Confidence Intervals for Projections of Partially Identified Parameters*

Hiroaki Kaido† Francesca Molinari‡ Jörg Stoye§

February 21, 2019

Abstract

We propose a bootstrap-based *calibrated projection* procedure to build confidence intervals for single components and for smooth functions of a partially identified parameter vector in moment (in)equality models. The method controls asymptotic coverage uniformly over a large class of data generating processes. The extreme points of the calibrated projection confidence interval are obtained by extremizing the value of the function of interest subject to a proper relaxation of studentized sample analogs of the moment (in)equality conditions. The degree of relaxation, or critical level, is calibrated so that the function of θ, not θ itself, is uniformly asymptotically covered with prespecified probability. This calibration is based on repeatedly checking feasibility of linear programming problems, rendering it computationally attractive.

Nonetheless, the program defining an extreme point of the confidence interval is generally nonlinear and potentially intricate. We provide an algorithm, based on the response surface method for global optimization, that approximates the solution rapidly and accurately, and we establish its rate of convergence. The algorithm is of independent interest for optimization problems with simple objectives and complicated constraints. An empirical application estimating an entry game illustrates the usefulness of the method. Monte Carlo simulations confirm the accuracy of the solution algorithm, the good statistical as well as computational performance of calibrated projection (including in comparison to other methods), and the algorithm's potential to greatly accelerate computation of other confidence intervals.

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

*We are grateful to Elie Tamer and three anonymous reviewers for very useful suggestions that substantially improved the paper. We thank for their comments Ivan Canay and seminar and conference participants at Amsterdam, Bonn, BC/BU joint workshop, Brown, Cambridge, Chicago, Cologne, Columbia, Cornell, CREST, Duke, ECARES, Harvard/MIT, Kiel, Kobe, Luxembourg, Mannheim, Maryland, Michigan, Michigan State, NUS, NYU, Penn, Penn State, Rochester, Royal Holloway, SMU, Syracuse, Toronto, Toulouse, UCL, UCLA, UCSD, Vanderbilt, Vienna, Yale, Western, and Wisconsin as well as CEME, Cornell-Penn State IO/Econometrics 2015 Conference, ES Asia Meeting 2016, ES European Summer Meeting 2017, ES North American Winter Meeting 2015, ES World Congress 2015, Frontiers of Theoretical Econometrics Conference (Konstanz), KEA-KAEA International Conference, Notre Dame Second Econometrics Workshop, Verein für Socialpolitik Ausschuss für Ökonometrie 2017. We are grateful to Undral Byambadalai, Zhonghao Fu, Debi Mohapatra, Sida Peng, Talal Rahim, Matthew Thirkettle, and Yi Zhang for excellent research assistance. A MATLAB package implementing the method proposed in this paper, Kaido, Molinari, Stoye, and Thirkettle (2017), is available at https://molinari.economics.cornell.edu/programs/KMSportable_V3.zip. We are especially grateful to Matthew Thirkettle for his contributions to this package. We gratefully acknowledge financial support through NSF grants SES-1230071 and SES-1824344 (Kaido), SES-0922330 and SES-1824375 (Molinari), and SES-1260980 and SES-1824375 (Stoye).

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

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

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

1 Introduction

This paper provides novel confidence intervals for projections and smooth functions of a parameter vector $\theta \in \Theta \subset \mathbb{R}^d$, $d < \infty$, that is partially or point identified through a finite number of moment (in)equalities. In addition, we develop a new algorithm for computing these confidence intervals and, more generally, for solving optimization problems with "black box" constraints, and obtain its rate of convergence.

Until recently, the rich literature on inference for moment (in)equalities focused on confidence sets for the entire vector θ, usually obtained by test inversion as

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

where the test statistic $T_n(\theta)$ aggregates violations of the sample analog of the moment (in)equalities and the critical value $c_{1-\alpha}(\theta)$ controls asymptotic coverage, often uniformly over a large class of data generating processes (DGPs). However, applied researchers are frequently interested in a specific component (or function) of θ, e.g., the returns to education. Even if not, they may simply want to report separate confidence intervals for components of a vector, as is standard practice in other contexts. Thus, consider inference on the projection $p'\theta$, where p is a known unit vector. To date, it is common to report as confidence set the corresponding projection of $\mathcal{C}_n(c_{1-\alpha})$ or the interval

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

which will miss any "gaps" in a disconnected projection but is much easier to compute. This approach yields asymptotically valid but typically conservative and therefore needlessly large confidence regions. The potential severity of this effect is easily appreciated in a point identified example. Given a $\sqrt{n}$-consistent estimator $\hat{\theta}_n \in \mathbb{R}^d$ with limiting covariance matrix equal to the identity matrix, the usual 95% confidence interval for θ_k equals $[\hat{\theta}_{n,k} - 1.96, \hat{\theta}_{n,k} + 1.96]$. Yet the analogy to CI_n^{proj} would be projection of a 95% confidence ellipsoid, which with $d = 10$ yields $[\hat{\theta}_{n,k} - 4.28, \hat{\theta}_{n,k} + 4.28]$ and a true coverage of essentially 1.

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

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

or, for computational simplicity and presentational convenience, the interval

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

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

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

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

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

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

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

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

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

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

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

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

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

2 Detailed Explanation of the Method

2.1 Setup and Definition of CI_n

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

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

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

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

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

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

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

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

with

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

and similarly for $(-p)$. Henceforth, to simplify notation, we write $\hat{c}_n$ for $\hat{c}_{n,1-\alpha}$. We also define $J \equiv J_1 + 2J_2$ moments, where $\bar{m}_{n,J_1+J_2+k}(\theta) = -\bar{m}_{J_1+k}(\theta)$ for $k = 1, \ldots, J_2$. That is, we treat moment equality constraints as two opposing inequality constraints.

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

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

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

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

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

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

depends on (i) the asymptotic behavior of the sample moments entering the inequality restrictions, which can change discontinuously depending on whether they bind at θ or not, and (ii) the local geometry of the identification region at θ, i.e. the shape of the constraint set formed by the moment restrictions. Features (i) and (ii) can be quite different at different points in $\Theta_I(P)$, making uniform inference challenging. In particular, (ii) does not arise if one only considers inference for the entire parameter vector, and hence is a new challenge requiring new methods.

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

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

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

Our goal is to set the probability of (2.8) equal to $1-\alpha$. To ease computation, we approximate (2.8) by linear expansion in λ of the constraint set. For each j, add and subtract $\sqrt{n}E_P[m_j(X_i, \theta+\lambda/\sqrt{n})]/\hat{\sigma}_{n,j}(\theta+\lambda/\sqrt{n})$ and apply the mean value theorem to obtain

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

Here $\mathbb{G}_{n,j}(\cdot) \equiv \sqrt{n}(\bar{m}_{n,j}(\cdot) - E_P[m_j(X_i,\cdot)])/\sigma_{P,j}(\cdot)$ is a normalized empirical process indexed by $\theta \in \Theta$, $D_{P,j}(\cdot) \equiv \nabla_\theta\{E_P[m_j(X_i,\cdot)]/\sigma_{P,j}(\cdot)\}$ is the gradient of the normalized moment, $\gamma_{1,P,j}(\cdot) \equiv E_P(m_j(X_i,\cdot))/\sigma_{P,j}(\cdot)$ is the studentized population moment, and the mean value $\bar{\theta}$ lies componentwise between θ and $\theta+\lambda/\sqrt{n}$.[5]

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

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

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

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

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

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

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

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

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

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

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

where P^* denotes the law of the random set $\Lambda_n^b(\theta, \rho, c)$ induced by the bootstrap sampling process, i.e. by the distribution of $(X_1^b, \ldots, X_n^b)$ conditional on the data. Expression (2.13) uses convexity of $\Lambda_n^b(\theta, \rho, c)$ and reveals that the probability inside curly brackets can be assessed by repeatedly checking feasibility of a linear program.[10] We describe in detail in Online Appendix D.4 how we compute $\hat{c}_n(\theta)$ through a root-finding algorithm.

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

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

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

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

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

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

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

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

2.3 Computation of CI_n and of Similar Confidence Intervals

Projection based methods as in (1.2) and (1.4) have nonlinear constraints involving a critical value which in general is an unknown function, with unknown gradient, of θ. Similar considerations often apply to critical values used to build confidence intervals for optimal values of optimization problems with estimated constraints. When the dimension of the parameter vector is large, directly solving optimization problems with such constraints can be expensive even if evaluating the critical value at each θ is cheap.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The advantages of E-A-M are as follows. First, we control the number of points at which we evaluate the critical value; recall that this evaluation is the expensive step. Also, the initial k evaluations can easily be parallelized. For any additional E-step, one needs to evaluate $c(\cdot)$ only at a single point $\theta^{(L+1)}$. The M-step is crucial for reducing the number of additional evaluation points. To determine the next evaluation point, it trades off "exploitation" (i.e. the benefit of drawing a point at which the optimal value is high) against "exploration" (i.e. the benefit of drawing a point in a region in which the approximation error of c is currently large) through maximizing expected improvement.[16] Finally, the algorithm simplifies the M-step by providing constraints and their gradients for program (2.21) in closed form, thus greatly aiding fast and stable numerical optimization. The price is the additional approximation step. In the empirical application in Section 4 and in the numerical exercises of Appendix C, this price turns out to be low.

2.4 Choice of Tuning Parameters

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

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

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

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

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

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

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

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

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

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

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

3 Theoretical Results

3.1 Asymptotic Validity of Inference

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

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

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

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

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

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

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

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

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

where $\hat{c}_n^f(\theta) \equiv \inf\{c \geqslant 0 : P^*(\Lambda_n^b(\theta,\rho,c) \cap \{\|\nabla_\theta f(\theta)\|^{-1}\nabla_\theta f(\theta)\lambda = 0\} \neq \varnothing) \geqslant 1-\alpha\}$. *Suppose that there exist* $\varpi > 0$ *and* $M < \infty$ *such that* $\inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} \|\nabla f(\theta)\| \geqslant \varpi$ *and* $\sup_{\theta,\bar{\theta}\in\Theta} \|\nabla f(\theta) - \nabla f(\bar{\theta})\| \leqslant M\|\theta - \bar{\theta}\|$*, where* $\nabla_\theta f(\theta)$ *is the gradient of* $f(\theta)$*.*[19] *Let* $0 < \alpha < 1/2$*. Then:*

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

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

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

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

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

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

3.2 Convergence of the E-A-M Algorithm

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

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

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

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

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

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

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

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

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

To apply Theorem 3.2 to calibrated projection, we provide low level conditions (Assumption D.1 in Online Appendix D.1.1) under which the map $\theta \mapsto \hat{c}_n(\theta)$ uniformly stochastically satisfies a Lipschitz-type condition. To get smoothness, we work with a mollified version of $\hat{c}_n$, denoted $\hat{c}_{n,\tau_n}$ in equation (D.1), where $\tau_n = o(n^{-1/2})$.[23] Theorem D.1 in the Online Appendix shows that $\hat{c}_n$ and $\hat{c}_{n,\tau_n}$ can be made uniformly arbitrarily close, and that $\hat{c}_{n,\tau_n}$ yields valid inference as in (3.1). In practice, we directly apply the E-A-M steps to $\hat{c}_n$.

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

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

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

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

4 Empirical Illustration: Estimating a Binary Game

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

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

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

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

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

of the parameter space.

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

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

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

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

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

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

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

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

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

5 Conclusion

This paper proposes a confidence interval for linear functions of parameter vectors that are partially identified through finitely many moment (in)equalities. The extreme points of our *calibrated projection* confidence interval are obtained by minimizing and maximizing $p'\theta$ subject to properly relaxed sample analogs of the moment conditions. The relaxation amount, or critical level, is computed to insure uniform asymptotic coverage of $p'\theta$ rather than θ itself. Its calibration is computationally attractive because it is based on repeatedly checking feasibility of (bootstrap) linear programming problems. Computation of the extreme points of the confidence intervals is furthermore attractive thanks to an application of the response surface method for global optimization; this is a novel contribution of independent interest. Indeed, one key result is a convergence rate for this algorithm when applied to constrained optimization problems in which the objective function is easy to evaluate but the constraints are "black box" functions. The result is applicable to any instance when the researcher wants to compute confidence intervals for optimal values of constrained optimization problems. Our empirical application and Monte Carlo analysis show that, in the DGPs that we considered, calibrated projection is fast and accurate, and also that the E-A-M algorithm can greatly improve computation of other confidence intervals.

References

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

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

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

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

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

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

Bugni, F. A., I. A. Canay, and X. Shi (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8(1), 1–38.

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

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

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

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

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

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

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

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

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

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

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

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

Kaido, H., F. Molinari, and J. Stoye (2017): "Confidence Intervals for Projections of Partially Identified Parameters," CeMMAP Working Paper CWP 49/17, available at https://www.cemmap.ac.uk/publication/id/10139.

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

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

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

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

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

——— (2015): "Moment Inequalities and Their Application," *Econometrica*, 83, 315–334.

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

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

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

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

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

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

A Convergence of the E-A-M Algorithm

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

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

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

Our algorithm iteratively determines evaluation points based on the *expected improvement* criterion (Jones, Schonlau, and Welch, 1998). For this, we formally introduce a model that describes the uncertainty associated with the values of c outside the current evaluation points. Specifically, the unknown function c is modeled as a Gaussian process such that[31]

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

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

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

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

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

Given this, the expected improvement function can be written as

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

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

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

in Section 2.3).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A.1 Proof of Theorem 3.2

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

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

Proof of Theorem 3.2. First, note that

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

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

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

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

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

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

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

For each $\omega \in D_L$, let

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Further, by the definition of B_L,

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

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

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

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

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

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

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

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

Therefore

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

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

A.2 Auxiliary Lemmas for the Proof of Theorem 3.2

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

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

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

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

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

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

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

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

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

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

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

By Assumption A.3,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

sufficiently large.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

for all L sufficiently large.

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

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

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

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

For evaluation points θ_L such that $|\bar{g}(\theta_L) - c(\theta_L)| > \varepsilon_L$, the following lemma is an analog of Lemma 8 in Bull (2011), which links the expected improvement to the actual improvement achieved by a new evaluation point θ.

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

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

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

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

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

A.6(i), we can write

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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

Similarly, for the lower bound, we have

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

Note that we may write

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

by $|\bar{g}(\theta_L)-c(\theta_L)|>\varepsilon_L$. Arguing as in (A.77) and noting that Φ is increasing, one has

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

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

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

where arguing as in (A.65),

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

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

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

by the mean value theorem, where $\tilde{\eta}_L$ is a point between $M_1\eta_L$ and $M_1\eta_L-M_2$. The claim of the lemma then follows from (A.77)-(A.72), and $(p'\theta_L - p'\theta_L^{*,L})$ being bounded because Θ is bounded. □

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

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

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

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

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

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

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

Similarly,

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

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

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

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

B Applying the E-A-M Algorithm to Profiling

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

C An Entry Game Model and Some Monte Carlo Simulations

We evaluate the statistical and numerical performance of calibrated projection and E-A-M in comparison with BCS-profiling in a Monte Carlo experiment run on a server with two Intel Xeon X5680 processors rated at 3.33GHz with 6 cores each and with a memory capacity of 24Gb rated at 1333MHz. The experiment simulates a two-player entry game in the Monte Carlo exercise of BCS, using their

code to implement their method.[36]

C.1 The General Entry Game Model

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

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

Here, Y_ℓ, Z_ℓ, and u_ℓ denote player ℓ's binary action, observed characteristics, and unobserved characteristics. The strategic interaction effects $Z_\ell'\Delta_\ell \leqslant 0$ measure the impact of the opponent's entry into the market. We let $X \equiv (Y_1, Y_2, Z_1', Z_2')'$. We generate $Z = (Z_1, Z_2)$ as an i.i.d. random vector taking values in a finite set whose distribution $p_z = P(Z = z)$ is known. We let $u = (u_1, u_2)$ be independent of Z and such that $Corr(u_1, u_2) \equiv r \in [0, 1]$ and $Var(u_\ell) = 1, \ell = 1, 2$. We let $\theta \equiv (\vartheta_1', \vartheta_2', \Delta_1', \Delta_2', r)'$. For a given set $A \subset \mathbb{R}^2$, we define $G_r(A) \equiv P(u \in A)$. We choose G_r so that the c.d.f. of u is continuous, differentiable, and has a bounded p.d.f. The outcome $Y = (Y_1, Y_2)$ results from pure strategy Nash equilibrium play. For some value of Z and u, the model predicts monopoly outcomes $Y = (0, 1)$ and $(1, 0)$ as multiple equilibria. When this occurs, we select outcome $(0, 1)$ by independent Bernoulli trials with parameter $\mu \in [0, 1]$. This gives rise to the following restrictions:

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

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

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

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

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

C.2 A Comparison to BCS-Profiling

BCS specialize this model as follows. First, u_1, u_2 are independently uniformly distributed on $[0, 1]$ and the researcher knows $r = 0$. Equality (C.1) disappears because $(0, 0)$ is never an equilibrium. Next, $Z_1 = Z_2 = [1; \{W_k\}_{k=0}^{d_W}]$, where W_k are observed market type indi-

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

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

cators, $\Delta_\ell = [\delta_\ell; 0_{d_W}]$ for $\ell = 1, 2$, and $\vartheta_1 = \vartheta_2 = \vartheta = [0; \{\vartheta^{[k]}\}_{k=0}^{d_W}]$.[38] The parameter vector is $\theta = [\delta_1; \delta_2; \vartheta]$ with parameter space $\Theta = \{\theta \in \mathbb{R}^{2+d_W} : (\delta_1, \delta_2) \in [0,1]^2,\ \vartheta_k \in [0, \min\{\delta_1, \delta_2\}],\ k = 1, \ldots, d_W\}$. This leaves 4 moment equalities and 8 moment inequalities (so $J = 16$); compare equation (5.1) in BCS. We set $d_W = 3$, $P(W_k = 1) = 1/4, k = 0, 1, 2, 3$, $\theta = [0.4; 0.6; 0.1; 0.2; 0.3]$, and $\mu = 0.6$. The implied true bounds on parameters are $\delta_1 \in [0.3872, 0.4239]$, $\delta_2 \in [0.5834, 0.6084]$, $\vartheta^{[1]} \in [0.0996, 0.1006]$, $\vartheta^{[2]} \in [0.1994, 0.2010]$, and $\vartheta^{[3]} \in [0.2992, 0.3014]$.

The BCS-profiling confidence interval CI_n^{prof} inverts a test of $H_0 : p'\theta = \tau$ over a grid for τ. We do not in practice exhaust the grid but search inward from the extreme points of Θ in directions $\pm p$. At each τ that is visited, we use BCS code to compute a profiled test statistic and the corresponding critical value $\hat{c}_n^{MR}(\tau)$. The latter is a quantile of the minimum of two distinct bootstrap approximations, each of which solves a nonlinear program for each bootstrap draw. Computational cost quickly increases with grid resolution, bootstrap size, and the number of starting points used to solve the nonlinear programs.

Calibrated projection computes $\hat{c}_n(\theta)$ by solving a series of linear programs for each bootstrap draw.[39] It computes the extreme points of CI_n by solving the nonlinear program (2.6) twice, a task that is much accelerated by the E-A-M algorithm. Projection of Andrews and Soares (2010) operates very similarly but computes its critical value $\hat{c}_n^{proj}(\theta)$ through bootstrap simulation without any optimization.

We align grid resolution in BCS-profiling with the E-A-M algorithm's convergence threshold of 0.005.[40] We run all methods with $B = 301$ bootstrap draws, and calibrated and "uncalibrated" (i.e., based on Andrews and Soares (2010)) projection also with $B = 1001$.[41] Some other choices differ: BCS-profiling is implemented with their own choice to multi-start the nonlinear programs at 3 oracle starting points, i.e. using knowledge of the true DGP; our implementation of both other methods multi-starts the nonlinear programs from 30 data dependent random points (see Kaido, Molinari, Stoye, and Thirkettle (2017) for details).

Table 2 displays results for (δ_1, δ_2) and for 300 Monte Carlo repetitions of all three methods. All confidence intervals are conservative, reflecting the effect of GMS. As expected, uncalibrated projection is most conservative, with coverage of essentially 1. Also, BCS-profiling is more conservative than calibrated projection. The most striking contrast is in computational effort. Here, uncalibrated projection is fastest – indeed, in contrast to received

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

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

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

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

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

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

Tables

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

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

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

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

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

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

Notes: (1) Projections of Θ_I are: $\delta_1 \in [0.3872, 0.4239]$, $\delta_2 \in [0.5834, 0.6084]$, $\zeta_1 \in [0.0996, 0.1006]$, $\zeta_2 \in [0.1994, 0.2010]$, $\zeta_3 \in [0.2992, 0.3014]$. (2) "Upper" coverage is for $\max_{\theta \in \Theta_I(P)} p'\theta$, and similarly for "Lower". (3) "Average time" is computation time in seconds averaged over MC replications. (4) CI_n^{prof} results from BCS-profiling, CI_n is calibrated projection, and CI_n^{proj} is uncalibrated projection. (5) "Implementation" refers to the method used to compute the extreme points of the confidence interval.

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

	$1-\alpha$	Median CI		CI_n Coverage		CI_n^{proj} Coverage		Average Time	
		CI_n	CI_n^{proj}	Lower	Upper	Lower	Upper	CI_n	CI_n^{proj}
$\delta_1 = 0.4$	0.95	[0.333,0.478]	[0.288,0.555]	0.988	0.982	1	1	42.41	22.23
	0.90	[0.341,0.470]	[0.296,0.542]	0.976	0.957	1	1	41.56	22.11
	0.85	[0.346,0.464]	[0.302,0.534]	0.957	0.937	1	1	40.47	19.79
$\delta_2 = 0.6$	0.95	[0.525,0.653]	[0.466,0.683]	0.969	0.983	1	1	42.11	24.39
	0.90	[0.538,0.646]	[0.478,0.677]	0.947	0.960	1	1	40.15	28.13
	0.85	[0.545,0.642]	[0.485,0.672]	0.925	0.941	1	1	41.38	26.44
$\zeta^{[1]} = 0.1$	0.95	[0.054,0.142]	[0.020,0.180]	0.956	0.958	1	1	40.31	22.53
	0.90	[0.060,0.136]	[0.028,0.172]	0.911	0.911	1	1	36.80	24.15
	0.85	[0.064,0.132]	[0.032,0.167]	0.861	0.860	0.999	0.999	39.10	21.81
$\zeta^{[2]} = 0.2$	0.95	[0.156,0.245]	[0.121,0.281]	0.952	0.952	1	1	39.23	24.66
	0.90	[0.162,0.238]	[0.128,0.273]	0.914	0.910	0.998	0.998	41.53	21.66
	0.85	[0.165,0.234]	[0.133,0.268]	0.876	0.872	0.996	0.996	39.44	22.83
$\zeta^{[3]} = 0.3$	0.95	[0.257,0.344]	[0.222,0.379]	0.946	0.946	1	1	41.45	22.91
	0.90	[0.263,0.338]	[0.230,0.371]	0.910	0.909	0.997	0.999	42.09	22.83
	0.85	[0.267,0.334]	[0.235,0.366]	0.882	0.870	0.994	0.993	42.19	23.69

Notes: Same DGP and conventions as in Table 2.