Simple Inference for Constrained Optima[*]
Preliminary and Incomplete

Hiroaki Kaido[†] Francesca Molinari[‡] Jörg Stoye[§]

February 14, 2022

Abstract

We provide one-sided confidence regions for the *value* of optimization problems with estimated objectives and constraints. The motivating application is (one- or two-sided) inference on components of partially identified parameter vectors. The novelty of the method is that it is (i) very simple to execute while (ii) valid under reasonably general conditions.

Regarding (i), a main innovation is to compute confidence intervals locally at a solution to the optimization problem's sample analog. No global optimization is performed after this first optimization step, and all remaining optimizations are linear. Regarding (ii), we assume (and provide low-level sufficient conditions) that *any* solution to the sample problem is at most $O(n^{-1/2})$ away from *some* solution to the population problem. However, we allow for singleton feasible sets (i.e., point identification), low dimensional feasible sets, and maxima that are at least partially characterized by first-order conditions; all of these are excluded by at least some other approaches in the literature. We provide STATA and Python implementation packages. We illustrate the method's simplicity and power by re-evaluating the empirical findings in several published papers.

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

[*] We thank audiences at Bristol/Warwick, Johns Hopkins, and Penn State for their feedback. We gratefully acknowledge financial support through NSF grants SES-1230071 and SES-1824344 (Kaido), SES-1824375 (Molinari), 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 simple one-sided confidence intervals for the value γ^* of the optimization problem

$$\max g(\theta) \text{ s.t. } \theta \in \Theta_I, \tag{1.1}$$

where $g(\cdot)$ is a known, scalar-valued, smooth function and the feasible set Θ_I is characterized through moment inequalities that must be estimated.[1]

The motivating application is inference on components of partially identified paramater vectors, i.e. when Θ_I is an *identified set*. In this context, our method immediately yields two-sided confidence intervals that can be interpreted exactly like the confidence intervals for OLS which are reported by any regression package's output. To underscore the new method's simplicity, we provide implementation packages in STATA and Python. Potential applications of the method in the existing literature include Ho (2009), Ho, Ho, and Mortimer (2012), Ho and Pakes (2014), Holmes (2011), Kawai and Watanabe (2013), Kline and Tartari (2016), Lee (2013), and Wollmann (2018); we replicate several of these using the STATA command.

We are not the first to approach this question and have written on it before (Kaido, Molinari, and Stoye, 2019). We argue that our new proposal strikes an attractive balance between computational simplicity and general validity. Regarding simplicity, the new confidence interval is constructed locally around one (not necessarily unique) estimator $\hat{\theta}^*$ of the (not necessarily unique) optimal argument θ^* in (1.1). To do so, we bootstrap a local (to $\hat{\theta}^*$) linear approximation to (1.1). Conditionally on having computed $\hat{\theta}^*$ (as would typically be required to even get an estimator of γ^*), the computational cost of the entire procedure is negligible. Regarding generality, we impose more structure than in our previous work but allow for point identification and "near point identification," for feasible sets that have no interior, and for the solution to the (sample or population) optimization problem to be nonunique and/or be at least partially characterized by first-order conditions. These features are shared by some recent proposals (Andrews, Roth, and Pakes, 2021; Cox and Shi, 2020) but these put more structure on $g(\cdot)$. To the best of our knowledge, no other approach allows for all of these features; the pioneering approach in Pakes, Porter, Ho, and Ishii (2011), which is employed in the aforecited applications, allows for none of them.[2]

Two auxiliary contributions of independent interest are as follows. First, we must ensure that any solution of the empirical optimization problem is "$\sqrt{n}$-close" to some (not necessarily unique) solution of the population problem. To do so, we generalize the "Argmax Theorem" (van der Vaart and Wellner, 2000) to show that m-estimators of partially identified parameters

[1] It actually suffices for $g(\cdot)$ to be smooth and its gradient uniformly estimable at parametric rate. We will formalize this in future iterations.

[2] We reference the unpublished version of Pakes, Porter, Ho, and Ishii (2015) because only that version contains the estimation and inference strategy discussed here and used in much subsequent work.

achieve this particular notion of consistency under weak conditions. Second, we provide exact conditions under which the approach in Pakes, Porter, Ho, and Ishii (2011) can be justified. These are "morally" weaker than in the original work and also correct an oversight therein.

The remainder of this paper is structured as follows. Section 2 gives an algorithmic explanation of the inference procedure, introducing only the minimal amount of notation needed. Section 3 provides theoretical justification. Section 4 discusses the relation to the literature in more depth and provides theoretical justification for Pakes, Porter, Ho, and Ishii (2011). Section 5 contains several empirical applications, and Section 6 concludes.

2 Explanation of Method

This section gives further motivation and a precise description of the inference procedure. Consider a parameter $\theta \in \Theta \subset \boldsymbol{R}^k$ that is partially identified by moment inequalities:

$$\theta \in \Theta_I \equiv \big\{\theta \in \Theta : E(m_j(\theta, X_i)) \leqslant 0, j = 1, \ldots, J\big\}. \tag{2.1}$$

Here, $m_j(\theta, X_i)$ are known smooth functions, and we assume that corresponding sample averages $\overline{m}_j(\theta)$ are available. The setting allows for equalities as well, which will be identified with paired inequalities.

Inference on γ^* can be of interest for at least two reasons. First, $g(\cdot)$ might be an objective function of immediate substantive interest, for example the social welfare induced by some treatment or policy. Note, however, that in this context we only procide one-sided inference from above (or from below for the objective function's minimum).

Our motivating application is inference on scalar functions of partially identified parameter vectors. To this end, note that for any known function $g(\cdot)$, the partial identification scenario (2.1) induces partial identification of $\gamma \equiv g(\theta)$ through

$$\gamma \in \Gamma_I \equiv \{g(\theta) : \theta \in \Theta_I\}.$$

In many cases, researchers will be willing to restrict attention to the smallest interval containing Γ_I, that is, to

$$[\gamma_L, \gamma_U] \equiv [\min \Gamma_I, \max \Gamma_I].$$

This might be because Θ_I is known to be connected, in which case $\Gamma_I = [\gamma_L, \gamma_U]$; it might be because the researcher is mainly interested in the extreme values that γ can take, in which case it can be statistically advantageous to consider $[\gamma_L, \gamma_U]$;[3] or it might be just for computational reasons, since finding gaps in Γ_I can be hard.

[3] If Γ_I is a finite union of singletons, e.g. because one has moment equalities with non-unique locally identifiable solutions, then a confidence region for Γ_I will locally resemble two-sided inference on γ_U, even though one-sided inference would do if γ_L is sufficiently below γ_U. Our approach adaptively implements this.

We define estimators of $[\gamma_L, \gamma_U]$ as

$$\begin{aligned}
\hat{\gamma_L} &\equiv \min_{\theta \in \hat{\Theta}_I} g(\theta) \\
\hat{\gamma_U} &\equiv \max_{\theta \in \hat{\Theta}_I} g(\theta) \\
\hat{\Theta}_I &\equiv \arg\min_{\theta \in \Theta} \max \left\{ \max_{j=1,\ldots,J} \overline{m}_j(\theta), 0 \right\},
\end{aligned}$$

where $\overline{M}_j(\cdot)$ is a sample average. Note that the definition of $\hat{\Theta}_I$ ensures its nonemptiness, whereas $\{\theta : \max_{j=1,\ldots,J} \overline{m}_j(\theta) \leqslant 0\}$ may be empty. An instance of particular interest, which the reader may want to keep in mind, is that γ is simply a component of θ.

Consider now the one-sided left-unbounded confidence interval for γ. Following Imbens and Manski (2004) and most of the subsequent literature, we want this interval to accurately cover γ uniformly over its possible values. That is,

$$\min_{\theta \in \Theta_I} \Pr(\gamma \in CI) \geqslant 1 - \alpha$$

at least in an asymptotic sense.[4] For left-unbounded CI's, it is clear that this probability decreases in γ, so that the one-sided CI is in practice a CI for γ_U. Similarly, the one-sided right-unbounded CI is in practice a CI for γ_L. Two-sided $(1-\alpha)$-CI's for γ will be constructed by intersecting one-sided $(1-\alpha/2)$-CI's.[5] Thus, inference on $g(\theta)$ reduces to inference on maximization problem (1.1).

The algorithm for computing a one-sided $(1-\alpha)$-CI is as follows:[6]

1. Compute $\hat{\gamma}_U$ and pick an arbitrary $\hat{\theta}^* \in \arg\max_{\theta \in \hat{\Theta}_I} g(\theta)$.

   (This step, i.e. estimating γ_U, will typically be the hardest.)

2. Define the index set

   $$\mathcal{J}^* \equiv \left\{ j \in \{1, \ldots, J\} : \overline{m}_j(\hat{\theta}^*)/\hat{\sigma}_j(\hat{\theta}^*) \leqslant \sqrt{\log(n)/n} \right\},$$

   where $\hat{\sigma}_j(\theta)$ is an estimator of $\sigma_j(\theta)$, the standard deviation of moment condition j at θ.

   (Intuitively, we henceforth restrict attention to constraints that are plausibly binding at $\hat{\theta}^*$. Note that equality constraints automatically pass this test.)

[4] Our theoretical justification also establishes uniformity of size control over a large set of true data generating processes. We omit this uniformity here for simplicity of notation.

[5] As is typical in this literature, the resultant interval can be empty, namely if the data suggest misspecification of the model. See Ponomareva and Tamer (2011), Andrews and Kwon (2019), and Stoye (2020) for recent contributions to this conversation. We leave connecting it to the present method to future research.

[6] For simplicity, we impute simple values for some tuning parameters. These values are commonly used in the related literature and are the ones that we later implement.

3. For each $j \in \mathcal{J}^*$, compute $\hat{D}_j(\hat{\theta}^*)$, an estimator of the moment condition's gradient at $\hat{\theta}^*$
$$D_j(\hat{\theta}^*) \equiv \nabla_\theta E(m_j(X_i, \hat{\theta}^*)/\sigma_j(\hat{\theta}^*)).$$
(There are many such estimators; our implementation provides one.)

4. Implement the following linear program parameterized by scalar c and vector $\boldsymbol{\mu} \equiv (\mu_j^b)_{j \in \mathcal{J}^*}$, where each μ_j is scalar.
$$\begin{aligned} \psi(c, \boldsymbol{\mu}) \quad &\equiv \quad \max_{\vartheta \in \boldsymbol{R}^k} \nabla_\theta g(\hat{\theta}^*)'\vartheta \\ \text{s.t.} \quad & \quad \hat{D}_j(\hat{\theta}^*)'\vartheta - \sqrt{n}\mu_j \leqslant c, j \in \mathcal{J}^* \\ & \quad -\rho \leqslant e_j'\vartheta \leqslant \rho, j = 1, \ldots, k, \end{aligned}$$
where $(e_1, \ldots, e_k)$ is an orthonormal basis of $\boldsymbol{R}^k$ s.t. $e_1 = \nabla_\theta g(\hat{\theta}^*)/\|\nabla_\theta g(\hat{\theta}^*)\|$.

   (Intutitively, we replace the optimization problem with a linear approximation that is uniformly valid in a neighborhood of $\hat{\theta}^*$. The parameters of the program are a relaxation c that we will calibrate to insure coverage and a slackness vector $\boldsymbol{\mu}$ that will simulate sampling uncertainty. The second set of constraints restricts ϑ to a hypercube over which the linear approximation is uniformly valid. They are governed by a tuning parameter $\rho > 0$ for which we give a suggestion. Note that, if γ is a component of θ, then one can take $(e_1, \ldots, e_k)$ to be the canonical basis.)

5. Let $\hat{c}$ be the smallest value of c s.t.
$$\Pr(\psi(c, \boldsymbol{\mu}^b) \geqslant 0) \geqslant 1 - \alpha,$$
where
$$\mu_j^b = \frac{\overline{m}_j^b(\hat{\theta}^*) - \overline{m}_j(\hat{\theta}^*)}{\hat{\sigma}_j(\hat{\theta}^*)}$$
and $\left(\overline{m}_j^b(\hat{\theta}^*)\right)_{j \in \mathcal{J}^*}$ is is an i.i.d. nonparametric bootstrap resample of $(\overline{m}_j(\hat{\theta}^*))_{j \in \mathcal{J}^*}$.

   (Intuitively, in our bootstrap approximation, this is by how much we would have to relax sample constraints so that the relaxed maximization problem covers the bootstrap population problem's true value –which equals 0 due to recentering– with the desired probability.)

6. The confidence interval equals
$$CI_\alpha \quad = \quad \left(-\infty, \hat{\gamma}_U + \psi(\hat{c}, \mathbf{0})/\sqrt{n}\right],$$
where $\mathbf{0}$ is a vector of zeros.

(Intuitively, we report the value of the sample optimization problem but relaxed by $\hat{c}$. An additional flourish is that, to avoid a second global optimization, we apply our local linear approximation to this last step. As the notation clarifies, from a computational point of view this renders it equivalent to one additional bootstrap iteration.)

Remark 2.1: We kept things as simple as possible by not studentizing constraints in the definition of $\hat{\Theta}_I$. We can do this because we will only require a limited notion (to be formalized later) of $\sqrt{n}$-consistency of $\hat{\theta}^*$. Not studentizing constraints can be attractive because, for example, studentization can turn linear constraints into nonlinear ones. That said, one might conjecture that studentization yields more efficient estimators. We provide the option to studentize at this stage and recommend it if computationally feasible.

Note, however, that local linearization allows us to studentize moment inequalities for inference purposes. This is reflected in the definition of μ_j^b and adds negligible computational burden even if solving the global problem with studentized constraints would be hard. See, in particular, Andrews and Soares (2010) for a discussion of why studentization is advisable for inference.

Remark 2.2: While we defer formal discussion to Section 3, we will now briefly clarify relation to our previous work (Kaido, Molinari, and Stoye, 2019). In that work, $\hat{c}$ is (at least in principle) computed at every possible value of θ. The global optimization problem is then revisited and solved subject to constraints that are relaxed by $\hat{c}(\theta)$; that is, the relaxation itself changes with θ. While we provide novel algorithms that improve computation and also supply a MATLAB implementation (Kaido, Molinari, Stoye, and Thirkettle, 2017), this is computationally involved. We here avoid it, albeit at the price of additional assumptions.

For an intuition of what changes and what kind of assumptions can motivate it, think of Kaido, Molinari, and Stoye (2019) as providing a test of the null hypothesis that, for a given θ, $g(\theta) \in [\gamma_L, \gamma_U]$. This test is then inverted, which is involved because its critical value depends on θ. In this analogy, we here compute the critical value exactly once, namely at $\hat{\theta}^*$. Intuitively, this is justified if (i) $\hat{\theta}^*$ is close to a true solution to the problem, (ii) on a vanishing neighborhood of $\hat{\theta}^*$ that (with high probability) includes said true solution, the test statistic whose quantiles we implicitly compute is asymptotically pivotal. The assumptions that we will impose beyond our previous work effectively impose (i); (ii) then turns out to be implied.

3 Detailed Justification of Method

This section first discusses "background assumptions" that we lift from our own previous work and that are also common in the literature, then discusses novel assumptions that we introduce, then provides a theorem justifying the new approach. For readability and comparability, we use notation that has become standard in the partial identification literature;

however, the substantive interpretation of Θ_I as identified set plays no role in the technical development.

3.1 Background Assumptions

The following assumptions are exactly as in Kaido, Molinari, and Stoye (2019) and are also closely related to others in the literature. We give a brief discussion after stating them. All assumptions refer to the set $\mathcal{P}$ of data generating processes over which uniformity is claimed.

Because this paper is rooted in the literature on partial identification and, in particular, moment inequalities, we will verbally refer to Θ_I as identified set characterized through moment inequalities. However, this is just a convenient interpretation; all that matters for the results is that the feasible set for some optimization problem can be described in the way that Θ_I is described here.

Assumption 3.1: *(a) $\Theta \subset \mathbb{R}^d$ is a compact hyperrectangle with nonempty interior.*
(b) All distributions $P \in \mathcal{P}$ satisfy the following:

(i) $\Theta_I \equiv \{\theta \in \Theta : E_P[m_j(X_i,\theta)] \leqslant 0,\ j = 1,\dots,J\} \neq \varnothing$.

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

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

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

Assumption 3.2: *All distributions $P \in \mathcal{P}$ satisfy* ***one*** *of the following two conditions for some constants $\omega > 0, \underline{\sigma} > 0, \epsilon > 0, \varepsilon > 0, M < \infty$:*

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

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

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

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

$$m_{j+R_1}(X_i,\theta) = -m_j(X_i,\theta) - t_j(X_i,\theta). \tag{3.1}$$

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

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

one has

$$\inf_{\theta \in \Theta_I(P)} \mathrm{eig}(\tilde{\Omega}_P(\theta)) \geqslant \omega. \tag{3.2}$$

Finally,

$$\inf_{\theta \in \Theta_I(P)} \sigma_{P,j}(\theta) > \underline{\sigma} \textit{ for } j = 1,\ldots,R_1. \tag{3.3}$$

Assumption 3.1 goes back to Andrews and Soares (2010). It clarifies that the parameter of interest θ is partially identified through finitely many moment conditions which can be individually regularly estimated. Importantly, the shape of Θ_I is not otherwise constrained at this point and will only be minimally constrained later.

Assumption 3.2 restricts the correlation between moment conditions. In particular, it excludes the possibility that some moment conditions are (almost) perfectly correlated *without the econometrician knowing this.* If this happens, the optimization problem's value mey be estimated superconsistently, i.e. at a rate faster than $O(n^{-1/2})$; this may sound like a good thing, but it invalidates the nonparametric bootstrap at the heart of our approach.[7] Note that the assumption allows for moment conditions that are perfectly correlated with the researcher's knowledge, e.g., an equality constraint that is entered as two "opposing" inequalities or (in certain applications to partial identification) interval data with fixed interval width.

Finally, a fully general statement of the procedure involves tuning parameters (φ_j, κ_n) governing Generalized Moment Selection (GMS; see Andrews and Soares (2010) and also Bugni (2010), Canay (2010), and Stoye (2009)). Our treatment of these enitrely follows the previous literature.

Assumption 3.3: *The function* φ_j *(whose use will be explained later) is continuous at all* $x \geqslant 0$ *and* $\varphi_j(0) = 0$; $\kappa_n \to \infty$ *and* $\kappa_n = o(n^{1/2})$. *If Assumption 3.2-2 is imposed,* $\kappa_n = o(n^{1/4})$.

3.2 Additional Assumption and High-Level Theorem

Our only additional assumption is that $\hat{\theta}^*$ must be $\sqrt{n}$-close to some solution to the true maximization problem. It is not necessary that either the sample or the population solu-

[7] The aforementioned oversight in PPHI is to not impose this or a similar assumption.

tion are unique, and it is also not required that the sample solutions converge to a specific true solution (or, indeed, anywhere, though of course all accumulation points must be true solutions). To state this formally, for closed $\Theta^* \subseteq \Theta$ we define

$$S(g, \Theta^*) = \arg\max_{\theta \in \Theta^*} g(\theta).$$

The notation is a reminder that, if g is a projection, then $S(\cdot)$ is the support set.) Then we have:

Assumption 3.4:

$$\max_{\hat{\theta}^* \in S(g, \hat{\Theta}_I)} \min_{\theta^* \in S(g, \Theta_I)} \|\hat{\theta}^* - \theta^*\| = O_{\mathcal{P}}(n^{-1/2}). \tag{3.4}$$

Theorem 3.1: *Suppose that Assumptions 3.1-3.2 and also 3.4 hold. Then Simple Calibrated projection is valid.*

Assumption 3.4 restricts *directed* Hausdorff distance between the estimated solution set $S(g, \hat{\Theta}_I)$ and the true solution set $S(g, \Theta_I)$: The former must be asymptotically contained in the latter but not necessarily conversely. This directed notion suffices because it implies that any selection $\hat{\theta}^* \in S(g, \hat{\Theta}_I)$ is asymptotically close to some true solution to the problem. Sufficiency of this directed notion of convergence is important for two reasons: First, it is much easier to verify this directed notion of convergence; indeed, if Assumption 3.4 were stated using Hausdorff distance, it would fail in many cases of interest. Second, while we technically define $\hat{\theta}^*$ as arbitrary selection from $S(p, \hat{\Theta}_I)$, computing it does not require one to compute $S(p, \hat{\Theta}_I)$ but only to find some element of it, which can be the considerably easier task.

3.3 Low-Level Conditions Justifying Simple Calibrated Projection

We next derive Assumption 3.4 from lower level conditions on $\mathcal{P}$. The first step is a novel result that clarifies "inner" (in the sense of directed Hausdorff distance) $\sqrt{n}$-consistency of m-estimators under partial identification.

3.3.1 A Rate Result for Set-Identified M-Estimators

Consider the standard "m-estimation" setup with[8]

$$\begin{aligned}\Theta^* &= \arg\min_{\theta\in\Theta} Q(\theta),\\ \hat{\Theta}^* &= \arg\min_{\theta\in\Theta} Q_n(\theta),\end{aligned}$$

with the only nonstandard aspect that Θ^* will not be assumed to be a singleton. (Notation anticipates that Θ^* need not be Θ_I.) We are interested in conditions under which

$$\max_{\theta\in\hat{\Theta}^*} d(\theta,\Theta^*) = O_P(n^{-1/2}), \tag{3.5}$$

i.e. the estimator asymptotically hits the set but need not explore it (nor converge to any particular element). We will henceforth denote this notion of consistency by *inner consistency.*

Such inner consistency (but without a rate) is implied by the usual consistency conditions less uniqueness. That is, if Θ^* is a well-separated minimum of $Q(\cdot)$, then $\hat{\Theta}^*$ is asymptotically contained in it. This is briefly mentioned in Newey and McFadden (1994), and the basic insight goes back at least to Redner (1981). However, we need a rate at which this "inner consistency" is assured. To this purpose, we provide the following, novel result.

THEOREM 3.2: *Suppose that $d(\hat{\theta}^*,\Theta^*) \xrightarrow{p} 0$ (i.e., inner consistency but without a rate) and that*

1. $Q(\theta) \gtrsim d^2(\theta,\Theta^*)$,
2. $\exists\epsilon>0\forall\delta\leqslant\epsilon : E\big(\sup_{\theta\in\Theta,\theta^*\in\Theta^*:\|\theta-\theta^*\|\leqslant\delta}|\nu_n(\theta)-\nu_n(\theta^*)|\big)\lesssim\delta.$

Then (3.5) holds.

Proof. Define $\mathcal{S}_{j,n}\equiv\{\theta\in\Theta : 2^{j-1}\leqslant\sqrt{n}d(\theta,\Theta^*)\leqslant 2^j\}$ and write

$$\begin{aligned}\Pr(\sqrt{n}d(\hat{\theta}^*,\Theta^*)>2^M) &\leqslant \sum_{j\geqslant M,2^j\leqslant\eta\sqrt{n}}\Pr\left(\inf_{\theta\in\mathcal{S}_{j,n}}Q_n(\theta)\leqslant\inf_{\theta^*\in\Theta^*}Q_n(\theta^*)\right)+\Pr\big(2d(\hat{\theta}^*,\Theta^*)\geqslant\eta\big)\\ &= \sum_{j\geqslant M,2^j\leqslant\eta\sqrt{n}}\Pr\left(\inf_{\theta\in\mathcal{S}_{j,n}}Q_n(\theta)\leqslant\inf_{\theta^*\in\Theta^*}Q_n(\theta^*)\right)+o_P(1), \qquad (3.6)\end{aligned}$$

[8] This is also sometimes called extremum estimation, with m-estimation referring to the special case where $Q_n(\cdot)$ is a sample average.

using that $d(\hat{\theta}^*, \Theta^*) \xrightarrow{p} 0$. Next, observe that uniformly for θ on any fixed $\mathcal{S}_{j,n}$, we have

$$Q(\theta) \gtrsim d^2(\theta, \Theta^*) \geqslant \frac{2^{2j-2}}{n},$$

where the first step is by assumption and the second one uses the definition of $\mathcal{S}_{j,n}$. For any $\theta \in \Theta$, define $\text{proj}_\theta \equiv \arg\min_{\theta^* \in \Theta^*} \|\theta - \theta^*\|$, the closest parameter value to θ that lies in Θ^* (with arbitrary selection if the arg min is not unique). We can then write

$$\begin{aligned}
&\Pr\left(\inf_{\theta \in \mathcal{S}_{j,n}} Q_n(\theta) - \inf_{\theta^* \in \Theta^*} Q_n(\theta^*) \leqslant 0\right) \\
\overset{(1)}{\leqslant}\; &\Pr\left(\inf_{\theta \in \mathcal{S}_{j,n}} (Q_n(\theta) - Q_n(\text{proj}_\theta)) \leqslant 0\right) \\
\overset{(2)}{\leqslant}\; &\Pr\left(\inf_{\theta \in \mathcal{S}_{j,n}} \{(Q_n - Q)(\theta) - (Q_n - Q)(\text{proj}_\theta)\} \leqslant -\frac{2^{2j-2}}{n}\right) \\
\overset{(3)}{=}\; &\Pr\left(\inf_{\theta \in \mathcal{S}_{j,n}} \{\nu_n(\theta) - \nu_n(\text{proj}_\theta)\} \leqslant -\frac{2^{2j-2}}{\sqrt{n}}\right) \\
\overset{(4)}{\leqslant}\; &\Pr\left(\sup_{\theta \in \mathcal{S}_{j,n}} |\nu_n(\theta) - \nu_n(\text{proj}_\theta)| \geqslant \frac{2^{2j-2}}{\sqrt{n}}\right) \\
\overset{(5)}{\leqslant}\; &\frac{\sqrt{n}}{2^{2j-2}} \cdot \frac{2^j}{\sqrt{n}} = 2^{2-j},
\end{aligned}$$

where (1) holds because the term inside the inf became smaller for each θ; (2) uses that $Q(\text{proj}_\theta) = 0$, whereas $Q(\theta) \geqslant 2^{2j-2}/n$; (3) plugs in the definition of ν_n; (4) is elementary; (5) uses Markov's inequality, the definition of $\mathcal{S}_{j,n}$, and the second condition, keeping in mind that $\|\theta - \text{proj}_\theta\| = d(\theta, \Theta^*)$. Conclude that the sum in (3.6) vanishes as $M \to \infty$, hence that $\sqrt{n} d(\hat{\theta}^*, \Theta^*) = O_P(1)$. □

Both the statament and the proof of this result are inspired, and heavily owe to, van der Vaart and Wellner (2000, Theorem 3.2.5.). The main difference is that we consider point-set distance $d(\theta, \Theta^*)$. This necessitates some modifications as the modulus of continuity must be invoked comparing a generic θ to the nearest element proj_θ of Θ^*, taking care of the fact that proj_θ is a function of θ. This requires reformulating condition 2 (in particular, note the supremum also being over $\theta^* \in \Theta^*$) and adding a few additional steps (e.g., compare (1) above).

3.3.2 Connection to Inference on Maxima

We next apply this result to inner consistent estimation of the support set. For this purpose, we identify $S(g, \Theta_I)$ as the "identified set" Θ^* in the extension of our original model where the condition

$$g(\theta) = \gamma_U$$

with sample analog

$$g(\theta) = \hat{\gamma}_U$$

is added. Note that the new condition is not a moment restriction, and we will not, for example, be able to verify pointwise asymptotic normality of the corresponding error process. However, all we need is to verify Assumption 3.4. We do this by imposing:

Assumption 3.5: *We have that:*

1. $\sqrt{n}(\hat{\gamma}_U - \gamma_U) = O_P(1)$.
2. *For any* $\theta^* \in S(g, \Theta_I)$, *let* $H(\theta^*) \equiv \{\theta^*\} \bigoplus \{\theta \in \Theta : \nabla_\theta g(\theta^*)'\theta = 0\}$ *be the corresponding supporting hyperplane of* θ_I *in direction* $\nabla_\theta g(\theta^*)$. *Note that, if* $g(\cdot)$ *is linear, then* $H(\theta^*)$ *is the same for all* $\theta^* \in S(g, \Theta_I)$.

   There exist $C, \epsilon > 0$ *s.t.* $\theta \in H(p, \Theta_I), \|\theta - \theta^*\| \leqslant \epsilon \implies d(\theta, \Theta_I) \geqslant C d^2(\theta, S(g, \Theta_I))$.

We then have:

Theorem 3.3: *The above assumption imply Assumption 3.4.*

Proof. As hinted at in text above. □

Part 1 of the condition will hold if it does so in the original moment inequalities model. We will derive it from lower level conditions later. The second condition forces the solution set to be a well-identified (if not necessarily unique) maximum of $g(\cdot)$ on Θ_I in the sense that Θ_I has nonvanishing curvature relative to the supporting hyperplane. For the salinet case where $g(\cdot)$ is linear and $H(\cdot)$ the usual supporting hyperplane, one can visualize violations of this assumption as follows: (i) Θ_I has a flat face that is almost orthogonal to the gradient of $g(\cdot)$; (ii) a smooth maximum with vanishing curvature, (iii) a distinct (possibly "far away") local maximum that attains close to the globally optimal value. In all of these cases, $\sqrt{n}$-consistency of $\hat{\theta}^*$ will generally fail, and in this sense the lower-level Assumption 3.5 appears rather tight. Along the same lines, note that issue (ii) is closely analogous to vanishing curvature of a likelihood function at its maximum, which would preclude $\sqrt{n}$-consistency of the Maximum Likelihood estimator.

3.3.3 Sufficient Conditions for $\sqrt{n}$-Consistency of $\hat{\gamma}_U$

We next provide an additional layer by giving low-level conditions that ensure $\sqrt{n}$-consistency of $\hat{\gamma}_U$. To this purpose, define

$$\boldsymbol{Q}(\gamma) \equiv \max\left\{ \max_{j=1,\ldots,J} E_P(m_j(X_i, \theta)/\sigma_j(\theta)) : g(\theta) = \gamma \right\}.$$

In words, $\boldsymbol{Q}(\cdot)$ is the criterion function but (i) with $g(\cdot)$ concentrated out and (ii) allowing for strictly negative values. The latter is needed to formalize conditions under which $\hat{\Theta}_I$ will explore the interior of Θ_I. Specifically, consider:

1. There exist $C, \epsilon > 0$ s.t. $|\gamma - \gamma_U| \leqslant \epsilon \implies \boldsymbol{Q}(\gamma) \geqslant C \cdot (\gamma - \gamma_U)$.
2. Γ_I is either one interval or is the finite union of intervals of positive length.

We then have:

Theorem 3.4: *Above Condition ensures that* $\sqrt{n}(s(p, \hat{\Theta}_I) - s(p, \Theta_I)) = O_P(1)$.

Proof. Omitted. This relates to rate conditions for level set estimation as in Molchanov (1998); however, we here state a uniform version of his assumption. □

Intuitively, the assumption states that the boundary of the projection $\mathcal{P}(\Theta_I)$ is well-identified: The criterion both increases sufficiently fast away from the projection but also dips below zero sufficiently quickly on its interior so that it can be approximateted sufficiently quickly from either direction. While somewhat high-level, the assumption is stated in a way that accommodates point identification, partial identification, and also cases for "near" point identification. Thus, it does not rely on a case distinction between point and partial identification, whereas lower level assumption frequently presuppose one or the other.

The following remark relates our assumption to the literature.[9]

Remark 3.1: Suppose background assumptions hold. If Θ_I is a singleton, above condition is implied by the minorant condition for moment inequality models in Chernozhukov, Hong, and Tamer (2007, display (4.5)). Otherwise, the assumption is implied by the same paper's degeneracy condition for moment inequalities (display 4.6). It is also implied by Pakes, Porter, Ho, and Ishii (2011, Assumption 4(a)).

We can similarly relate Assumption 3.5(2) to the literature.

Remark 3.2: Suppose background assumptions hold. Then Assumption 3.5(2) is implied by Bugni, Canay, and Shi (2017, Assumption 3(a)) as well as Pakes, Porter, Ho, and Ishii (2011, Assumption 3).

An important addendum to this remark is that both of these conditions from the literature are considerably stronger than what we need. In particular, in both cases, $d(\theta, S(g, \Theta_I))$ is not squared. This is meaningfully more restrictive because it excludes the possibility of "smooth maxima," i.e. the surface of Θ_I being (locally) a differentiable manifold of which θ^* is an extreme point.

[9] See the appendix for formal statements of assumptions alluded to as well as proofs of remarks. The remarks draw on insights developed in Kaido, Molinari, and Stoye (2022).

3.3.4 Justifying Further Simplification

We next provide further restrictions that allow one to set ρ equal to ∞ or, equivalently, to drop the "ρ-box" constraints altogether. In principle, this is attractive because it removes a remaining source of conservatism, as well as a tuning parameter. However, we caution that the assumptions justifying it are restrictive. An auxiliary contribution of this section's analysis is to connect Simple Calibrated Projection to the pioneering bootstrap approach of Pakes, Porter, Ho, and Ishii (2011) and to improve our understanding of the latter.

Our strongest restriction on the shape of Θ_I is as follows.

ASSUMPTION 3.6: *The solution set $S(g, \Theta_I)$ is a finite union of singletons. Furthermore, there exists $\epsilon > 0$ s.t. for each $\theta^* \in S(g, \Theta_I)$, there exist k constraints $\tilde{\mathcal{J}}(\theta^*) \equiv \{j_1, \dots, j_k\} \subset \mathcal{J}^*(\theta^*)$ s.t.*

$$\min \operatorname{eig} \begin{pmatrix} D_{j_1}(\theta^*) \\ \vdots \\ D_{j_k}(\theta^*) \end{pmatrix} > \epsilon$$

$$\max_{t \in \mathcal{T}(\theta^*) \setminus \{0\}} p' t / \|t\| \leqslant -\epsilon.$$

The assumption has two parts. First, any support point must be characterized as intersection of linearly independent constraints. This comes with two clarifications: The restriction on eigenvalues ensures linear independence in a way that is uniform over $\mathcal{P}$; also, recall that an equality constraint consists of two inequality constraints (but can enter the above index set only once as the corresponding gradients are collinear). Second, the polyhedral cone that is locally defined by these constraints is well-separated from the supporting halfspace; that is, no direction in the supporting halfspace is (near) tangential to it. Geometrically, if the optimization problem were defined by these constraints only, the tangent cone would be pointy (while having an interior) and would point uniformly away from the supporting halfspace.

The assumption is obviously restrictive. That said, it has important precedent in the literature.

REMARK 3.3: Above assumption is implied by Pakes, Porter, Ho, and Ishii (2011, Assumption 3 and 4a).

Once again, the implication actually holds with slack. The aforementioned assumption forces the support set to be a global singleton and also for Θ_I to be contained in $\mathcal{T}(\theta^*)$.

Our final major result is:

THEOREM 3.5: ***No ρ-box.*** *If all of the above assumptions hold, then one can set $\rho = +\infty$; equivalently, the "ρ-box constraints" can be discarded.*

4 Relation to Literature

This section relates Simple Calibrated Projection to other approaches in the literature including Andrews, Roth, and Pakes (2021), Bugni, Canay, and Shi (2017), Cho and Russell (2021), Cox and Shi (2020), Kaido, Molinari, and Stoye (2019), and Pakes, Porter, Ho, and Ishii (2011). A comparison of particular interest is to the latter because (i) Simple Calibrated Projection is operationally similar to that prioneering proposal[10] and (ii) even our strongest assumptions are implied by theirs. Indeed, from our last result we can also derive a precise justification of their approach under assumptions that are still weaker –notably, they allow for point identification or near point identification–, with the exception of the addition of Assumption 3.2, which is however needed for the result to obtain.

5 Empirical Applications

We are in the process of validating our Stata/Python implementation and of replicating several papers from the literature.

6 Conclusion

We provide a novel method for inference on constrained maxima, and in particular for inference of scalar fucntions of partially identified parameter vectors, that is rather generally valid and exceptionally easy to implement. We demonstrate this by re-evaluating empirical claims in several published papers.

References

Andrews, D. W. K., and S. Kwon (2019): "Inference in Moment Inequality Models That Is Robust to Spurious Precision under Model Misspecification," *Cowles Foundation Discussion Paper CFDP 2184R*.

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

Andrews, I., J. Roth, and A. Pakes (2021): "Inference for Linear Conditional Moment Inequalities," Papers 1909.10062, arXiv.org.

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

[10] As a reminder, that proposal is to bootstrap the value function of a local linearized analog of the recentered optimization problem and then report a percentile confidence interval by subtracting the α-quantile of the bootstrapped distribution from the initial estimator (this quantile is negative due to centering).

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.

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

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

Cho, J., and T. M. Russell (2021): "Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments," .

Cox, G., and X. Shi (2020): "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models," .

Ho, K. (2009): "Insurer-Provider Networks in the Medical Care Market," *American Economic Review*, 99(1), 393–430.

Ho, K., J. Ho, and J. H. Mortimer (2012): "The Use of Full-Line Forcing Contracts in the Video Rental Industry," *American Economic Review*, 102(2), 686–719.

Ho, K., and A. Pakes (2014): "Hospital Choices, Hospital Prices, and Financial Incentives to Physicians," *American Economic Review*, 104(12), 3841–84.

Holmes, T. J. (2011): "The Diffusion of Wal-Mart and Economies of Density," *Econometrica*, 79(1), 253–302.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence Intervals for Projections of Partially Identified Parameters," *Econometrica*, 87(4), 1397–1432.

——— (2022): "Constraint Qualifications in Partial Identification," *Econometric Theory*, pp. 1–24.

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.

Kawai, K., and Y. Watanabe (2013): "Inferring Strategic Voting," *American Economic Review*, 103(2), 624–62.

Kline, P., and M. Tartari (2016): "Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach," *American Economic Review*, 106(4), 972–1014.

Lee, R. S. (2013): "Vertical Integration and Exclusivity in Platform and Two-Sided Markets," *American Economic Review*, 103(7), 2960–3000.

Molchanov, I. (1998): "A limit theorem for solutions of inequalities," *Scand. J. Statist.*, 25, 235–242.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4, chap. 36. New York: Elsevier.

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.

Ponomareva, M., and E. Tamer (2011): "Misspecification in moment inequality models: back to moment equalities?," *The Econometrics Journal*, 14(2), 186–203.

Redner, R. (1981): "Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions," *The Annals of Statistics*, 9(1), 225 – 228.

Stoye, J. (2009): "More on Confidence Regions for Partially Identified Parameters," *Econometrica*, 77(4), 1299–1315.

——— (2020): "A Simple, Short, but Never-Empty Confidence Interval for Partially Identified Parameters," .

van der Vaart, A., and J. Wellner (2000): *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer-Verlag, Berlin.

Wollmann, T. G. (2018): "Trucks without Bailouts: Equilibrium Product Characteristics for Commercial Vehicles," *American Economic Review*, 108(6), 1364–1406.