Confidence Intervals for Projections of Partially Identified Parameters*

Hiroaki Kaido† Francesca Molinari‡ Jörg Stoye§

October 25, 2017

Abstract

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

The extreme points of the calibrated projection confidence interval are obtained by extremizing the value of the component (or function) of interest subject to a proper relaxation of studentized sample analogs of the moment (in)equality conditions. The degree of relaxation, or critical level, is calibrated so that the component (or function) of 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. The algorithm is of independent interest for inference on optimal values of stochastic nonlinear programs. We establish its convergence under conditions satisfied by canonical examples in the moment (in)equalities literature.

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

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

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

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

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

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

1 Introduction

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

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

However, applied researchers are frequently primarily interested in a specific component (or function) of 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 the projection $p'\theta$, where p is a known unit vector. To date, it has been common to report as confidence interval for $p'\theta$ the projection of $\mathcal{C}_n(c_{1-\alpha})$:

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

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

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

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

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

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

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

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

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

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

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

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

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

In an influential paper, Pakes, Porter, Ho, and Ishii (2011) also use linearization but, subject to this approximation, directly bootstrap the sample projection.[3] This is valid only under stringent conditions, and we show that calibrated projection can be much simplified under those conditions. Other related papers that explicitly consider inference on projections include Andrews, Berry, and Jia (2004), Beresteanu and Molinari (2008), Bontemps, Magnac, and Maurin (2012), Chen, Tamer, and Torgovitsky (2011), Kaido (2016), Kitagawa (2012), Kline and Tamer (2015), and Wan (2013). However, some are Bayesian, as opposed to our frequentist approach, and none of them establish uniform validity of confidence sets. Chen, Christensen, and Tamer (2017) establish uniform validity of MCMC-based confidence intervals for projections, but these are aimed at covering the entire set $\{p'\theta : \theta \in \Theta_I(P)\}$, whereas we aim at covering the projection of θ. Finally, Gafarov, Meier, and Montiel-Olea (2016) have used our insight in the context of set identified spatial VARs.

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

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

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

2 Detailed Explanation of the Method

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[5] Under Assumption 4.3-(II), in equation (2.5) instead of $\hat{\sigma}_{n,j}$ we use the estimator $\hat{\sigma}_{n,j}^M$ specified in (E.188) in Lemma E.10 p.51 of the Online Appendix for $j = 1, \ldots, 2R_1$ (with $R_1 \leqslant J_1/2$ defined in the assumption). In equation (3.2) we use $\hat{\sigma}_{n,j}$ for all $j = 1, \ldots, J$. To ease notation, we distinguish the two only where needed.

[6] For a simple analogy, consider the point identified model defined by the single moment equality $E_P(m_1(X_i, \theta)) = E_P(X_i) - \theta = 0$, where θ is a scalar. In this case, $\mathcal{C}_n(\hat{c}_n) = \bar{X} \pm \hat{c}_n \hat{\sigma}_n / \sqrt{n}$. The upper endpoint of the confidence interval can be written as $\sup_\theta \{p'\theta \text{ s.t. } -\hat{c}_n \leqslant \sqrt{n}(\bar{X} - \theta)/\hat{\sigma}_n \leqslant \hat{c}_n\}$, with $p = 1$, and similarly for the lower endpoint.

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

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

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

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

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

Here $\mathbb{G}_{n,j}(\cdot) \equiv \sqrt{n}(\bar{m}_{n,j}(\cdot) - E_P[m_j(X_i, \cdot)])/\sigma_{P,j}(\cdot)$ is a normalized empirical process indexed by $\theta \in \Theta$, $D_{P,j}(\cdot) \equiv \nabla_\theta\{E_P[m_j(X_i, \cdot)]/\sigma_{P,j}(\cdot)\}$ is the gradient of the normalized moment, $\gamma_{1,P,j}(\cdot) \equiv E_P(m_j(X_i, \cdot))/\sigma_{P,j}(\cdot)$ is the studentized population moment, and the mean value $\bar{\theta}$ lies componentwise between θ and $\theta + \lambda/\sqrt{n}$.[7]

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

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

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

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

3 Computing Calibrated Projection Confidence Intervals

3.1 Computing the Critical Level

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

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

where $\mathbb{G}_{n,j}^b(\cdot) = n^{-1/2}\sum_{i=1}^n (m_j(X_i^b, \cdot) - \bar{m}_{n,j}(\cdot))/\hat{\sigma}_{n,j}(\cdot)$ is a bootstrap analog of $\mathbb{G}_{P,j}$,[11] $\hat{D}_{n,j}(\cdot)$ is a consistent estimator of $D_{P,j}(\cdot)$, $\rho > 0$ is a constant chosen by the researcher (see Section 4.3), $B^d = [-1,1]^d$, and $\hat{\xi}_{n,j}$ is defined by

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

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

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

Restrictions on φ and the rate at which κ_n diverges are imposed in Assumption 4.2.

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

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

[11] Bugni, Canay, and Shi (2017) approximate the stochastic process $\mathbb{G}_{P,j}$ using $n^{-1/2}\sum_{i=1}^n [(m_j(X_i, \cdot) - \bar{m}_{n,j}(\cdot))/\hat{\sigma}_{n,j}(\cdot)]\chi_i$ with $\{\chi_i \sim N(0,1)\}_{i=1}^n$ i.i.d. This approximation is equally valid in our approach, and can be computationally faster as it avoids repeated evaluation of $m_j(X_i^b, \cdot)$ across bootstrap replications.

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

Heuristically, the random convex polyhedral set $\Lambda_n^b(\theta, \rho, c)$ in (3.1) is a local (to θ) linearized bootstrap approximation to the random constraint set in (2.7). To see this, note that the bootstrapped empirical process and the estimator of the gradient approximate the first two terms in the constraint in (2.7) as linearized in (2.8). Next, for $\theta \in \Theta_I(P)$, the GMS function conservatively approximates the local slackness parameter $\sqrt{n}\gamma_{1,P,j}(\theta)$. This is needed because the scaling of $\sqrt{n}\gamma_{1,P,j}(\theta)$ precludes consistent estimation. The problem is resolved by shrinking estimated intercepts toward zero, thereby tightening constraints and hence increasing $\hat{c}_n(\theta)$. As with other uses of GMS, the resulting conservative distortion vanishes pointwise but not uniformly. Finally, restricting λ to the "ρ-box" ρB^d has a strong regularizing effect: It ensures uniform validity in challenging situations, including several that are assumed away in most of the literature. We discuss this point in more detail in Section 4.3.

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

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

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

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

so that we can equivalently define

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

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

[12] These are $\varphi^1 - \varphi^4$ in Andrews and Soares (2010), all of which depend on $\kappa_n^{-1}\sqrt{n}\bar{m}_{n,j}(\theta)/\hat{\sigma}_{n,j}(\theta)$. We do not consider GMS function φ^5 in Andrews and Soares (2010), which depends also on the covariance matrix of the moment functions.

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

3.2 Computation of the Outer Maximization Problem

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3 Convergence of the E-A-M Algorithm

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

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

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

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

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

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

REMARK 3.3: The requirement that 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 $\mathbb{R}^d$ such that c satisfies the smoothness requirement in Theorem 3.1 on that rectangle.

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

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

4 Asymptotic Validity of Inference

4.1 Assumptions

We posit that P, the distribution of the observed data, belongs to a class of distributions denoted by $\mathcal{P}$. We write stochastic order relations that hold uniformly over $P \in \mathcal{P}$ using the notations $o_{\mathcal{P}}$ and $O_{\mathcal{P}}$; see Online Appendix D.1 for the formal definitions. Below, 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 4.1: *(a) $\Theta \subset \mathbb{R}^d$ is a compact hyperrectangle with nonempty interior. (b) All distributions $P \in \mathcal{P}$ satisfy the following:*

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

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

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

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

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

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

ASSUMPTION 4.2: *The function* 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 4.3-(II) is imposed,* $\kappa_n = o(n^{1/4})$.

Assumption 4.1-(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 (2.8). Assumption 4.1-(b) defines our moment (in)equalities model. Assumption 4.2 constrains the GMS function and the rate at which its tuning parameter diverges. Both 4.1-(b) and 4.2 are based on Andrews and Soares (2010) and are standard in the literature,[21] although typically with $\kappa_n = o(n^{1/2})$. The slower rate $\kappa_n = o(n^{1/4})$ is satisfied for the popular choice, recommended by Andrews and Soares (2010), of $\kappa_n = \sqrt{\ln n}$.

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

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

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

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

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

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

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

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

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

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

one has

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

Finally,

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

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

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

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

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

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

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

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

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

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

we replace $\mathbb{G}^b_{n,j+R_1}(\theta)$ with $-\mathbb{G}^b_{n,j}(\theta)$ and $\hat{D}_{n,j+R_1}(\theta)$ with $-\hat{D}_{n,j}(\theta)$, so that inequality $\mathbb{G}^b_{n,j+R_1}(\theta) + \hat{D}_{n,j+R_1}(\theta)\lambda \leqslant c$ is replaced with $-\mathbb{G}^b_{n,j}(\theta) - \hat{D}_{n,j}(\theta)\lambda \leqslant c$ in equation (3.1). In words, when hard threshold GMS indicates that both paired inequalities bind, we pick one of them, treat it as an equality, and drop the other one. In the proof of Theorem 4.1, we show that this tightens the stochastic program.[23] The rest of the procedure is unchanged.

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

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

[23] When paired inequalities are present, in equation (2.5) instead of $\hat{\sigma}_{n,j}$ we use the estimator $\hat{\sigma}^M_{n,j}$ specified in (E.188) in Lemma E.10 p.51 of the Online Appendix for $\sigma_{P,j}, j = 1, \ldots, 2R_1$ (with $R_1 \leqslant J_1/2$ defined in the assumption). In equation (3.2) we use $\hat{\sigma}_{n,j}$ for all $j = 1, \ldots, J$. To ease notation, we do not distinguish the two unless it is needed.

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

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

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

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

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

(i) For each j, there exist $D_{P,j}(\cdot) \equiv \nabla_\theta\{E_P[m_j(X,\cdot)]/\sigma_{P,j}(\cdot)\}$ and its estimator $\hat{D}_{n,j}(\cdot)$ such that $\sup_{\theta\in\Theta^\epsilon} \|\hat{D}_{n,j}(\theta) - D_{P,j}(\theta)\| = o_{\mathcal{P}}(1)$.

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

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

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

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

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

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

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

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

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

(iii) Q_P satisfies

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

[26] The requirements are imposed on Θ^ϵ. Under Assumption 4.3-(I) it suffices they hold on Θ.

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

4.2 Theoretical Results

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

The following theorem establishes the asymptotic validity of the proposed confidence interval $CI_n \equiv [-s(-p, \mathcal{C}_n(\hat{c}_n)), s(p, \mathcal{C}_n(\hat{c}_n))]$, where $s(p, \mathcal{C}_n(\hat{c}_n))$ was defined in equation (2.5) and $\hat{c}_n$ in (3.5).

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

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

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

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

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

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

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

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

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

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

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

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

Second set of results: Simplifications for special cases.

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

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

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

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

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

we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.3 Role of the ρ-box Constraints and Heuristics for Choosing ρ

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

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

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

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

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

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

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

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

5 Monte Carlo Simulations

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

5.1 The General Entry Game Model

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

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

Here, Y_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

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

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

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

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

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

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

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

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

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

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

5.2 Specific Implementations and Results

Set 1: A comparison with BCS-Profiling

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

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

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

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

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

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

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

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

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

Set 2: Heterogeneous interaction effects and potentially correlated errors

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

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

[40] Based on some trial runs of BCS-profiling for δ_1, we estimate that running it with $B = 1001$ throughout would take 3.14-times longer than the computation times reported in Table 1. By comparison, calibrated projection takes only 1.75-times longer when implemented with $B = 1001$ instead of $B = 301$.

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

In our Set 2-DGP1, we set $\zeta_1 = (.5, .25)'$, $\Delta_1 = (-1,-1)'$, and $r = 0$. Set 2-DGP2 differs by setting $\Delta_1 = (-1,-.75)'$. In both cases, $(\zeta_2, \Delta_2) = (\zeta_1, \Delta_1)$ and $\mu = 0.5$; we only report results for (ζ_1, Δ_1). Although parameter values are similar, there is a qualitative difference: In DGP1, parameters are point identified; in DGP2, they are not but the true bounds ($\zeta_1^{[1]} \in [0.405, 0.589]$, $\zeta_1^{[2]} \in [0.236, 0.266]$, $\Delta_1^{[1]} \in [-1.158, -0.832]$, $\Delta_1^{[2]} \in [-0.790, -0.716]$) are not wide compared to sampling uncertainty. We therefore expect all methods that use GMS to be conservative in DGP2.[41] In both Set 2-DGP1& DGP2 we use knowledge that $r = 0$, so that $d = 8$. Our Set 2-DGP3 preserves the same payoff parameters values as in Set 2-DGP2 but sets $r = 0.5$ and this parameter is also unknown, so that $d = 9$.

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

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

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

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

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

6 Conclusions

This paper introduces a computationally attractive confidence interval for linear functions of parameter vectors that are partially identified through finitely many moment (in)equalities. The extreme points of our *calibrated projection* confidence interval are obtained by minimizing and maximizing $p'\theta$ subject to properly relaxed sample analogs of the moment conditions. The relaxation amount, or critical level, is computed to insure uniform asymptotic coverage of $p'\theta$ rather than 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 also computationally attractive thanks to an application, novel to this paper, of the response surface method for global optimization that is of independent interest in the partial identification literature. Indeed, a key contribution of the paper is to establish convergence of this algorithm.

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

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

A Convergence of the E-A-M Algorithm

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

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

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

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

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

where $\beta = (\beta_1, \cdots, \beta_d) \in \mathbb{R}^d$ controls the length-scales of the process. Two values $c(\theta)$ and $c(\theta')$ are highly correlated when $\theta_k - \theta'_k$ is small relative to 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 $k = 1, \cdots, d$. We let $\bar{\beta} = (\bar{\beta}_1, \cdots, \bar{\beta}_d)' \in \mathbb{R}^d$. Specific suggestions on the forms of K_β are given in Appendix B.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):

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

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

Given this, the expected improvement function can be written as

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

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

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

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

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

A.1 Proof of Theorem 3.1

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

Below, we let $L \geqslant 2k$. Let $0 < \nu < \infty$. Let $0 < \eta < \epsilon$ and $A_L \in \mathcal{F}$ be the event that at least $\lfloor \eta L \rfloor$ of the points $\theta^{(k+1)}, \cdots, \theta^{(L)}$ are drawn independently from a uniform distribution on 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:

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

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

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

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

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

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

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

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

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

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

optimal solution to (3.6). [**HK: Below I changed all θ^*_ℓ to $\theta^{*,\ell}$.**] Then,

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

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

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

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

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

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

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

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

Let Z_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),

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

Further, by the definition of B_L,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Tables

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

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

Notes: (1) Projections of 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.

[31]

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

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

Notes: Same DGP and conventions as in Table 1.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References

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

Andrews, D. W. K., S. T. Berry, and P. Jia (2004): "Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location," mimeo.

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

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

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

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

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set Identified Linear Models," *Econometrica*, 80, 1129–1155.

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

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

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

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

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

Chen, X., E. Tamer, and A. Torgovitsky (2011): "Sensitivity Analysis in Semiparametric Likelihood Models," Working paper.

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

Ciliberto, F., and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77, 1791–1828.

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

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

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

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

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

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

Jones, D. R. (2001): "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *Journal of Global Optimization*, 21(4), 345–383.

Jones, D. R., M. Schonlau, and W. J. Welch (1998): "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, 13(4), 455–492.

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

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

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

Kitagawa, T. (2012): "Inference and Decision for Set Identified Parameters Using Posterior Lower Probabilities," CeMMAP Working Paper.

Kline, B., and E. Tamer (2015): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, forthcoming.

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

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

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

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

Pakes, A., J. Porter, K. Ho, and J. Ishii (2011): "Moment Inequalities and Their Application," Discussion Paper, Harvard University.

——— (2015): "Moment Inequalities and Their Application," *Econometrica*, 83, 315334.

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

Romano, J. P., and A. M. Shaikh (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, 138, 2786–2807.

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

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

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

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