for $j \in \mathcal{J}^*$. For each $P \in \mathcal{P}$, let

$$K_P(\theta,\rho) \equiv \begin{bmatrix} [\rho D_{P,j}(\theta)]_{j=1}^{J_1+J_2} \\ [-\rho D_{P,j-J_2}(\theta)]_{j=J_1+J_2+1}^{J} \\ I_d \\ -I_d \\ p' \\ -p' \end{bmatrix}. \tag{H.58}$$

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

$$\phi_n(\lambda) \equiv \left(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n,\lambda),\rho)\right)^{-1} g_n^{\mathcal{C}}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}), \tag{H.59}$$

where the notation $\bar{\theta}(\theta_n,\lambda)$ emphasizes that $\bar{\theta}$ depends on θ_n and λ because it lies component-wise between θ_n and $\theta_n + \frac{\lambda\rho}{\sqrt{n}}$. We show that ϕ_n is a contraction mapping and hence has a fixed point.

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

$$\begin{aligned} \|\phi_n(\lambda) - \phi_n(\lambda')\| &= \left\| \left(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n,\lambda),\rho)\right)^{-1} g_n^{\mathcal{C}}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) - \left(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n,\lambda'),\rho)\right)^{-1} g_n^{\mathcal{C}}(\theta_n + \tfrac{\lambda'\rho}{\sqrt{n}}) \right\| \\ &\leqslant \left\| \left(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n,\lambda),\rho)\right)^{-1} \right\|_2 \left\| g_n^{\mathcal{C}}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) - g_n^{\mathcal{C}}(\theta_n + \tfrac{\lambda'\rho}{\sqrt{n}}) \right\| \\ &\quad + \left\| \left(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n,\lambda),\rho)\right)^{-1} - \left(K_{P_n}^{\mathcal{C}}(\bar{\theta}(\theta_n,\lambda'),\rho)\right)^{-1} \right\|_2 \left\| g_n^{\mathcal{C}}(\theta_n + \tfrac{\lambda'\rho}{\sqrt{n}}) \right\|, \end{aligned} \tag{H.60}$$

where $\|\cdot\|_2$ denotes the spectral norm (induced by the Euclidean norm).

By Assumption E.5 (ii), for any $\eta > 0$, $k > 0$, there is $N \in \mathbb{N}$ such that

$$\mathbf{P}\left(\left\| g_n^{\mathcal{C}}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) - g_n^{\mathcal{C}}(\theta_n + \tfrac{\lambda'\rho}{\sqrt{n}}) \right\| \leqslant k\|\lambda - \lambda'\| \right) \tag{H.61}$$

$$= \mathbf{P}\left(\left\| \mathbb{G}_n^{*,\mathcal{C}}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) - \mathbb{G}_n^{*,\mathcal{C}}(\theta_n + \tfrac{\lambda'\rho}{\sqrt{n}}) \right\| \leqslant k\|\lambda - \lambda'\| \right) \geqslant 1-\eta, \ \forall n \geqslant N. \tag{H.62}$$

Moreover, by arguing as in equation (G.20), for any η there exist $0 < L < \infty$ and $N \in \mathbb{N}$ such that $\forall n \geqslant N$

$$\mathbf{P}\left(\sup_{\lambda' \in B^d} \left\| g_n^{\mathcal{C}}(\theta_n + \tfrac{\lambda'\rho}{\sqrt{n}}) \right\| \leqslant L \right) \geqslant 1-\eta. \tag{H.63}$$

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 H.7 and the definition of $\mathcal{C}$, for any $\eta > 0$, there exist $0 < L < \infty$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\left(\left\| \left(K^{\mathcal{C}}\right)^{-1} \right\|_2 \leqslant L \right) \geqslant 1-\eta, \ \forall n \geqslant N, \tag{H.64}$$

By Horn and Johnson (1985, ch. 5.8), for any invertible matrices $K, \tilde{K}$ such that $\|\tilde{K}^{-1}(K-\tilde{K})\|_2 < 1$,

$$\|K^{-1} - \tilde{K}^{-1}\|_2 \leqslant \frac{\|\tilde{K}^{-1}(K-\tilde{K})\|_2}{1 - \|\tilde{K}^{-1}(K-\tilde{K})\|_2} \|\tilde{K}^{-1}\|_2. \tag{H.65}$$

By the assumption that $D_{P_n}(\theta_n) \to D$ and Assumption E.4, for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that

$$\sup_{\lambda \in B^d} \|K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho) - K^{\mathcal{C}}\|_2 \leqslant \eta, \ \forall n \geqslant N. \tag{H.66}$$

By (H.65), 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

$$\begin{aligned}
&\mathbf{P}\big(\sup_{\lambda \in B^d} \big\| \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\big)^{-1} \big\|_2 \leqslant 2L_1 \big) \\
&\geqslant \mathbf{P}\big(\big\| \big(K^{\mathcal{C}}\big)^{-1} \big\|_2 + \sup_{\lambda \in B^d} \|K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)^{-1} - (K^{\mathcal{C}})^{-1}\|_2 \leqslant 2L_1 \big) \\
&\geqslant \mathbf{P}\left(\big\| \big(K^{\mathcal{C}}\big)^{-1} \big\|_2 \leqslant L_1, \frac{\big\| \big(K^{\mathcal{C}}\big)^{-1} \big\|_2^2}{1 - \big\| \big(K^{\mathcal{C}}\big)^{-1} (K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho) - K^{\mathcal{C}}) \|_2} \leqslant L_2, \ \sup_{\lambda \in B^d} \|K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho) - K^{\mathcal{C}}\|_2 \leqslant \frac{L_1}{L_2} \right) \\
&\geqslant 1 - 2\eta, \ \forall n \geqslant N,
\end{aligned} \tag{H.67}$$

Again by applying (H.65), for any $k > 0$, there exists $N \in \mathbb{N}$ such that

$$\mathbf{P}\big(\big\| \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda))\big)^{-1} - \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda'))\big)^{-1} \big\|_2 \leqslant k\|\lambda - \lambda'\| \big) \tag{H.68}$$

$$\geqslant \mathbf{P}\Big(\sup_{\lambda \in B^d} \big\| \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda))\big)^{-1} \big\|_2^2 M\rho \|\bar{\theta}(\theta_n, \lambda) - \bar{\theta}(\theta_n, \lambda')\| \leqslant k\|\lambda - \lambda'\| \Big) \geqslant 1 - \eta, \ \forall n \geqslant N, \tag{H.69}$$

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 E.4 (ii), and the last inequality follows from (H.67).

By (H.60)-(H.63) and (H.67)-(H.69), it then follows that there exists $\beta \in [0, 1)$ such that for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that

$$\mathbf{P}\left(|\phi_n(\lambda) - \phi_n(\lambda')| \leqslant \beta\|\lambda - \lambda'\|, \ \ \forall \lambda, \lambda' \in B^d \right) \geqslant 1 - \eta, \ \forall n \geqslant N. \tag{H.70}$$

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

$$\mathbf{P}\left(\exists \lambda_n^f : \lambda_n^f = \phi_n(\lambda_n^f) \right) \geqslant 1 - \eta, \ \forall n \geqslant N. \tag{H.71}$$

Next, define the mapping

$$\psi_n(\lambda) \equiv \big(K^{\mathcal{C}}\big)^{-1} g^{\mathcal{C}}. \tag{H.72}$$

This map is constant in λ 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 (H.59) and (H.72) arguing as in (H.60), it follows that for any $\lambda \in B^d$,

$$\|\psi_n(\lambda) - \phi_n(\lambda)\| \leqslant \Big\| \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\big)^{-1} \Big\|_2 \Big\| g^{\mathcal{C}} - g^{\mathcal{C}}_n(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) \Big\| + \Big\| \big(K^{\mathcal{C}}\big)^{-1} - \big(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n, \lambda), \rho)\big)^{-1} \Big\|_2 \|g^{\mathcal{C}}\|. \tag{H.73}$$

By (H.53) and (H.57)

$$\begin{aligned}
\Big\| g^{\mathcal{C}} - g^{\mathcal{C}}_n(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) \Big\| &\leqslant \max_{j \in \mathcal{J}^*} | - \mathbb{Z}^*_j - c^*_n/(1 + \eta^*_{n,j}) + \mathbb{G}^*_{n,j}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}})| \\
&\leqslant \max_{j \in \mathcal{J}^*} |\mathbb{Z}^*_j - \mathbb{G}^*_{n,j}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}})| + \max_{j \in \mathcal{J}^*} |c^*_n/(1 + \eta^*_{n,j})|.
\end{aligned} \tag{H.74}$$

We note that when Assumption E.3-2 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 μ_j were defined in (G.11)-(G.12) and (H.10)-(H.11) respectively. Moreover, $\mathbb{G}^*_{n,j}(\theta_n + \frac{\lambda\rho}{\sqrt{n}}) \xrightarrow{a.s.} \mathbb{Z}^*$ and (H.49) implies $c^*_n \to 0$ so that we have

$$\sup_{\lambda\in B^d}\left\| g^{\mathcal{C}} - g^{\mathcal{C}}_n(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}})\right\| \xrightarrow{a.s.} 0. \tag{H.75}$$

Further, by (H.65), $D_{P_n} \to D$ and, Assumption E.4-(ii), for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that

$$\sup_{\lambda\in B^d}\left\| \left(K^{\mathcal{C}}\right)^{-1} - \left(K^{\mathcal{C}}_{P_n}(\bar{\theta}(\theta_n,\lambda),\rho)\right)^{-1}\right\|_2 \leqslant \eta,\ \forall n \geqslant N. \tag{H.76}$$

In sum, by (H.63), (H.67), and (H.74)-(H.76), for any $\eta, \nu > 0$, there exists $N \geqslant \mathbb{N}$ such that

$$\mathbf{P}\left(\sup_{\lambda\in B^d}\|\psi_n(\lambda) - \phi_n(\lambda)\| < \nu\right) \geqslant 1-\eta,\ \forall n \geqslant \mathbb{N}. \tag{H.77}$$

Hence, for a specific choice of $\nu = \kappa(1-\beta)$, where β is defined in equation (H.70), we have that $\sup_{\lambda\in B^d}\|\psi_n(\lambda) - \phi_n(\lambda)\| < \kappa(1-\beta)$ implies

$$\begin{aligned}\|\lambda^{\mathcal{C}}_n - \lambda^f_n\| &= \|\psi_n(\lambda^{\mathcal{C}}_n) - \phi_n(\lambda^f_n)\| \\ &\leqslant \|\psi_n(\lambda^{\mathcal{C}}_n) - \phi_n(\lambda^{\mathcal{C}}_n)\| + \|\phi_n(\lambda^{\mathcal{C}}_n) - \phi_n(\lambda^f_n)\| \\ &\leqslant \kappa(1-\beta) + \beta\|\lambda^{\mathcal{C}}_n - \lambda^f_n\|\end{aligned} \tag{H.78}$$

Rearranging terms, we obtain $\|\lambda^{\mathcal{C}}_n - \lambda^f_n\| \leqslant \kappa$. Note that by Assumptions E.4 (i) and E.5 (i), for any $\delta > 0$, there exists $\kappa_\delta > 0$ and $N \in \mathbb{N}$ such that

$$\mathbf{P}\Big(\sup_{\|\lambda-\lambda'\|\leqslant\kappa_\delta}|u^*_{n,j,\theta_n}(\lambda) - u^*_{n,j,\theta_n}(\lambda')| < \delta\Big) \geqslant 1-\eta,\ \forall n \geqslant \mathbb{N}. \tag{H.79}$$

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

$$\mathfrak{w}^*_j(\lambda^{\mathcal{C}}_n) + \delta \leqslant 0,\ j \in \{1,\cdots,J_1\} \cap \mathcal{J}^*. \tag{H.80}$$

Hence, by (H.39), (H.49), and (H.79)-(H.80), $\|\lambda^{\mathcal{C}}_n - \lambda^f_n\| \leqslant \kappa_{\delta/4}$, for each $j \in \{1,\cdots,J_1\} \cap \mathcal{J}^*$ we have

$$u^*_{n,j,\theta_n}(\lambda^f_n) - c^*_n(\theta_n) \leqslant u^*_{n,j,\theta_n}(\lambda^{\mathcal{C}}_n) - c^*_n(\theta_n) + \delta/4 \leqslant \mathfrak{w}^*_j(\lambda^{\mathcal{C}}_n) + \delta/2 \leqslant 0. \tag{H.81}$$

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

$$\begin{aligned}\mathbf{P}\Big(\{U^*_n(\theta_n, c^*_n) = \varnothing\} \cap \{\mathfrak{W}^{*,-\delta}(0) \neq \varnothing\}\Big) &\leqslant \mathbf{P}\Big(\nexists \lambda^f_n \in U^*_n(\theta_n, c^*_n), \exists \lambda^{\mathcal{C}}_n \in \mathfrak{W}^{*,-\delta}(0)\Big) \\ &\leqslant \mathbf{P}\left(\left\{\sup_{\lambda\in B^d}\|\psi_n(\lambda) - \phi_n(\lambda)\| < \kappa_\delta(1-\beta) \cap A_n\right\}^c\right) \leqslant \eta/3,\end{aligned} \tag{H.82}$$

where A^c denotes the complement of the set A, and the last inequality follows from (H.39) and (H.77). □

Lemma H.3: *Suppose Assumptions E.1, E.2, E.3, E.4, and E.5 hold. Let $\{P_n, \theta_n\} \in \{(P,\theta) : P \in \mathcal{P}, \theta \in \Theta_I(P)\}$ be a sequence satisfying (H.1)-(H.3). For each j, let*

$$v^I_{n,j,\theta_n}(\lambda) \equiv \mathbb{G}^b_{n,j}(\theta_n) + \rho\hat{D}_{n,j}(\theta_n)\lambda + \varphi^*_j(\hat{\xi}_{n,j}(\theta_n)), \tag{H.83}$$

$$\mathfrak{w}_j(\lambda) \equiv \mathbb{Z}_j + \rho D_j\lambda + \pi^*_{1,j}, \tag{H.84}$$

where

$$\varphi_j^*(\xi) = \begin{cases} \varphi_j(\xi) & \pi_{1,j} = 0 \\ -\infty & \pi_{1,j} < 0 \\ 0 & j = J_1 + 1, \cdots, J. \end{cases} \tag{H.85}$$

For each $c \geqslant 0$, define

$$V_n^I(\theta_n, c) \equiv \{\lambda \in B_{n,\rho}^d : p'\lambda = 0 \cap v_{n,j,\theta_n}^I(\lambda) \leqslant c, j = 1, \cdots, J\}, \tag{H.86}$$

$$\mathfrak{W}(c) \equiv \left\{\lambda \in \mathfrak{B}_\rho^d : p'\lambda = 0 \cap \mathfrak{w}_j(\lambda) \leqslant c, \ \forall j = 1, \ldots, J\right\}. \tag{H.87}$$

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,

$$P_n^*(V_n^I(\theta_n', c) \neq \varnothing) - \Pr(\mathfrak{W}(c) \neq \varnothing) \to 0, \tag{H.88}$$

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,

$$\hat{c}_n(\theta_n') \geqslant c_n^I(\theta_n') + o_{P_n}(1). \tag{H.89}$$

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 H.15, in $l^\infty(\Theta)$ uniformly in $\mathcal{P}$ conditional on $\{X_i\}_{i=1}^\infty$, and by Assumption E.4 $\|\hat{D}_n(\theta_n') - D_{P_n}(\theta_n)\| \xrightarrow{p} 0$. Further, by Lemma H.5, $\hat{\xi}_{n,j}(\theta_n') \xrightarrow{P_n} \pi_{1,j}$. Therefore,

$$(\mathbb{G}_n^b(\theta_n'), \hat{D}_n(\theta_n'), \hat{\xi}_n(\theta_n'))|\{X_i\}_{i=1}^\infty \xrightarrow{d} (\mathbb{Z}, D, \pi_1). \tag{H.90}$$

for almost all sample paths $\{X_i\}_{i=1}^\infty$. By Lemma H.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 $(\tilde{\mathbb{G}}_n^b(\theta_n'), \tilde{D}_n, \tilde{\xi}_n) \stackrel{d}{=} (\mathbb{G}_n^b(\theta_n'), \hat{D}_n(\theta_n'), \hat{\xi}_n(\theta_n'))$ 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 $\tilde{\mathbb{Z}} \stackrel{d}{=} \mathbb{Z}$. 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^∞ for notational simplicity (see Appendix H.3 for details). Using this representation, define

$$\tilde{v}_{n,j,\theta_n'}^I(\lambda) \equiv \tilde{\mathbb{G}}_{n,j}^b(\theta_n') + \rho\tilde{D}_n\lambda + \varphi_j^*(\tilde{\xi}_{n,j}), \tag{H.91}$$

and

$$\tilde{\mathfrak{w}}_j(\lambda) \equiv \tilde{\mathbb{Z}}_j + \rho D_j\lambda + \pi_{1,j}^*, \tag{H.92}$$

where $\tilde{\mathbb{Z}} \stackrel{d}{=} \mathbb{Z}$, 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

$$\begin{aligned} |P_n^*(V_n^I(\theta_n', c) \neq \varnothing) - \Pr(\mathfrak{W}(c) \neq \varnothing)| &= |\tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n', c) \neq \varnothing) - \tilde{\mathbf{P}}(\tilde{\mathfrak{W}}(c) \neq \varnothing)| \\ &\leqslant |\tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n', c) = \varnothing \cap \tilde{\mathfrak{W}}(c) \neq \varnothing) + \tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n', c) \neq \varnothing \cap \tilde{\mathfrak{W}}(c) = \varnothing)|, \end{aligned} \tag{H.93}$$

where the inequality is due to (H.33). First, we bound the first term on the right hand side of (H.93). Note that

$$\tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n', c) = \varnothing \cap \tilde{\mathfrak{W}}(c) \neq \varnothing) \leqslant \tilde{\mathbf{P}}(\tilde{V}_n^{I,+\delta}(\theta_n', c) = \varnothing \cap \tilde{\mathfrak{W}}(c) \neq \varnothing) + \tilde{\mathbf{P}}(\tilde{V}_n^{I,+\delta}(\theta_n', c) \neq \varnothing \cap \tilde{V}_n^I(\theta_n', c) = \varnothing), \tag{H.94}$$

where $\tilde{V}_n^{I,+\delta}$ is defined as

$$\tilde{V}_n^{I,+\delta} \equiv \Big\{\lambda \in B_{n,\rho}^d : p'\lambda = 0 \cap \tilde{v}_{n,j,\theta_n'}^I(\lambda) \leqslant c + \delta, j \in \mathcal{J}^*\Big\}. \tag{H.95}$$

Let

$$A_n \equiv \Big\{\tilde{\omega} \in \tilde{\Omega} : \sup_{\lambda \in B^d} \max_{j \in \mathcal{J}^*} |\tilde{v}_{n,j,\theta_n'}^I(\lambda) - \tilde{\mathfrak{w}}_j(\lambda)| \geqslant \delta\Big\}. \tag{H.96}$$

Let

$$E \equiv \{\{x_i\}_{i=1}^{\infty} : \|\hat{D}_n(\theta_n') - D\| < \eta, \max_{j \in \mathcal{J}^*} |\varphi_j^*(\hat{\xi}_{n,j}(\theta_n')) - \pi_{1,j}^*| < \eta\}. \tag{H.97}$$

Note that, $P_n(E) \geqslant 1 - \eta$ for all n sufficiently large by Assumption E.4 and Lemma H.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 $j \in \mathcal{J}^*$,

$$|\tilde{v}_{n,j,\theta_n'}^I(\lambda) - \tilde{\mathfrak{w}}_j(\lambda)| \leqslant |\tilde{\mathbb{G}}_{n,j}^b(\theta_n') - \tilde{\mathbb{Z}}_j| + \rho\|\tilde{D}_{j,n} - D_j\|\|\lambda\| + |\varphi_j^*(\tilde{\xi}_{n,j}) - \pi_{1,j}^*| \leqslant (2 + \rho)\eta, \tag{H.98}$$

uniformly in $\lambda \in B^d$, where we used $\tilde{\mathbb{G}}_n^b \to \tilde{\mathbb{Z}}, \tilde{\mathbf{P}} - a.s.$ Since η 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

$$L_n \equiv \Big\{\tilde{\omega} \in \tilde{\Omega} : \tilde{\mathfrak{W}}(c) \subseteq \tilde{V}_n^{I,+\delta}(\theta_n', c)\Big\}. \tag{H.99}$$

Using this,

$$\tilde{\mathbf{P}}(\tilde{V}_n^{I,+\delta}(\theta_n', c) = \varnothing \cap \tilde{\mathfrak{W}}(c) \neq \varnothing) \leqslant \tilde{\mathbf{P}}(\tilde{\mathfrak{W}}(c) \nsubseteq \tilde{V}_n^{I,+\delta}(\theta_n', c)) = \tilde{\mathbf{P}}(L_n^c) \leqslant \tilde{\mathbf{P}}(A_n) < \eta/2, \tag{H.100}$$

for all n sufficiently large. Also, by Lemma H.6,

$$\tilde{\mathbf{P}}(\tilde{V}_n^{I,+\delta}(\theta_n', c) \neq \varnothing \cap \tilde{V}_n^I(\theta_n', c) = \varnothing) < \eta/2, \tag{H.101}$$

for all n sufficiently large.

Combining (H.94), (H.96), (H.100), (H.101), and using $P_n(E) \geqslant 1 - \eta$ for all n, we have

$$\int_E \tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n', c) = \varnothing \cap \tilde{\mathfrak{W}}(c) \neq \varnothing) dP_n + \int_{E^c} \tilde{\mathbf{P}}(\tilde{V}_n^I(\theta_n', c) = \varnothing \cap \tilde{\mathfrak{W}}(c) \neq \varnothing) dP_n \leqslant \eta(1 - \eta) + \eta \leqslant 2\eta. \tag{H.102}$$

The second term of the right hand side of (H.93) 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

$$P_n^*(V_n^I(\theta_n', c) \neq \varnothing) - \Pr(\mathfrak{W}(c) \neq \varnothing) \to 0. \tag{H.103}$$

Fix $c > 0$, and set

$$g_j = \begin{cases} c - \mathbb{Z}_j, & j = 1, \ldots, J, \\ 1, & j = J+1, \ldots, J+2d, \\ 0, & j = J+2d+1, J+2d+2. \end{cases} \tag{H.104}$$

Mimic the argument following (H.141). Then, this yields

$$|\Pr(\mathfrak{W}(c) \neq \varnothing) - \Pr(\mathfrak{W}(c-\delta) \neq \varnothing)| = \Pr(\{\mathfrak{W}(c) \neq \varnothing\} \cap \{\mathfrak{W}(c-\delta) = \varnothing\}) \leqslant \eta, \tag{H.105}$$

$$|\Pr(\mathfrak{W}(c+\delta) \neq \varnothing) - \Pr(\mathfrak{W}(c) \neq \varnothing)| = \Pr(\{\mathfrak{W}(c+\delta) \neq \varnothing\} \cap \{\mathfrak{W}(c) = \varnothing\}) \leqslant \eta, \tag{H.106}$$

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

$$K = \begin{bmatrix} D^{[1:J^*,:]} \\ I_d \\ -I_d \\ p' \\ -p' \end{bmatrix},\ g = \begin{bmatrix} e^{[1:J^*]} \\ \rho \cdot \mathbf{1}_d \\ \rho \cdot \mathbf{1}_d \\ 0 \\ 0 \end{bmatrix},\ \tau = \begin{bmatrix} \mathbf{1}_{J^*} \\ \mathbf{0}_d \\ \mathbf{0}_d \\ 0 \\ 0 \end{bmatrix}. \tag{H.107}$$

By Farkas' lemma (Rockafellar, 1970, Theorem 22.1) and arguing as in (H.146),

$$\Pr(\{\mathfrak{W}(c) \neq \varnothing\} \cap \{\mathfrak{W}(c-\delta) = \varnothing\}) = \Pr\left(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu'(g - \delta\tau) < 0, \exists \mu \in \mathcal{M}\}\right), \tag{H.108}$$

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

$$\mathcal{M} = \{\mu : \mu = b \sum_{t=1}^{T} a_t \nu^t, b > 0, a_t \geqslant 0, \sum_{t=1}^{T} a_t = 1\}. \tag{H.109}$$

This implies

$$\mu' g \geqslant 0,\ \forall \mu \in \mathcal{M} \ \Leftrightarrow\ \nu^{t\prime} g \geqslant 0,\ \forall t \in \{1, \cdots, T\} \tag{H.110}$$

$$\mu'(g - \delta\tau) < 0,\ \exists \mu \in \mathcal{M} \ \Leftrightarrow\ \nu^{t\prime} g < \delta \nu^{t\prime} \tau,\ \exists t \in \{1, \cdots, T\}. \tag{H.111}$$

Hence,

$$\Pr\left(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu'(g - \delta\tau) < 0, \exists \mu \in \mathcal{M}\}\right) = \Pr\left(0 \leqslant \nu^{s\prime} g,\ 0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau,\ \forall s, \exists t\right) \tag{H.112}$$

Note that by (H.107), for each $s \in \{1, \cdots, T\}$,

$$\nu^{s\prime} g = \nu^{s,[1:J^*]\prime}(c 1_{\mathcal{J}^*} - \mathbb{Z}_{\mathcal{J}^*}) + \rho \sum_{j=J^*+1}^{J^*+2d} \nu^{s,[j]}, \tag{H.113}$$

$$\nu^{s\prime} \tau = \sum_{j=1}^{J^*} \nu^{s,[j]}. \tag{H.114}$$

For each $s \in \{1, \cdots, T\}$, let

$$h_s^U \equiv c \sum_{j=1}^{J^*} \nu^{s,[j]} + \rho \sum_{j=J^*+1}^{J^*+2d} \nu^{s,[j]} \tag{H.115}$$

$$h_s^L \equiv (c-\delta) \sum_{j=1}^{J^*} \nu^{s,[j]}, \tag{H.116}$$

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 (H.112) as

$$\Pr\left(0 \leqslant \nu^{s\prime} g,\ 0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau,\ \forall s, \exists t\right) = \Pr\left(\nu^{s,[1:J^*]\prime} \mathbb{Z}_{\mathcal{J}^*} \leqslant h_s^U, h_t^L < \nu^{t,[1:J^*]\prime} \mathbb{Z}_{\mathcal{J}^*} \leqslant h_t^U\ \forall s, \exists t\right) > 0, \tag{H.117}$$

where the last inequality follows because $\mathbb{Z}_{\mathcal{J}^*}$'s correlation matrix Ω has an eigenvalue bounded away from 0 by Assumption E.3. By (H.108), (H.112), and (H.117), $c \mapsto \Pr(\mathfrak{W}(c) \neq \varnothing)$ 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 (2.13)) 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 E.3-1 holds. In this case, there are no paired inequalities, and V_n^I differs from V_n^b only in terms of the function φ_j^* in (H.85) used in place of the GMS function φ_j. In particular, $\varphi_j^*(\xi) \leqslant \varphi_j(\xi)$ for any j and ξ, and therefore $\hat{c}_n(\theta_n) \geqslant c_n^I(\theta_n)$ by construction.

Next, suppose Assumption E.3-2 holds and $V_n^I(\theta_n', c)$ is defined with hard threshold GMS, i.e. with GMS function φ^1 in AS. The only case that might create concern is one in which

$$\pi_{1,j} \in [-1, 0) \text{ and } \pi_{1,j+R_1} = 0. \tag{H.118}$$

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

$$\pi_{1,j}^* = -\infty, \text{ and } \pi_{1,j+R_1}^* = 0, \tag{H.119}$$

$$\varphi_j(\hat{\xi}_{n,j}(\theta_n')) = 0, \text{ and } \varphi_{j+R_1}(\hat{\xi}_{n,j+R_1}(\theta_n')) = 0, \tag{H.120}$$

so that in $V_n^I(\theta_n', c)$, inequality $j + R_1$, which is

$$\mathbb{G}_{n,j+R_1}^b(\theta_n') + \rho \hat{D}_{n,j+R_1}(\theta_n')\lambda \leqslant c, \tag{H.121}$$

is replaced with inequality

$$-\mathbb{G}_{n,j}^b(\theta_n') - \rho \hat{D}_{n,j}(\theta_n')\lambda \leqslant c, \tag{H.122}$$

as explained in Section E.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 (H.83) 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 (H.86) 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 H.9-(3) establishes

$$\sup_{\lambda \in B_{n,\rho}^d} \|\mathbb{G}_{n,j+R_1}^b(\theta_n') + \rho \hat{D}_{n,j+R_1}(\theta_n')\lambda + \mathbb{G}_{n,j}^b(\theta_n') + \rho \hat{D}_{n,j}(\theta_n')\lambda\| = o_{P^*}(1), \tag{H.123}$$

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

$$c_n^{IP}(\theta_n') \xrightarrow{P_n} c_{\pi*}. \tag{H.124}$$

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 E.2 and continuous in its argument, we can mimic the above argument using the replacements in (H.12)-(H.13) with $\hat{\mu}_{n,j+R_1}$ as defined in (H.14) and $\hat{\mu}_{n,j}(\theta_n')$ as in (H.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 μ_j, μ_{j+R_1} are defined in equations (H.10)-(H.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

$$\Delta(\mu, \hat{\mu}) \leqslant \left|\hat{\mu}_{n,j+R_1}(\theta_n') - \mu_{j+R_1}(\theta_n')\right| \left\|\{\mathbb{G}_{n,j+R_1}^b(\theta_n') + \rho\hat{D}_{n,j+R_1}(\theta_n')\lambda\} + \{\mathbb{G}_{n,j}^b(\theta_n') + \rho\hat{D}_{n,j}(\theta_n')\lambda\}\right\|. \tag{H.125}$$

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 H.9-(3)-(4), and by Assumption E.4-(i). The rest of the argument follows similarly as for the case of hard-threshold GMS. □

LEMMA H.4: *Let Assumptions E.1, E.2, E.4, and E.5 hold. Let (P_n, θ_n) be the sequence satisfying (H.1)-(H.3), let $\mathcal{J}^*$ be defined as in (H.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 $n \geqslant \max\{N', N''\}$,*

$$\mathbf{P}\left(\sup_{\lambda \in B^d}\left|\max_{j=1,\cdots,J}(u_{n,j,\theta_n}^*(\lambda) - c_n^*) - \max_{j=1,\cdots,J}(\mathfrak{w}_j^*(\lambda) - c_{\pi*})\right| \geqslant \varepsilon\right) < \eta, \tag{H.126}$$

$$\tilde{\mathbf{P}}\left(\sup_{\lambda \in B^d}\left|\max_{j=1,\cdots,J}\tilde{\mathfrak{w}}_j(\lambda) - \max_{j=1,\cdots,J}\tilde{v}_{n,j,\theta_n'(\lambda)}^I\right| \geqslant \varepsilon\right) < \eta, \ w.p.1, \tag{H.127}$$

where the functions $u_n^*, \mathfrak{w}^*, \tilde{v}_n, \tilde{\mathfrak{w}}$ are defined in equations (H.24),(H.25), (H.91), and (H.92).

Proof. We first establish (H.126). By definition, $\pi_{1,j}^* = -\infty$ for all $j \notin \mathcal{J}^*$ and therefore

$$\mathbf{P}\Big(\sup_{\lambda \in B^d}|\max_{j=1,\cdots,J}(u_{n,j,\theta_n}^*(\lambda) - c_n^*) - \max_{j=1,\cdots,J}(\mathfrak{w}_j^*(\lambda) - c_{\pi*})| \geqslant \varepsilon\Big) \tag{H.128}$$

$$= \mathbf{P}\Big(\sup_{\lambda \in B^d}|\max_{j \in \mathcal{J}^*}(u_{n,j,\theta_n}^*(\lambda) - c_n^*) - \max_{j \in \mathcal{J}^*}(\mathfrak{w}_j^*(\lambda) - c_{\pi*})| \geqslant \varepsilon\Big). \tag{H.129}$$

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

$$\begin{aligned}|r_{n,j,\theta_n}(\lambda)| &\leqslant |\mathbb{G}^*_{n,j}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) - \mathbb{Z}^*_j| + \rho\|D_{P_n,j}(\bar{\theta}_n) - D_j\|\|\lambda\| + |\mathbb{G}^*_{n,j}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) + \rho D_{P_n,j}(\bar{\theta}_n)\lambda|\ \eta^*_{n,j} + |c^*_n - c_{\pi*}| \\ &= |\mathbb{G}^*_{n,j}(\theta_n + \tfrac{\lambda\rho}{\sqrt{n}}) - \mathbb{Z}^*_j| + o(1) + \{O_{\mathcal{P}}(1) + O(1)\})\eta^*_{n,j} + o_{\mathcal{P}}(1) \\ &= o_{\mathcal{P}}(1) \end{aligned} \tag{H.130}$$

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 E.4-(ii), and $\bar{\theta}_n$ being a mean value between θ_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 E.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 H.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 μ_j were defined in (G.11)-(G.12) and (H.10)-(H.11) respectively.

By (H.130) and the fact that $j \in \mathcal{J}^*$, we have

$$\sup_{\lambda\in B^d} |\max_{j\in\mathcal{J}^*}(u^*_{n,j,\theta_n}(\lambda) - c^*_n) - \max_{j\in\mathcal{J}^*}(\mathfrak{w}^*_j(\lambda) - c_{\pi*})| \leqslant \sup_{\lambda\in B^d}\max_{j\in\mathcal{J}^*}|r_{n,j,\theta_n}(\lambda)| = o_{\mathcal{P}}(1). \tag{H.131}$$

The conclusion of the lemma then follows from (H.129) and (H.131).

The result in (H.127) follows from similar arguments. □

LEMMA H.5: *Let Assumptions E.1, E.2, E.4, and E.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*

$$\chi_j(\{Q_n,\vartheta_n\}) \equiv \begin{cases} 0, & \text{if } \lim_{n\to\infty}\kappa_n^{-1}\sqrt{n}\gamma_{1,Q_n,j}(\vartheta_n) = 0, \\ -\infty, & \text{if } \lim_{n\to\infty}\kappa_n^{-1}\sqrt{n}\gamma_{1,Q_n,j}(\vartheta_n) < 0. \end{cases} \tag{H.132}$$

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

$$\begin{aligned}\sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P),\theta'\in\theta+\rho/\sqrt{n}B^d}&\left|\frac{\sqrt{n}E_P(m_j(X,\theta))}{\kappa_n\sigma_{P,j}(\theta)} - \frac{\sqrt{n}E_P(m_j(X,\theta'))}{\kappa_n\sigma_{P,j}(\theta')}\right| \\ &\leqslant \sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P),\theta'\in\theta+\rho/\sqrt{n}B^d}\frac{\sqrt{n}\|D_{P,j}(\tilde{\theta})\|\|\theta'-\theta\|}{\kappa_n} = o(1), \end{aligned} \tag{H.133}$$

where $\tilde{\theta}$ represents a mean value that lies componentwise between θ and θ' and where we used the fact that $D_{P,j}(\theta)$ is Lipschitz continuous and $\sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P)}\|D_{P,j}(\theta)\| \leqslant \bar{M}$. Result (ii) then follows immediately from (H.132).

For (iii), note that

$$\begin{aligned}
&\sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d}\left|\kappa_n^{-1}\frac{\sqrt{n}\bar{m}_{n,j}(\theta_n')}{\hat{\sigma}_{n,j}(\theta_n')}-\kappa_n^{-1}\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')}\right| \\
&\quad\leqslant \sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d}\left|\kappa_n^{-1}\frac{\sqrt{n}(\bar{m}_{n,j}(\theta_n')-E_{P_n}[m_j(X_i,\theta_n')])}{\sigma_{n,j}(\theta_n')}(1+\eta_{n,j}(\theta_n'))+\kappa_n^{-1}\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')}\eta_{n,j}(\theta_n')\right| \\
&\quad\leqslant \sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d}|\kappa_n^{-1}\mathbb{G}_n(\theta_n')(1+\eta_{n,j}(\theta_n'))|+\left|\frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\kappa_n\sigma_{P_n,j}(\theta_n')}\eta_{n,j}(\theta_n')\right|=o_{\mathcal{P}}(1), \quad \text{(H.134)}
\end{aligned}$$

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 (2015b), 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 H.10-(i). □

LEMMA H.6: *Let Assumptions E.1, E.2, E.3, E.4, and E.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

$$\sup_{c\geqslant 0}\Pr(\{\mathfrak{W}(c)\neq\varnothing\}\cap\{\mathfrak{W}^{-\delta}(c)=\varnothing\})<\eta. \quad \text{(H.135)}$$

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

$$\sup_{c\geqslant 0}P_n^*(\{V_n^I(\theta_n',c)\neq\varnothing\}\cap\{V_n^{I,-\delta}(\theta_n',c)=\varnothing\})<\eta,\ \forall n\geqslant N. \quad \text{(H.136)}$$

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

$$\mathfrak{W}^{-\delta}(c)\equiv\left\{\lambda\in\mathfrak{B}_\rho^d : p'\lambda=0\cap\mathfrak{w}_j(\lambda)\leqslant c-\delta,\ \forall j=1,\ldots,J\right\}, \quad \text{(H.137)}$$

and

$$V_n^{I,-\delta}(\theta_n',c)\equiv\left\{\lambda\in B_{n,\rho}^d : p'\lambda=0\cap v_{n,j,\theta_n'}^I(\lambda)\leqslant c-\delta,\ \forall j=1,\ldots,J\right\}. \quad \text{(H.138)}$$

Then for any $\eta>0$, there exists $\delta>0$ such that

$$\sup_{c\geqslant\underline{c}}\Pr(\{\mathfrak{W}(c)\neq\varnothing\}\cap\{\mathfrak{W}^{-\delta}(c)=\varnothing\})<\eta. \quad \text{(H.139)}$$

with $\mathfrak{W}^{-\delta}(c)$ defined in (H.137). Moreover, for any $\eta>0$, there exist $\delta>0$ and $N\in\mathbb{N}$ such that

$$\sup_{c\geqslant\underline{c}}P_n^*(\{V_n^I(\theta_n',c)\neq\varnothing\}\cap\{V_n^{I,-\delta}(\theta_n',c)=\varnothing\})<\eta,\ \forall n\geqslant N, \quad \text{(H.140)}$$

with $V_n^{-\delta}(\theta_n',c)$ defined in (H.138).

Proof. We first show (H.135). If $\mathcal{J}^*=\varnothing$, with $\mathcal{J}^*$ as defined in (H.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 (G.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 (H.16) and (H.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

$$g_j = \begin{cases} c - \mathbb{Z}_j, & j = 1, \ldots, J, \\ 1, & j = J+1, \ldots, J+2d, \\ 0, & j = J+2d+1, J+2d+2, \end{cases} \tag{H.141}$$

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

$$\tau_j = \begin{cases} 1, & j = 1, \ldots, J_1, \\ 0, & j = J_1 + 1, \ldots, J + 2d + 2. \end{cases} \tag{H.142}$$

Then we can express the sets of interest as

$$\mathfrak{W}(c) = \{\lambda : K\lambda \leqslant g\}, \tag{H.143}$$

$$\mathfrak{W}^{-\delta}(c) = \{\lambda : K\lambda \leqslant g - \delta\tau\}. \tag{H.144}$$

By Farkas' Lemma, e.g. Rockafellar (1970, Theorem 22.1), a solution to the system of linear inequalities in (H.143) 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 (H.144) 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

$$\mathcal{M} \equiv \{\mu \in \mathbb{R}_+^{J+2d+2} : \mu' K = 0\}. \tag{H.145}$$

Then, one may write

$$\begin{aligned} & \Pr(\{\mathfrak{W}(c) \neq \varnothing\} \cap \{W^{-\delta}(\theta_n', c) = \varnothing\}) \\ = \; & \Pr(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu'(g - \delta\tau) < 0, \exists \mu \in \mathcal{M}\}) \\ = \; & \Pr(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu' g < \delta \mu' \tau, \exists \mu \in \mathcal{M}\}). \end{aligned} \tag{H.146}$$

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

$$\mu = b \sum_{t=1}^{T} a_t \nu^t, \tag{H.147}$$

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 ν^t, we have

$$\sum_{t=1}^{T} a_t \nu^{t\prime} g < \delta \sum_{t=1}^{T} a_t \nu^{t\prime} \tau. \tag{H.148}$$

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,

$$\Pr\left(\{\mu' g \geqslant 0, \forall \mu \in \mathcal{M}\} \cap \{\mu' g < \delta \mu' \tau, \exists \mu \in \mathcal{M}\}\right)$$
$$\leqslant \Pr\left(0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau, \exists t \in \{1, \cdots, T\}\right) \leqslant \sum_{t=1}^{T} \Pr\left(0 \leqslant \nu^{t\prime} g < \delta \nu^{t\prime} \tau\right). \quad \text{(H.149)}$$

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

$$\nu^{t\prime} g = \sum_{j=J+1}^{J+2d} \nu_j^t,$$
$$\delta \nu^{t\prime} \tau = \delta \sum_{j=J+1}^{J+2d+2} \nu_j^t \tau_j = 0,$$

where the last equality follows by (H.142). 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 ν^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:

$$g = \begin{cases} c - \mathbb{Z}_j & j = J_1 + 1, \ldots, J_1 + J_2, \\ c + \mathbb{Z}_{j - J_2} & j = J_1 + J_2 + 1, \ldots, J. \end{cases} \quad \text{(H.150)}$$

For each ν^t, (H.150) implies

$$\sum_{j=J_1+1}^{J_1+2J_2} \nu_j^t g_j = c \sum_{j=J_1+1}^{J_1+2J_2} \nu_j^t + \sum_{j=J_1+1}^{J_1+J_2} (\nu_j^t - \nu_{j+J_2}^t) \mathbb{Z}_j. \quad \text{(H.151)}$$

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

$$\tilde{\nu}_j^t \equiv \begin{cases} \nu_j^t & j = 1, \cdots, J_1 \\ \nu_j^t - \nu_{j+J_2}^t & j = J_1 + 1, \cdots, J_1 + J_2. \end{cases}. \quad \text{(H.152)}$$

We then let $\tilde{\nu}^t \equiv (\tilde{\nu}_{n,1}^t, \cdots, \tilde{\nu}_{n,J_1+J_2}^t)'$ and have

$$\nu^{t\prime} g = \sum_{j=1}^{J_1+J_2} \tilde{\nu}_j^t \mathbb{Z}_j + c \sum_{j=1}^{J} \nu_j^t + \sum_{j=J+1}^{J+2d} \nu_j^t. \quad \text{(H.153)}$$

Case 2-a. Suppose $\tilde{\nu}^t \neq 0$. Then, by (H.153), $\frac{\nu^{t\prime} g}{\nu^{t\prime} \tau}$ is a normal random variable with variance $(\tilde{\nu}^{t\prime} \tau)^{-2} \tilde{\nu}'^t \Omega \tilde{\nu}^t$. By Assumption E.3, there exists a constant $\omega > 0$ such that the smallest eigenvalue of Ω is bounded from below by ω for all θ_n'. Hence, letting $\|\cdot\|_p$ denote the p-norm in $\mathbb{R}^{J+2d+2}$, we have

$$\frac{\tilde{\nu}'^t \Omega \tilde{\nu}^t}{(\tilde{\nu}^{t\prime} \tau)^2} \geqslant \frac{\omega \|\tilde{\nu}^t\|_2^2}{(J+2d+2)^2 \|\tilde{\nu}^t\|_2^2} \geqslant \frac{\omega}{(J+2d+2)^2}. \quad \text{(H.154)}$$

Therefore, the variance of the normal random variable in (H.149) 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 ν^t assigns positive weight also to constraints in $\{1, \ldots, J\}$, this must be because $\nu_j^t = 0$ for all $j = 1, \cdots, J_1$ and $\nu_j^t = \nu_{j+J_2}^t$ for all $j = J_1 +$

$1, \cdots, J_1 + J_2$, while $\nu_j^t \neq 0$ for some $j = J_1 + 1, \cdots, J_1 + J_2$. Then we have $\sum_{j=1}^J \nu_j^t g \geqslant 0$, 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 ν^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 (H.135).

To see why equation (H.136) 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 (H.58) 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 (H.145) 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 (2015b) together with Lemma H.17 (through an argument similar to that following equation (H.90)), $\mathbb{G}_n^b \xrightarrow{d} \mathbb{G}_P$ in $l^\infty(\Theta)$ uniformly in $\mathcal{P}$ conditional on $\{X_1, \cdots, X_n\}$, and by Assumption E.4 $\hat{D}_n(\theta_n') \xrightarrow{P_3} D$, for almost all sample paths. Set

$$g_{P_n,j}(\theta_n') = \begin{cases} c - \varphi_j^*(\xi_{n,j}(\theta_n')) - \mathbb{G}_{n,j}^b(\theta_n'), & j = 1, \ldots, J, \\ 1, & j = J+1, \ldots, J+2d, \\ 0, & j = J+2d+1, J+2d+2, \end{cases} \tag{H.155}$$

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 (H.141) to conclude (H.136).

The results in (H.139)-(H.140) follow by similar arguments, with proper redefinition of τ in equation (H.142). □

Lemma H.7: *Let Assumptions E.3 and E.5 hold. Let (P_n, θ_n) have the almost sure representations given in Lemma H.1, let $\mathcal{J}^*$ be defined as in (H.29), and assume that $\mathcal{J}^* \neq \varnothing$. Let $\tilde{\mathcal{C}}$ 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 $\tilde{\mathcal{C}}$ 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 H.2. Then, for any $\eta > 0$, there exist $0 < \varepsilon_\eta < \infty$ and $N \in \mathbb{N}$ s.t. $n \geqslant N$ implies*

$$\mathbf{P}\left(\mathfrak{W}^{*,-\delta}(0) \neq \varnothing, \left|\det K^{\mathcal{C}}\right| \leqslant \varepsilon_\eta\right) \leqslant \eta. \tag{H.156}$$

Proof. We bound the probability in (H.156) as follows:

$$\begin{aligned} \mathbf{P}\left(\mathfrak{W}^{*,-\delta}(0) \neq \varnothing, \left|\det K^{\mathcal{C}}\right| \leqslant \varepsilon_\eta\right) &\leqslant \mathbf{P}\left(\exists C \in \tilde{\mathcal{C}} : \lambda^C \in B^d, \left|\det K^C\right| \leqslant \varepsilon_\eta\right) && \text{(H.157)} \\ &\leqslant \sum_{C \in \tilde{\mathcal{C}} : \left|\det K^C\right| \leqslant \varepsilon_\eta} \mathbf{P}\left(\lambda^C \in B^d\right) && \text{(H.158)} \\ &\leqslant \sum_{C \in \tilde{\mathcal{C}} : |\alpha^C| \leqslant \varepsilon_\eta^{2/d}} \mathbf{P}\left(\lambda^C \in B^d\right), && \text{(H.159)} \end{aligned}$$

where α^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 $\tilde{\mathcal{C}}$ 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 $C \in \tilde{\mathcal{C}}$. Let q^C denote the eigenvector associated with α^C and recall that because $K^C K^{C\prime}$ is symmetric,

$\|q^C\| = 1$. Thus the claim is equivalent to:

$$|q^{C\prime}K^C K^{C\prime}q^C| \leqslant \varepsilon_\eta^{2/d} \implies \mathbf{P}((K^C)^{-1}g^C \in \mathfrak{B}_\rho^d) \leqslant \overline{\eta}. \tag{H.160}$$

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

$$\left|q^{C\prime}g_{P_n}^C\right| = \left|q^{C\prime}K^C\left(K^C\right)^{-1}g^C\right| < \sqrt{d}\varepsilon_\eta^{1/d}, \tag{H.161}$$

hence

$$\mathbf{P}((K^C)^{-1}g^C \in \mathfrak{B}_\rho^d) \leqslant \mathbf{P}\Big(|q^{C\prime}g^C| < \sqrt{d}\varepsilon_\eta^{1/d}\Big). \tag{H.162}$$

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 E.3 and E.5 (iii) yield

$$\begin{aligned} &\operatorname{eig}(\tilde{\Omega}) \geqslant \omega \\ \implies &Var_{\mathbf{P}}(q^{C\prime}g^C) \geqslant \omega \\ \implies &\mathbf{P}\left(|q^{C\prime}g^C| < \sqrt{d}\varepsilon_\eta^{1/d}\right) = \mathbf{P}\left(-\sqrt{d}\varepsilon_\eta^{1/d} < q^{C\prime}g^C < \sqrt{d}\varepsilon_\eta^{1/d}\right) \\ &< \frac{2\sqrt{d}\varepsilon_\eta^{1/d}}{\sqrt{2\omega\pi}}, \end{aligned} \tag{H.163}$$

where the result in (H.163) 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 H.8: *Let Assumptions E.1, E.2, E.3, E.4, and E.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 θ. 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 (H.164) to (H.165) below.) Then the coverage

probability $P^*(V_n^I(\theta,c) \neq \varnothing)$ is asymptotically bounded above by

$$P^*\Big(\sup_{\lambda\in\rho B^d_{n,\rho}} \Big\{p'\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j \in \mathcal{J}^*\Big\} \geqslant 0\Big) \leqslant P^*\Big(\sup_{\lambda\in\mathbb{R}^d} \Big\{p'\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j \in \bar{C}\Big\} \geqslant 0\Big) \tag{H.164}$$

$$= P^*\Big(p'(\hat{D}^{\bar{C}}_n)^{-1}(c\mathbf{1}_d - \mathbb{G}^{b,\bar{C}}_n) \geqslant 0\Big) \tag{H.165}$$

$$= P^*\left(\frac{p'(\hat{D}^{\bar{C}}_n)^{-1}(c\mathbf{1}_d - \mathbb{G}^{b,\bar{C}}_n)}{\sqrt{p'(\hat{D}^{\bar{C}}_n)^{-1}\Omega^C_P(\hat{D}^{\bar{C}}_n)^{-1}p}} \geqslant 0\right) \tag{H.166}$$

$$= P^*\left(\frac{p'adj(\hat{D}^{\bar{C}}_n)(c\mathbf{1}_d - \mathbb{G}^{b,\bar{C}}_n)}{\sqrt{p'(adj(\hat{D}^{\bar{C}}_n)\Omega^C_P adj(\hat{D}^{\bar{C}}_n)p}} \geqslant 0\right) \tag{H.167}$$

$$= \Phi\left(\frac{p'adj(\hat{D}^{\bar{C}}_n)c\mathbf{1}_d}{\sqrt{p'(adj(\hat{D}^{\bar{C}}_n)\Omega^C_P adj(\hat{D}^{\bar{C}}_n)p}}\right) + o_{\mathcal{P}}(1) \tag{H.168}$$

$$\leqslant \Phi(d\omega^{-1/2}c) + o_{\mathcal{P}}(1). \tag{H.169}$$

Here, (H.164) removes constraints and hence enlarges the feasible set; (H.165) solves in closed form; (H.166) divides through by a positive scalar; (H.167) 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; (H.168) follows by Assumption E.5, using that the term multiplying $\mathbb{G}^{b,\bar{C}}_n$ is $O_{\mathcal{P}}(1)$; and (H.169) uses that by Assumption E.3, there exists a constant $\omega > 0$ that does not depend on θ such that the smallest eigenvalue of Ω_P is bounded from below by ω. The result follows for any choice of $\underline{c} \in (0, \Phi^{-1}(1-\alpha) \times \omega^{1/2}/d)$.

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

$$P^*\Big(\sup_{\lambda\in\rho B^d_{n,\rho}} \Big\{p'\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j \in \bar{C} \cup \{J_1+J_2+1\}\Big\} \geqslant 0\Big) \tag{H.170}$$

$$\leqslant P^*\Big(\exists\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j \in \bar{C} \cup \{J_1+J_2+1\}\Big) \tag{H.171}$$

$$\leqslant P^*\Big(\sup_{\lambda\in\rho B^d_{n,\rho}} \Big\{\hat{D}_{n,J_1+1}\lambda : \hat{D}_{n,j}\lambda \leqslant c - \mathbb{G}^b_{n,j}, j \in \bar{C}\Big\} \geqslant \inf_{\lambda\in\rho B^d_{n,\rho}} \Big\{\hat{D}_{n,J_1+1}\lambda : \hat{D}_{n,J_1+J_2+1}\lambda \leqslant c - \mathbb{G}^b_{n,J_1+J_2+1}\Big\}\Big) \tag{H.172}$$

$$= P^*\Big(\hat{D}_{n,J_1+1}\hat{D}^{\bar{C}\prime}_n(\hat{D}^{\bar{C}}_n\hat{D}^{\bar{C}\prime}_n)^{-1}(c\mathbf{1}_{\bar{d}} - \mathbb{G}^{b,\bar{C}}_n) \geqslant -c + \mathbb{G}^b_{n,J_1+J_2+1}\Big). \tag{H.173}$$

Here, the reasoning from (H.170) to (H.172) holds because we evaluate the probability of increasingly larger events; in particular, if the event in (H.172) 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 (H.173) resembles (H.166), and the argument can be concluded in analogy to (H.167)-(H.169). □

LEMMA H.9: *Let Assumptions E.1, E.2, E.3-2, E.4, and E.5 hold. Suppose that both $\pi_{1,j}$ and $\pi_{1,j+R_1}$ are finite, with $\pi_{1,j}$, $j = 1, \ldots, J$, defined in (G.4). Let (P_n, θ_n) be the sequence satisfying the conditions of Lemma H.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 H.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.

$$\lim_{n\to\infty} \frac{\kappa_n}{\sqrt{n}} \sigma_{P_n,j}(\theta_n') > 0, \tag{H.174}$$

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

$$\frac{\sigma^2_{P_n,j+R_1}(\theta_n')}{\sigma^2_{P_n,j}(\theta_n')} = \frac{\sigma^2_{P_n,j}(\theta_n') + Var_{P_n}(t_j(X_i,\theta_n')) + 2Cov_{P_n}(m_j(X_i,\theta_n'), t_j(X_i,\theta_n'))}{\sigma^2_{P_n,j}(\theta_n')} \to 1, \tag{H.175}$$

where the convergence follows because $Var_{P_n}(t_j(X_i,\theta_n'))$ is bounded due to Assumption E.3-2,

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

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

Case 2.

$$\lim_{n\to\infty} \frac{\kappa_n}{\sqrt{n}} \sigma_{P_n,j}(\theta_n') = 0. \tag{H.176}$$

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 (H.175), it also follows that

$$\lim_{n\to\infty} \frac{\kappa_n}{\sqrt{n}} \sigma_{P_n,j+R_1}(\theta_n') = 0, \tag{H.177}$$

and hence that $E_{P_n}(t_j(X_i,\theta_n')) \to 0$. We then have, using Assumption E.3-2 again,

$$\begin{aligned} Var_{P_n}(t_j(X_i,\theta_n')) &= \int t_j(x,\theta_n')^2 dP_n(x) - E_{P_n}[t_j(X_i,\theta_n')]^2 \\ &\leqslant M \int t_j(x,\theta_n') dP_n(x) - E_{P_n}[t_j(X_i,\theta_n')]^2 \to 0. \end{aligned} \tag{H.178}$$

Hence,

$$\begin{aligned} \frac{\sigma^2_{P_n,j+R_1}(\theta_n')}{\sigma^2_{P_n,j}(\theta_n')} &= \frac{\sigma^2_{P_n,j}(\theta_n') + Var_{P_n}(t_j(X_i,\theta_n')) + 2Cov_{P_n}(m_j(X_i,\theta_n'), t_j(X_i,\theta_n'))}{\sigma^2_{P_n,j}(\theta_n')} \\ &\leqslant \frac{\sigma^2_{P_n,j}(\theta_n') + Var_{P_n}(t_j(X_i,\theta_n'))}{\sigma^2_{P_n,j}(\theta_n')} + \frac{2(Var_{P_n}(t_j(X_i,\theta_n')))^{1/2}}{\sigma_{P_n,j}(\theta_n')} \\ &\to 1, \end{aligned} \tag{H.179}$$

and the first claim follows.

To obtain claim 2, note that

$$Corr_{P_n}(m_j(X_i,\theta_n'),m_{j+R_1}(X_i,\theta_n')) = \frac{-\sigma^2_{P_n,j}(\theta_n') - Cov_{P_n}(m_j(X_i,\theta_n'),t_j(X_i,\theta_n'))}{\sigma_{P_n,j}(\theta_n')\sigma_{P_n,j+R_1}(\theta_n')}$$
$$\to -1, \tag{H.180}$$

where the result follows from (H.178) and (H.179).

To establish Claim 3, consider $\mathbb{G}_n$ below. Note that, for $j = 1, \cdots, R_1$,

$$\begin{bmatrix} \mathbb{G}_{n,j}(\theta_n') \\ \mathbb{G}_{n,j+R_1}(\theta_n') \end{bmatrix} = \begin{bmatrix} \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n (m_j(X_i,\theta_n') - E_{P_n}[m_j(X_i,\theta_n')])}{\sigma_{P_n,j}(\theta_n')} \\ -\frac{\sum_{i=1}^n (m_j(X_i,\theta_n') - E_{P_n}[m_j(X_i,\theta_n')]) + \frac{1}{\sqrt{n}}\sum_{i=1}^n (t_j(X_i,\theta_n') - E_{P_n}[t_j(X_i,\theta_n')])}{\sigma_{P_n,j+R_1}(\theta_n')} \end{bmatrix}. \tag{H.181}$$

Under the conditions of Case 1 above, we immediately obtain

$$|\mathbb{G}_{n,j}(\theta_n') + \mathbb{G}_{n,j+R_1}(\theta_n')| \xrightarrow{P_n} 0. \tag{H.182}$$

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, (H.182) obtains again. These results imply that $\mathbb{Z}_j + \mathbb{Z}_{j+R_1} = 0, a.s.$ By Lemma H.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 (H.182) 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

$$E_{P_n}\left(\frac{m_j(X,\theta_n')}{\sigma_{P_n,j}(\theta_n')} + \frac{m_{j+R_1}(X,\theta_n')}{\sigma_{P_n,j+R_1}(\theta_n')}\right) = O_{\mathcal{P}}\left(\frac{\kappa_n}{\sqrt{n}}\right). \tag{H.183}$$

Define the $1 \times d$ vector

$$q_n \equiv D_{P_n,j+R_1}(\theta_n') + D_{P_n,j}(\theta_n'). \tag{H.184}$$

Suppose by contradiction that

$$\rho q_n \to \varsigma \neq 0,$$

where $\|\varsigma\|$ might be infinite. Write

$$\tilde{r}_n = \frac{q_n'}{\|q_n\|}. \tag{H.185}$$

Let

$$r_n = \tilde{r}_n \rho \kappa_n^2/\sqrt{n}. \tag{H.186}$$

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

lie component-wise between θ_n' and $\theta_n' + r_n$, we obtain

$$\begin{aligned}
&E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+R_1}(X,\theta_n'+r_n)}{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}\right)\\
&= E_{P_n}\left(\frac{m_j(X,\theta_n')}{\sigma_{P_n,j}(\theta_n')}+\frac{m_{j+R_1}(X,\theta_n')}{\sigma_{P_n,j+R_1}(\theta_n')}\right)+\big(D_{P_n,j}(\bar\theta_n)+D_{P_n,j+R_1}(\tilde\theta_n)\big)r_n\\
&= O_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}})+\big(D_{P_n,j}(\theta_n')+D_{P_n,j+R_1}(\theta_n')\big)\,r_n+\big(D_{P_n,j}(\bar\theta_n)-D_{P_n,j}(\theta_n')\big)\,r_n+\big(D_{P_n,j+R_1}(\tilde\theta_n)-D_{P_n,j+R_1}(\theta_n')\big)r_n\\
&= O_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}})+\frac{\rho\kappa_n^2}{\sqrt{n}}+O_{\mathcal{P}}(\frac{\rho^2\kappa_n^4}{n}).
\end{aligned}\tag{H.187}$$

It then follows that there exists $N\in\mathbb{N}$ such that for all $n\geqslant N$, the right hand side in (H.187) is strictly greater than zero.

Next, observe that

$$\begin{aligned}
&E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+R_1}(X,\theta_n'+r_n)}{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}\right)\\
&= E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+R_1}(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}\right)-\left(\frac{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}-1\right)\frac{E_{P_n}(m_{j+R_1}(X,\theta_n'+r_n))}{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}\\
&= E_{P_n}\left(\frac{m_j(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}+\frac{m_{j+R_1}(X,\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}\right)-o_{\mathcal{P}}(\frac{\rho\kappa_n^2}{\sqrt{n}}).
\end{aligned}\tag{H.188}$$

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 E.3-2, we have

$$\frac{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}{\sigma_{P_n,j}(\theta_n'+r_n)}-1=\frac{\sigma_{P_n,j+R_1}(\theta_n')}{\sigma_{P_n,j}(\theta_n')}-1+o_{\mathcal{P}}(1)=o_{\mathcal{P}}(1),\tag{H.189}$$

where we used Claim 1. Second, using Assumption E.4, we have that

$$\frac{E_{P_n}(m_{j+R_1}(X,\theta_n'+r_n))}{\sigma_{P_n,j+R_1}(\theta_n'+r_n)}=\frac{E_{P_n}(m_{j+R_1}(X,\theta_n'))}{\sigma_{P_n,j+R_1}(\theta_n')}+D_{P_n,j+R_1}(\tilde\theta_n)r_n=O_{\mathcal{P}}(\frac{\kappa_n}{\sqrt{n}})+O_{\mathcal{P}}(\frac{\rho\kappa_n^2}{\sqrt{n}}).\tag{H.190}$$

The product of (H.189) and (H.190) is therefore $o_{\mathcal{P}}(\frac{\rho\kappa_n^2}{\sqrt{n}})$ and (H.188) 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 E.3-2 applies. Therefore, there exists $N\in\mathbb{N}$ such that for all $n\geqslant N$, the left hand side in (H.187) 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 H.10: *Suppose Assumptions E.1, E.2, and E.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 $j=1,\ldots,J_1+J_2$*

$$\inf_{P\in\mathcal{P}}P\Big(\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)|\to 0\Big)=1.\tag{H.191}$$

(ii) For any $j=1,\ldots,R_1$ let

$$\hat\sigma_{n,j}^M(\theta)=\hat\sigma_{n,j+R_1}^M(\theta)\equiv\hat\mu_{n,j}(\theta)\hat\sigma_{n,j}(\theta)+(1-\hat\mu_{n,j}(\theta))\hat\sigma_{n,j+R_1}(\theta).\tag{H.192}$$

Let (P_n,θ_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 (H.29). *Then, for any* $\eta > 0$, *there exists* $N \in \mathbb{N}$ *such that*

$$P_n\Big(\max_{j \in (\mathcal{R}_1 \cup \mathcal{R}_2) \cap \mathcal{J}^*} \Big| \frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1 \Big| > \eta \Big) < \eta \tag{H.193}$$

for all $n \geqslant N$.

Proof. We first show that, for any $\epsilon > 0$ and for any $j = 1, \ldots, J_1 + J_2$,

$$\inf_{P \in \mathcal{P}} P\Big(\sup_{m \geqslant n} \sup_{\theta \in \Theta} \Big| \frac{\hat{\sigma}_{n,j}(\theta)}{\sigma_{P,j}(\theta)} - 1 \Big| \leqslant \epsilon \Big) \to 1. \tag{H.194}$$

For this, define the following sets:

$$\mathcal{M}_j \equiv \{ m_j(\cdot, \theta)/\sigma_{P,j}(\theta) : \theta \in \Theta, P \in \mathcal{P} \} \tag{H.195}$$

$$\mathcal{S}_j \equiv \{ (m_j(\cdot, \theta)/\sigma_{P,j}(\theta))^2 : \theta \in \Theta, P \in \mathcal{P} \}. \tag{H.196}$$

By Assumptions E.1-(a), E.1 (iv), E.5 (i), (iii), and arguing as in the proof of Lemma D.2.2 (and D.2.1) in Bugni, Canay, and Shi (2015b), 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 $\epsilon > 0$,

$$\inf_{P \in \mathcal{P}} P\Big(\sup_{m \geqslant n} \sup_{\theta \in \Theta} \Big| \frac{n^{-1} \sum_{i=1}^n m_j(X_i, \theta)^2}{\sigma^2_{P,j}(\theta)} - \frac{E_P[m_j(X, \theta)^2]}{\sigma^2_{P,j}(\theta)} \Big| \leqslant \epsilon \Big) \to 1 \tag{H.197}$$

$$\inf_{P \in \mathcal{P}} P\Big(\sup_{m \geqslant n} \sup_{\theta \in \Theta} \Big| \frac{\bar{m}_{n,j}(\theta) - E_P[m_j(X, \theta)]}{\sigma_{P,j}(\theta)} \Big| \leqslant \epsilon \Big) \to 1. \tag{H.198}$$

Note that, by Assumption E.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 (H.198), for any $\epsilon > 0$, it follows that

$$\inf_{P \in \mathcal{P}} P\Big(\sup_{m \geqslant n} \sup_{\theta \in \Theta} \Big| \frac{\bar{m}_{n,j}(\theta)^2 - E_P[m_j(X, \theta)]^2}{\sigma^2_{P,j}(\theta)} \Big| \leqslant \epsilon \Big) \to 1. \tag{H.199}$$

By the uniform continuity of $x \mapsto \sqrt{x}$ on $\mathbb{R}_+$, for any $\epsilon > 0$, there is a constant $\eta > 0$ such that

$$\Big| \frac{\hat{\sigma}^2_{n,j}(\theta)}{\sigma^2_{P,j}(\theta)} - 1 \Big| \leqslant \eta \;\Rightarrow \Big| \frac{\hat{\sigma}_{n,j}(\theta)}{\sigma_{P,j}(\theta)} - 1 \Big| \leqslant \epsilon. \tag{H.200}$$

By the definition of $\sigma^2_{P,j}(\theta)$ and the triangle inequality,

$$\Big| \frac{\hat{\sigma}^2_{n,j}(\theta)}{\sigma^2_{P,j}(\theta)} - 1 \Big| \leqslant \Big| \frac{n^{-1} \sum_{i=1}^n m(X_i, \theta)^2 - E[m_j(X_i, \theta)^2]}{\sigma^2_{P,j}(\theta)} \Big| + \Big| \frac{\bar{m}_{n,j}(\theta)^2 - E[m_j(X_i, \theta)]^2}{\sigma^2_{P,j}(\theta)} \Big|. \tag{H.201}$$

By (H.200)-(H.201), bounding each of the terms on the right hand side of (H.201) by $\eta/2$ implies $|\hat{\sigma}_{n,j}(\theta)/\sigma_{P,j}(\theta) - 1| \leqslant \epsilon$. This, together with (H.197) and (H.199), ensures that, for any $\epsilon > 0$, (H.194) 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 (2015b) to conclude that

$$\inf_{P \in \mathcal{P}} P\Big(\sup_{m \geqslant n} \sup_{\theta \in \Theta} \Big| \frac{\sigma_{P,j}(\theta)}{\hat{\sigma}_{n,j}(\theta)} - 1 \Big| \leqslant \epsilon \Big) \to 1. \tag{H.202}$$

Finally, recall that $\eta_{n,j}(\theta) = \sigma_{P,j}(\theta)/\hat{\sigma}_{n,j}(\theta) - 1$ and note that for any $\epsilon > 0$,

$$\begin{aligned}
1 &= \lim_{n\to\infty} \inf_{P\in\mathcal{P}} P\Big(\sup_{m\geqslant n} \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant \epsilon \Big) \\
&\leqslant \inf_{P\in\mathcal{P}} \lim_{n\to\infty} P\Big(\bigcap_{m\geqslant n} \big\{ \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant \epsilon \big\} \Big) \\
&= \inf_{P\in\mathcal{P}} P\Big(\lim_{n\to\infty} \bigcap_{m\geqslant n} \{ \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant \epsilon \} \Big) \\
&= \inf_{P\in\mathcal{P}} P\Big(\sup_{\theta\in\Theta} |\eta_{n,j}(\theta)| \leqslant \epsilon, \text{ for almost all } n \Big), \qquad \text{(H.203)}
\end{aligned}$$

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 (H.14)-(H.15). Note that

$$\begin{aligned}
&\sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d} \Big| \kappa_n^{-1} \frac{\sqrt{n}\bar{m}_{n,j}(\theta_n')}{\hat{\sigma}_{n,j}(\theta_n')} - \kappa_n^{-1} \frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')} \Big| \\
&\leqslant \sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d} \Big| \kappa_n^{-1} \frac{\sqrt{n}(\bar{m}_{n,j}(\theta_n') - E_{P_n}[m_j(X_i,\theta_n')])}{\sigma_{n,j}(\theta_n')} (1+\eta_{n,j}(\theta_n')) + \kappa_n^{-1} \frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\sigma_{P_n,j}(\theta_n')} \eta_{n,j}(\theta_n') \Big| \\
&\leqslant \sup_{\theta_n'\in\theta_n+\rho/\sqrt{n}B^d} |\kappa_n^{-1}\mathbb{G}_n(\theta_n')(1+\eta_{n,j}(\theta_n'))| + \Big| \frac{\sqrt{n}E_{P_n}[m_j(X_i,\theta_n')]}{\kappa_n\sigma_{P_n,j}(\theta_n')} \eta_{n,j}(\theta_n') \Big| = o_{\mathcal{P}}(1), \qquad \text{(H.204)}
\end{aligned}$$

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 (2015b), 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,

$$\hat{\mu}_{n,j}(\theta_n) \xrightarrow{P_n} 1 - \min\Big\{ \max(0, \frac{\pi_{1,j}}{\pi_{1,j+R_1} + \pi_{1,j}}), 1 \Big\}, \qquad \text{(H.205)}$$

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

$$\hat{\mu}_{n,j}(\theta_n) \xrightarrow{P_n} 1. \qquad \text{(H.206)}$$

Similarly, if $\pi_{1,j} = -\infty$ and $\pi_{1,j+R_1} \in (-\infty, 0]$, one has

$$\hat{\mu}_{n,j+R_1}(\theta_n) \xrightarrow{P_n} 1. \qquad \text{(H.207)}$$

Now, one may write

$$\frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1 = \frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}_{n,j}(\theta_n)} \Big(\frac{\hat{\sigma}_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1 \Big) + \Big(\frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}_{n,j}(\theta_n)} - 1 \Big) = O_{P_n}(1) \Big(\frac{\hat{\sigma}_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1 \Big) + o_{P_n}(1), \qquad \text{(H.208)}$$

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,

$$\hat{\sigma}^M_{n,j}(\theta_n) = \hat{\mu}_{n,j}(\theta_n)\hat{\sigma}_{n,j}(\theta_n) + (1 - \hat{\mu}_{n,j}(\theta_n))\hat{\sigma}_{n,j+R_1}(\theta_n) \tag{H.209}$$

$$= (1 + o_{P_n}(1))\hat{\sigma}_{n,j}(\theta_n) + (1 - \hat{\mu}_{n,j}(\theta_n))O_{P_n}(\hat{\sigma}_{n,j}(\theta_n)), \tag{H.210}$$

where the second equality follows from (H.206) and the fact that

$$\hat{\sigma}_{n,j+R_1}(\theta_n) = \left(\hat{\sigma}^2_{n,j}(\theta_n) + 2\widehat{Cov}_n(m_j(X_i,\theta), t_j(X_i,\theta)) + \widehat{Var}_n(t_j(X_i,\theta))\right)^{1/2}$$

$$= \left(\hat{\sigma}^2_{n,j}(\theta_n) + O_{P_n}(\hat{\sigma}_{n,j}(\theta_n)) + O_{P_n}(1)\right)^{1/2} = O_{P_n}(\hat{\sigma}_{n,j}(\theta_n)), \tag{H.211}$$

where the second equality follows from, $Var_{P_n}(t_j(X_i,\theta))$ being bounded by Assumption E.3-(II) and

$$\widehat{Var}_n(t_j(X_i,\theta)) = Var_{P_n}(t_j(X_i,\theta)) + o_{P_n}(1) \tag{H.212}$$

$$\widehat{Cov}_n(m_j(X_i,\theta), t_j(X_i,\theta)) \leqslant \hat{\sigma}_{n,j}(\theta_n)\widehat{Var}_n(t_j(X_i,\theta))^{1/2}, \tag{H.213}$$

where the last inequality is due to the Cauchy-Schwarz inequality.

Therefore,

$$\frac{\hat{\sigma}_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1 = \frac{\hat{\sigma}_{n,j}(\theta_n) - \hat{\sigma}^M_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} = \frac{(1 - \hat{\mu}_{n,j}(\theta_n))O_{P_n}(\hat{\sigma}_{n,j}(\theta_n))}{(1 + o_{P_n}(1))\hat{\sigma}_{n,j}(\theta_n) + (1 - \hat{\mu}_{n,j}(\theta_n))O_{P_n}(\hat{\sigma}_{n,j}(\theta_n))} = o_{P_n}(1), \tag{H.214}$$

where we used $\hat{\sigma}^{-1}_{n,j}(\theta_n) = O_{P_n}(1)$ by equation (E.3) and part (i) of the lemma. By (H.208) and (H.214), $\sigma_{P_n,j}(\theta_n)/\hat{\sigma}^M_{n,j}(\theta_n) - 1 = o_{P_n}(1)$. 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 H.9 (1),

$$\left|\frac{\sigma_{P_n,j}(\theta_n)}{\sigma_{P_n,j+R_1}(\theta_n)} - 1\right| = o_{P_n}(1), \quad \text{for } j = 1, \cdots, R_1, \tag{H.215}$$

and therefore,

$$\frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)} - 1 = \frac{\sigma_{P_n,j}(\theta_n) - \hat{\sigma}^M_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)}$$

$$= \frac{[\hat{\mu}_{n,j}(\theta_n) + (1 - \hat{\mu}_{n,j}(\theta_n))]\sigma_{P_n,j}(\theta_n) - [\hat{\mu}_{n,j}(\theta_n)\hat{\sigma}_{n,j}(\theta_n) + (1 - \hat{\mu}_{n,j}(\theta_n))\hat{\sigma}_{n,j+R_1}(\theta_n)]}{\hat{\sigma}^M_{n,j}(\theta_n)}$$

$$= \frac{\hat{\mu}_{n,j}(\theta_n)[\sigma_{P_n,j}(\theta_n) - \hat{\sigma}_{n,j}(\theta_n)]}{\hat{\sigma}^M_{n,j}(\theta_n)} + \frac{(1 - \hat{\mu}_{n,j}(\theta_n))[\sigma_{P_n,j+R_1}(\theta_n) - \hat{\sigma}_{n,j+R_1}(\theta_n) + o_{P_n}(1)]}{\hat{\sigma}^M_{n,j}(\theta_n)}, \tag{H.216}$$

where the second equality follows from the definition of $\hat{\sigma}^M_{n,j}(\theta_n)$, and the third equality follows from (H.215) and $\sigma_{P_n,j+R_1}$ bounded away from 0 due to (E.3). Note that

$$\frac{\hat{\mu}_{n,j}(\theta_n)[\sigma_{P_n,j}(\theta_n) - \hat{\sigma}_{n,j}(\theta_n)]}{\hat{\sigma}^M_{n,j}(\theta_n)} = \hat{\mu}_{n,j}(\theta_n)\frac{\hat{\sigma}_{n,j}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)}\left(\frac{\sigma_{P_n,j}(\theta_n)}{\hat{\sigma}_{n,j}(\theta_n)} - 1\right) = o_{P_n}(1), \tag{H.217}$$

where the second equality follows from the first conclusion of the lemma. Similarly,

$$\frac{(1-\hat{\mu}_{n,j}(\theta_n))[\sigma_{P_n,j+R_1}(\theta_n)-\hat{\sigma}_{n,j+R_1}(\theta_n)+o_{P_n}(1)]}{\hat{\sigma}^M_{n,j}(\theta_n)}$$
$$=(1-\hat{\mu}_{n,j}(\theta_n))\frac{\hat{\sigma}_{n,j+R_1}(\theta_n)}{\hat{\sigma}^M_{n,j}(\theta_n)}\Big(\frac{\sigma_{P_n,j+R_1}(\theta_n)}{\hat{\sigma}_{n,j+R_1}(\theta_n)}-1+o_{P_n}(1)\Big)=o_{P_n}(1). \quad \text{(H.218)}$$

By (H.216)-(H.218), 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. □

H.2 Lemmas Used to Prove Theorem D.1

Let $\{X^b_i\}^n_{i=1}$ denote a bootstrap sample drawn randomly from the empirical distribution. Define

$$\mathfrak{G}^b_{n,j}(\theta)\equiv\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(m_j(X^b_i,\theta)-\bar{m}_n(\theta)\right)/\sigma_{P,j}(\theta)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^n(M_{n,i}-1)m_j(X_i,\theta)/\sigma_{P,j}(\theta), \quad \text{(H.219)}$$

where $\{M_{n,i}\}^n_{i=1}$ denotes the multinomial weights on the original sample, and we let P^*_n denote the conditional distribution of $\{M_{n,i}\}^n_{i=1}$ given the sample path $\{X_i\}^\infty_{i=1}$ (see Appendix H.3 for details on the construction of the bootstrapped empirical process).

LEMMA H.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*

$$\sup_Q N(\epsilon\|F\|_{L^2_Q},\mathcal{M}_P,L^2_Q)\leqslant K\epsilon^{-v},\ 0<\epsilon<1, \quad \text{(H.220)}$$

where the supremum is taken over all discrete distributions; (ii) There exists a positive constant $\gamma>0$ such that

$$\|(\theta_1,\tilde{\theta}_1)-(\theta_2,\tilde{\theta}_2)\|\leqslant\delta\quad\Rightarrow\quad\sup_{P\in\mathcal{P}}\|Q_P(\theta_1,\tilde{\theta}_1)-Q_P(\theta_2,\tilde{\theta}_2)\|\leqslant M\delta^\gamma. \quad \text{(H.221)}$$

Let δ_n be a positive sequence tending to 0 and let ϵ_n be a positive sequence such that $\epsilon_n/|\delta^\gamma_n\ln\delta_n|\to\infty$ as $n\to\infty$. Then,

$$\sup_{P\in\mathcal{P}}P\left(\sup_{\|\theta-\theta'\|\leqslant\delta_n}\|\mathbb{G}_n(\theta)-\mathbb{G}_n(\theta'))\|>\epsilon_n\right)=o(1). \quad \text{(H.222)}$$

Further,

$$\lim_{n\to\infty}P^*_n\left(\sup_{\|\theta-\theta'\|\leqslant\delta_n}\|\mathfrak{G}^b_n(\theta)-\mathfrak{G}^b_n(\theta'))\|>\epsilon_n|\{X_i\}^\infty_{i=1}\right)=0. \quad \text{(H.223)}$$

for almost all sample paths $\{X_i\}^\infty_{i=1}$ 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:

$$P\left(\sup_{\|\theta-\theta'\|\leqslant\delta_n}\max_{j=1,\cdots,J}|\mathbb{G}_{n,j}(\theta)-\mathbb{G}_{n,j}(\theta')|>\epsilon_n\right)=o(1). \tag{H.224}$$

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 (2015b), one has

$$\|\theta-\theta'\|\leqslant\delta_n \quad\Rightarrow\quad \varrho_P(\theta,\theta')\leqslant\tilde{\delta}_n, \tag{H.225}$$

where $\tilde{\delta}_n=O(\delta_n^\gamma)$ by assumption. Define

$$\mathcal{M}_{P,\tilde{\delta}_n}=\{\sigma_{P,j}(\theta)^{-1}m_j(\cdot,\theta)-\sigma_{P,j}(\theta')^{-1}m_j(\cdot,\theta')|\theta,\theta'\in\Theta,\varrho_P(\theta,\tilde{\theta})<\tilde{\delta}_n,j=1,\cdots,J\}. \tag{H.226}$$

Define $Z_n(\tilde{\delta}_n)\equiv\sup_{f\in\mathcal{M}_{\tilde{\delta}_n}}|\sqrt{n}(\mathbb{P}_n-P)f|$. Then, by (H.225), one has

$$P\left(\sup_{\|\theta-\theta'\|\leqslant\delta_n}\max_{j=1,\cdots,J}|\mathbb{G}_{n,j}(\theta)-\mathbb{G}_{n,j}(\theta'))|>\epsilon_n\right)\leqslant P(Z_n(\tilde{\delta}_n)>\epsilon_n\Big). \tag{H.227}$$

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

$$P(Z_n(\tilde{\delta}_n)>\epsilon_n)\leqslant\frac{2}{\epsilon_n}E_{P\times P^W}\left[\sup_{f\in\mathcal{M}_{P,\tilde{\delta}_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_if(X_i)\right|\right], \tag{H.228}$$

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 $f\mapsto\{n^{-1/2}\sum_{i=1}^n W_if(X_i)\}$ 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

$$\begin{aligned}
E_{P^W}\left[\sup_{f\in\mathcal{M}_{\tilde{\delta}_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_if(X_i)\right|\right]&\leqslant K\int_0^{\tilde{\delta}_n}\sqrt{\ln N(\epsilon,\mathcal{M}_{P,\tilde{\delta}_n},L^2_{\hat{P}_n})}d\epsilon\\
&\leqslant K\int_0^{\tilde{\delta}_n/\|F\|_{L^2_Q}}\sup_Q\sqrt{\ln N(\epsilon\|F\|_{L^2_Q},\mathcal{M}_P,L^2_Q)}d\epsilon\\
&\leqslant K'\int_0^{\tilde{\delta}_n/\|F\|_{L^2_Q}}\sqrt{-v\ln\epsilon}d\epsilon,
\end{aligned} \tag{H.229}$$

for some $K'>0$, where the last inequality follows from (H.220). Note that $\sqrt{-\ln\epsilon}\leqslant-\ln\epsilon$ for $\epsilon\leqslant\tilde{\delta}_n/\|F\|_{L^2_Q}$ with n sufficiently large. Hence,

$$E_{P^W}\left[\sup_{f\in\mathcal{M}_{\tilde{\delta}_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_if(X_i)\right|\right]\leqslant K'v^{1/2}\int_0^{\tilde{\delta}_n/\|F\|_{L^2_Q}}(-\ln\epsilon)d\epsilon=K'v^{1/2}(\tilde{\delta}_n-\tilde{\delta}_n\ln(\tilde{\delta}_n)). \tag{H.230}$$

By (H.228) and taking expectations with respect to P in (H.230), it follows that

$$P(Z_n(\tilde{\delta}_n)>\epsilon_n)\leqslant 2K'v^{1/2}(\tilde{\delta}_n-\tilde{\delta}_n\ln(\tilde{\delta}_n))/\epsilon_n=O(\delta_n^\gamma/\epsilon_n)+O(|\delta_n^\gamma\ln(\delta_n)|/\epsilon_n)=o(1), \tag{H.231}$$

where the last equality follows from the rate condition on ϵ_n. By (H.227) and (H.231), 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 (H.225), one has

$$P_n^*\left(\sup_{\|\theta-\theta'\|\leqslant\delta_n}\max_{j=1,\cdots,J}|\mathfrak{G}_{n,j}^b(\theta)-\mathfrak{G}_{n,j}^b(\theta')|>\epsilon_n\middle|\{X_i\}_{i=1}^\infty\right)\leqslant P_n^*\big(Z_n^*(\tilde{\delta}_n)>\epsilon_n\big|\{X_i\}_{i=1}^\infty\big). \tag{H.232}$$

By Markov's inequality and Lemma 2.3.1 (symmetrization lemma) in van der Vaart and Wellner (2000), one has

$$P_n^*\big(Z_n^*(\tilde{\delta}_n)>\epsilon_n\big|\{X_i\}_{i=1}^\infty\big)\leqslant\frac{2}{\epsilon_n}E_{P_n^*\times P^W}\left[\sup_{f\in\mathcal{M}_{P,\delta_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i f(X_i^b)\right|\middle|\{X_i\}_{i=1}^\infty\right] \tag{H.233}$$

$$=\frac{2}{\epsilon_n}E_{P_n^*}\left[E_{P^W}\left[\sup_{f\in\mathcal{M}_{P,\delta_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i f(X_i^b)\right|\middle|\{X_i^b\},\{X_i\}_{i=1}^\infty\right]\middle|\{X_i\}_{i=1}^\infty\right], \tag{H.234}$$

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 (H.228)-(H.231). 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 H.12: *Suppose Assumptions E.1, E.2, and E.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*

$$\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)|^*=O_{\mathcal{P}}(1/\sqrt{n}); \tag{H.235}$$

(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*

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}}P\Big(\max_{j=1,\cdots,J}\sup_{\|\theta-\theta'\|<\delta}|\eta_{n,j}(\theta)-\eta_{n,j}(\theta')|>C\delta\Big)<\eta. \tag{H.236}$$

Proof. We show the claim by first showing that, for any $\delta>0$, there exist $M>0$ and $N\in\mathbb{N}$ such that

$$\inf_{P\in\mathcal{P}}P^\infty\Big(\sup_{\theta\in\Theta}\Big|\frac{\hat{\sigma}_{n,j}(\theta)}{\sigma_{P,j}(\theta)}-1\Big|\leqslant M/\sqrt{n}\Big)\geqslant 1-\delta,\ \forall n\geqslant N. \tag{H.237}$$

By Assumptions E.1 (iv), E.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 $\epsilon>0$,

$$\Big|P\Big(\sqrt{n}\sup_{\theta\in\Theta}\Big|\frac{n^{-1}\sum_{i=1}^n m_j(X_i,\theta)^2}{\sigma_{P,j}^2(\theta)}-\frac{E_P[m_j(X,\theta)^2]}{\sigma_{P,j}^2(\theta)}\Big|\leqslant C_1\Big)-\Pr(\sup_{\theta\in\Theta}|\mathbb{H}_{P,j}(\theta)|\leqslant C_1)\Big|\leqslant\epsilon \tag{H.238}$$

$$\Big|P\Big(\sqrt{n}\sup_{\theta\in\Theta}\Big|\frac{\bar{m}_{n,j}(\theta)-E_P[m_j(X,\theta)]}{\sigma_{P,j}(\theta)}\Big|\leqslant C_2\Big)-\Pr(\sup_{\theta\in\Theta}|\mathbb{G}_{P,j}(\theta)|\leqslant C_2)\Big|\leqslant\epsilon. \tag{H.239}$$

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 H.10 (i), bounding each term of the right hand side of (H.201) by $C_1/\sqrt{n}$ and $C_2/\sqrt{n}$ implies that

$\sup_{\theta\in\Theta}\left|\frac{\hat\sigma^2_{n,j}(\theta)}{\sigma^2_{P,j}(\theta)}-1\right|\leqslant C/\sqrt{n}$ for some constant $C>0$. Now choose $C_1>0$ and $C_2>0$ so that

$$\Pr(\sup_{\theta\in\Theta}|\mathbb{H}_{P,j}(\theta)|\leqslant C_1)>1-\delta/3\quad\text{and}\quad\Pr(\sup_{\theta\in\Theta}|\mathbb{G}_{P,j}(\theta)|\leqslant C_2)>1-\delta/3 \tag{H.240}$$

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,

$$\begin{aligned}1-\delta&\leqslant P\Big(\sqrt{n}\sup_{\theta\in\Theta}\Big|\frac{n^{-1}\sum_{i=1}^n m_j(X_i,\theta)^2}{\sigma^2_{P,j}(\theta)}-\frac{E_P[m_j(X,\theta)^2]}{\sigma^2_{P,j}(\theta)}\Big|\leqslant C_1,\sqrt{n}\sup_{\theta\in\Theta}\Big|\frac{\bar m_{n,j}(\theta)-E_P[m_j(X,\theta)]}{\sigma_{P,j}(\theta)}\Big|\leqslant C_2\Big)\\&\leqslant P\Big(\sup_{\theta\in\Theta}\Big|\frac{\hat\sigma^2_{n,j}(\theta)}{\sigma^2_{P,j}(\theta)}-1\Big|\leqslant C/\sqrt{n}\Big),\end{aligned} \tag{H.241}$$

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 (H.241), for sufficiently large n,

$$1-\delta\leqslant P\Big(\sup_{\theta\in\Theta}\Big|\frac{\hat\sigma_{n,j}(\theta)}{\sigma_{P,j}(\theta)}-1\Big|\leqslant C/\sqrt{n}\Big), \tag{H.242}$$

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

$$P\Big(\sup_{\theta\in\Theta}|\eta_{n,j}(\theta)|\leqslant C'/\sqrt{n}\Big)\geqslant 1-\delta, \tag{H.243}$$

uniformly in $P\in\mathcal{P}$ for all n sufficiently large. This establishes the first claim of the lemma.

(ii) First, consider

$$\frac{\hat\sigma^2_{n,j}(\theta)}{\sigma^2_{P,j}(\theta)}=n^{-1}\sum_{i=1}^n\left(\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right)^2-\left(n^{-1}\sum_{i=1}^n\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right)^2. \tag{H.244}$$

We claim that this function is Lipschitz with probability approaching 1. To see this, note that, for any $\theta,\theta'\in\Theta$,

$$\begin{aligned}&\left|n^{-1}\sum_{i=1}^n\left(\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right)^2-n^{-1}\sum_{i=1}^n\left(\frac{m(X_i,\theta')}{\sigma_{P,j}(\theta')}\right)^2\right|\\&=\left|n^{-1}\sum_{i=1}^n\left(\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}+\frac{m(X_i,\theta')}{\sigma_{P,j}(\theta')}\right)\left(\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}-\frac{m(X_i,\theta')}{\sigma_{P,j}(\theta')}\right)\right|\\&\leqslant n^{-1}\sum_{i=1}^n 2\sup_{\theta\in\Theta}\left|\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right|\bar M(X_i)\|\theta-\theta'\|.\end{aligned} \tag{H.245}$$

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,

$$P(B_n>C)\leqslant\frac{E[B_n]}{C}\leqslant\frac{2E_P\left[\sup_{\theta\in\Theta}\left|\frac{m(X_i,\theta)}{\sigma_{P,j}(\theta)}\right|^2\right]^{1/2}E_P\left[\bar M(X_i)^2\right]^{1/2}}{C}\leqslant\frac{2M}{C}, \tag{H.246}$$

where the third inequality is due to Assumptions E.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 $\theta, \theta' \in \Theta$,

$$\begin{aligned}
&\left| \left(n^{-1} \sum_{i=1}^{n} \frac{m(X_i, \theta)}{\sigma_{P,j}(\theta)} \right)^2 - \left(n^{-1} \sum_{i=1}^{n} \frac{m(X_i, \theta')}{\sigma_{P,j}(\theta')} \right)^2 \right| \\
&= \left| n^{-1} \sum_{i=1}^{n} \frac{m(X_i, \theta)}{\sigma_{P,j}(\theta)} + n^{-1} \sum_{i=1}^{n} \frac{m(X_i, \theta')}{\sigma_{P,j}(\theta')} \right| \left| n^{-1} \sum_{i=1}^{n} \frac{m(X_i, \theta)}{\sigma_{P,j}(\theta)} - n^{-1} \sum_{i=1}^{n} \frac{m(X_i, \theta')}{\sigma_{P,j}(\theta')} \right| \\
&\leqslant n^{-1} \sum_{i=1}^{n} 2 \sup_{\theta \in \Theta} \left| \frac{m(X_i, \theta)}{\sigma_{P,j}(\theta)} \right| n^{-1} \sum_{i=1}^{n} \bar{M}(X_i) \|\theta - \theta'\|. \qquad \text{(H.247)}
\end{aligned}$$

Define $\tilde{B}_n \equiv n^{-1} \sum_{i=1}^{n} 2 \sup_{\theta \in \Theta} \left| \frac{m(X_i, \theta)}{\sigma_{P,j}(\theta)} \right| n^{-1} \sum_{i=1}^{n} \bar{M}(X_i)$. By Markov, Cauchy-Schwarz, and Jensen's inequalities,

$$\begin{aligned}
P(\tilde{B}_n > C) \leqslant \frac{E[\tilde{B}_n]}{C} &\leqslant \frac{2 E_P\Big[\Big(n^{-1} \sum \sup_{\theta \in \Theta} \left| \frac{m(X_i, \theta)}{\sigma_{P,j}(\theta)} \right| \Big)^2 \Big]^{1/2} E_P\Big[\Big(n^{-1} \sum \bar{M}(X_i) \Big)^2 \Big]^{1/2}}{C} \\
&\leqslant \frac{2 E_P\Big[\sup_{\theta \in \Theta} \left| \frac{m(X_i, \theta)}{\sigma_{P,j}(\theta)} \right|^2 \Big]^{1/2} E_P[\bar{M}(X_i)^2]^{1/2}}{C} \leqslant \frac{2M}{C}, \qquad \text{(H.248)}
\end{aligned}$$

where the last inequality is due to Assumptions E.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 (H.242), $\hat{\sigma}^2_{n,j}(\theta) / \sigma^2_{P,j}(\theta)$ converges to 1 in probability, and g is locally Lipschitz on a neighborhood of 1. Combining this with (H.244)-(H.248) yields the desired result. □

LEMMA H.13: *Suppose Assumption E.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,

$$\sup_{P \in \mathcal{P}} \|Q_P(\theta_1, \tilde{\theta}_1) - Q_P(\theta_2, \tilde{\theta}_2)\| \leqslant M \|(\theta_1, \tilde{\theta}_1) - (\theta_2, \tilde{\theta}_2)\|, \qquad \text{(H.249)}$$

for some $M > 0$ *and for all* $\theta_1, \tilde{\theta}_1, \theta_2, \tilde{\theta}_2 \in \Theta$.

Proof. Recall that

$$[Q_P(\theta_1, \tilde{\theta}_1)]_{j,k} = E_P\Big[\frac{m_j(X_i, \theta_1)}{\sigma_{P,j}(\theta_1)} \frac{m_k(X_i, \tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)} \Big] - E_P\Big[\frac{m_j(X_i, \theta_1)}{\sigma_{P,j}(\theta_1)} \Big] E_P\Big[\frac{m_k(X_i, \tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)} \Big]. \qquad \text{(H.250)}$$

For any $\theta_1, \tilde{\theta}_1, \theta_2, \tilde{\theta}_2 \in \Theta$,

$$\begin{aligned}
&\left| E_P\Big[\frac{m_j(X_i, \theta_1)}{\sigma_{P,j}(\theta_1)} \frac{m_k(X_i, \tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)} \Big] - E_P\Big[\frac{m_j(X_i, \theta_2)}{\sigma_{P,j}(\theta_2)} \frac{m_k(X_i, \tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)} \Big] \right| \\
&\leqslant \left| E_P\Big[\Big(\frac{m_j(X_i, \theta_1)}{\sigma_{P,j}(\theta_1)} - \frac{m_j(X_i, \theta_2)}{\sigma_{P,j}(\theta_2)} \Big) \frac{m_k(X_i, \tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)} \Big] \right| + \left| E_P\Big[\frac{m_j(X_i, \theta_1)}{\sigma_{P,j}(\theta_1)} \Big(\frac{m_k(X_i, \tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)} - \frac{m_k(X_i, \tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)} \Big) \Big] \right| \\
&\leqslant E_P\Big[\sup_{\theta \in \Theta} \left| \frac{m_k(X_i, \theta)}{\sigma_{P,k}(\theta)} \right| \bar{M}(X_i) \Big] \|\theta_1 - \theta_2\| + E_P\Big[\sup_{\theta \in \Theta} \left| \frac{m_j(X_i, \theta)}{\sigma_{P,j}(\theta)} \right| \bar{M}(X_i) \Big] \|\tilde{\theta}_1 - \tilde{\theta}_2\| \\
&\leqslant M(\|\theta_1 - \theta_2\| + \|\tilde{\theta}_1 - \tilde{\theta}_2\|), \qquad \text{(H.251)}
\end{aligned}$$

where the last inequality is due to the Cauchy-Schwarz inequality, Assumption E.1 (iv), and the assumption on $\bar{M}$.

Similarly, for any $\theta_1, \tilde{\theta}_1, \theta_2, \tilde{\theta}_2 \in \Theta$,

$$\begin{aligned}
&\Big|E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}\Big]E_P\Big[\frac{m_k(X_i,\tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)}\Big] - E_P\Big[\frac{m_j(X_i,\theta_2)}{\sigma_{P,j}(\theta_2)}\Big]E_P\Big[\frac{m_k(X_i,\tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)}\Big]\Big| \\
&\leqslant \Big|E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)} - \frac{m_j(X_i,\theta_2)}{\sigma_{P,j}(\theta_2)}\Big]\Big|\Big|E_P\Big[\frac{m_k(X_i,\tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)}\Big]\Big| + \Big|E_P\Big[\frac{m_j(X_i,\theta_1)}{\sigma_{P,j}(\theta_1)}\Big]\Big|\Big|E_P\Big[\frac{m_k(X_i,\tilde{\theta}_1)}{\sigma_{P,k}(\tilde{\theta}_1)} - \frac{m_k(X_i,\tilde{\theta}_2)}{\sigma_{P,k}(\tilde{\theta}_2)}\Big]\Big| \\
&\leqslant E_P\Big[\sup_{\theta\in\Theta}\Big|\frac{m_k(X_i,\theta)}{\sigma_{P,k}(\theta)}\Big|\Big]E_P[\bar{M}(X_i)]\|\theta_1-\theta_2\| + E_P\Big[\sup_{\theta\in\Theta}\Big|\frac{m_j(X_i,\theta)}{\sigma_{P,j}(\theta)}\Big|\Big]E_P[\bar{M}(X_i)]\|\tilde{\theta}_1-\tilde{\theta}_2\| \\
&\leqslant M(\|\theta_1-\theta_2\| + \|\tilde{\theta}_1-\tilde{\theta}_2\|), \qquad \text{(H.252)}
\end{aligned}$$

where the last inequality is due to the Cauchy-Schwarz inequality, Assumption E.1 (iv), and the assumption on $\bar{M}$. The conclusion of the lemma then follows from (H.250)-(H.252). □

H.3 Almost Sure Representation Lemma and Related Results

In this appendix, we provide details on the almost sure representation used in Lemmas H.3, H.4, H.6, and H.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^∞ as the coordinate projection on the first ∞ 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^∞. 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 Θ. We simply write $\ell^\infty(\Theta)$ when $J = 1$.

Using the multinomial weight, we rewrite the empirical bootstrap process as

$$\mathbb{G}^b_{n,j}(\cdot) = g_j(X^\infty, M_n) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n (M_{n,i} - 1)m_j(X_i, \cdot)/\hat{\sigma}_{n,j}(\cdot),\ j = 1, \cdots, J, \qquad \text{(H.253)}$$

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^∞, M_n) to the empirical bootstrap process $\mathbb{G}^b_{n,j}$. 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 $f(\mathbb{G}^b_n)$ given the sample path X^∞ is

$$E_M[f(\mathbb{G}^b_n)|X^\infty = x^\infty] = \int f \circ g(x^\infty, m_n) dQ(m_n), \qquad \text{(H.254)}$$

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:

$$\begin{aligned}
\mathfrak{G}^b_{n,j}(\theta) &\equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(m_j(X^b_i, \theta) - \bar{m}_n(\theta)\big)/\sigma_{P,j}(\theta) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n (M_{n,i} - 1)m_j(X_i, \theta)/\sigma_{P,j}(\theta). \qquad \text{(H.255)}
\end{aligned}$$

The following result was shown in the proof of Lemma D.2.8 in Bugni, Canay, and Shi (2015b), 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 H.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,*

$$\sup_{h\in BL_1} |E_M[h(\mathfrak{G}_n^b)|X^\infty] - E[h(\mathbb{G}_P)]| \xrightarrow{as*} 0, \tag{H.256}$$

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 H.15: *Suppose Assumptions E.1, E.2, and E.5 hold. Then,*

$$\sup_{h\in BL_1} |E_M[h(\mathbb{G}_n^b)|X^\infty] - E[h(\mathbb{G}_P)]| \xrightarrow{as*} 0, \tag{H.257}$$

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

Proof. By Assumptions E.1 (iv) and E.5, Assumptions A.1-A.4 in Bugni, Canay, and Shi (2015b) 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 E.1 (iv) again, $\sup_{P\in\mathcal{P}} P^* \|f - Pf\|_{\mathcal{F}} < \infty$. Hence, by Lemma H.14,

$$\inf_{P\in\mathcal{P}} P^\infty\Big(\sup_{h\in BL_1} |E_M[h(\mathfrak{G}_n^b)|X^\infty] - E[h(\mathbb{G}_P)]|^* \to 0\Big) = 1. \tag{H.258}$$

For later use, we define the following set of sample paths, which has probability 1 uniformly in $P \in \mathcal{P}$.

$$A \equiv \Big\{x^\infty \in \mathcal{X}^\infty : \sup_{h\in BL_1} |E_M[h(\mathfrak{G}_n^b)|X^\infty = x^\infty] - E[h(\mathbb{G}_P)]|^* \to 0\Big\}. \tag{H.259}$$

Note that $\mathbb{G}_{n,j}^b$ and $\mathfrak{G}_{n,j}^b$ are related to each other by the following relationship:

$$\mathbb{G}_{n,j}^b(\theta) - \mathfrak{G}_{n,j}^b(\theta) = \mathfrak{G}_{n,j}^b(\theta)\left(\frac{\sigma_{P,j}(\theta)}{\hat{\sigma}_{n,j}(\theta)} - 1\right) = \mathfrak{G}_{n,j}^b(\theta)\eta_{n,j}(\theta),\ \theta \in \Theta. \tag{H.260}$$

By Assumptions E.1, E.2, and E.5, Lemma H.10 applies. Hence,

$$\inf_{P\in\mathcal{P}} P^\infty\Big(\sup_{\theta\in\Theta} |\eta_{n,j}(\theta)|^* \to 0\Big) = 1. \tag{H.261}$$

Define the following set of sample paths:

$$B \equiv \Big\{x^\infty \in \mathcal{X}^\infty : \sup_{\theta\in\Theta} |\eta_{n,j}(\theta)|^* \to 0, \forall j = 1, \cdots, J\Big\}. \tag{H.262}$$

For any $x^\infty \in A \cap B$, it then follows that

$$\sup_{h\in BL_1} \big|E_M[h(\mathbb{G}_n^b)|X^\infty = x^\infty] - E[h(\mathbb{G}_P)]\big|^* \to 0, \tag{H.263}$$

due to (H.258) and (H.260), h being Lipschitz, $\mathfrak{G}_{n,j}^b$ being bounded (given x^∞), 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 (H.258), (H.261), and De Morgan's law. This establishes the conclusion of the lemma. □

The following lemma shows that, for almost all sample path x^∞, one can find an almost sure representation of

the bootstrapped empirical process that is convergent.

LEMMA H.16: *Suppose Assumptions E.1, E.2, and E.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*

$$\int h \circ g(x^\infty, m_n) dQ(m_n) = \int h(\tilde{G}_{n,x^\infty}(\tilde{\omega})) d\tilde{\mathbf{P}}^*(\tilde{\omega}), \ \forall h \in BL_1 \tag{H.264}$$

$$\int h(\mathbb{G}_P(\omega)) dP(\omega) = \int h(\tilde{G}_{P,x^\infty}(\tilde{\omega})) d\tilde{\mathbf{P}}^*(\tilde{\omega}), \ \forall h \in BL_1, \tag{H.265}$$

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

$$\inf_{P\in\mathcal{P}} P^\infty\left(\{x^\infty \in \mathcal{X}^\infty : \tilde{G}_{n,x^\infty} \overset{\tilde{\mathbf{P}}-as*}{\to} \tilde{G}_{P,x^\infty}\}\right) = 1. \tag{H.266}$$

Proof. Define the following set of sample paths:

$$C \equiv \left\{x^\infty \in \mathcal{X}^\infty : \sup_{h\in BL_1} |E_M[h(\mathbb{G}^b_{n,j})|X^\infty = x^\infty] - E[h(\mathbb{G}_P)]|^* \to 0\right\}. \tag{H.267}$$

By Lemma H.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 (H.253). 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 H.15,

$$\sup_{h\in BL_1} \left| \int h \circ g(x^\infty, m_n) dQ(m_n) - E[h(\mathbb{G}_P)] \right|^* \to 0. \tag{H.268}$$

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

$$\int h \circ g(x^\infty, m_n) dQ(m_n) = \int h(\tilde{G}_{n,x^\infty}(\tilde{\omega})) d\tilde{\mathbf{P}}^*(\tilde{\omega}), \ \forall h \in BL_1. \tag{H.269}$$

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

$$\tilde{G}_{n,x^\infty}(\tilde{\omega}) = g(x^\infty, M_n(\phi_n(\tilde{\omega}))), \tag{H.270}$$

for some perfect map $\phi_n : \tilde{\Omega} \to \mathcal{Z}$ (see the construction of ϕ_α 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. (H.263), and $\inf_{P\in\mathcal{P}} P(C) = 1$, it follows that

$$\inf_{P\in\mathcal{P}} P^\infty\left(\{x^\infty \in \mathcal{X}^\infty : \tilde{G}_{n,x^\infty} \overset{\tilde{\mathbf{P}}-as*}{\to} \tilde{G}_{P,x^\infty}\}\right) = 1. \tag{H.271}$$

This establishes the claim of the lemma. □

LEMMA H.17: *Suppose Assumptions E.1, E.2, and E.5 hold. Let $W_n \equiv (\mathbb{G}^b_n, Y_n)$ 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*

$$\inf_{P\in\mathcal{P}} P^\infty\big(\sup_{h\in BL_1} |E_M[h(W_n)|X^\infty = x^\infty] - E[h(W)]|^* \to 0\big) = 1, \tag{H.272}$$

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*

$$E_M\big[h(W_n)|X^\infty = x^\infty\big] = \int h(W^*_{n,x^\infty}(\tilde{\omega}))d\tilde{\mathbf{P}}^*(\tilde{\omega}),\ \forall h \in BL_1 \tag{H.273}$$

$$E\big[h(W)\big] = \int h(W^*_{x^\infty}(\tilde{\omega}))d\tilde{\mathbf{P}}^*(\tilde{\omega}),\ \forall h \in BL_1, \tag{H.274}$$

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*

$$\inf_{P\in\mathcal{P}} P^\infty\Big(\big\{x^\infty \in \mathcal{X}^\infty : W^*_{n,x^\infty} \overset{\tilde{\mathbf{P}}-as*}{\to} \tilde{W}^*_{x^\infty}\big\}\Big) = 1. \tag{H.275}$$

Proof. Let $C \equiv \{x^\infty : \sup_{h\in BL_1} |E_M\big[h(W_n)|X^\infty = x^\infty\big] - E\big[h(W)\big]|^* \to 0\}$. The rest of the proof is the same as the one for Lemma H.16 and is therefore omitted. □

REMARK H.1: When called by the Lemmas in Appendix H, Lemma H.17 is applied, for example, with $Y_n = (vec(\hat{D}_n(\theta'_n)), \hat{\xi}_n(\theta'_n))$ and $Y = (vec(D), \pi_1)$.

References

Adams, R. A., and J. J. Fournier (2003): *Sobolev spaces*, vol. 140. Academic press.

Andrews, D. W. (1994): "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.

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.

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 (2015a): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185(1), 259–282.

——— (2015b): "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.

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

Mammen, E. (1992): *When Does Bootstrap Work? Asymptotic Results and Simulations*. Spinger Verlag, New York, NY.

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*, 19(4), 541–564.

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.

Rasmussen, C. E., and C. K. I. Williams (2005): *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

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.