

Inference in high-dimensional set-identified affine models*

Bulat Gafarov[†]

2019-04-02

Abstract

This paper proposes both *point-wise* and *uniform* confidence sets (CS) for an element θ_1 of a parameter vector $\theta \in \mathbb{R}^d$ that is partially identified by affine moment equality and inequality conditions. The method is based on an estimator of a regularized support function of the identified set. This estimator is *half-median unbiased* and has an *asymptotic linear representation* which provides closed form standard errors and enables optimization-free multiplier bootstrap. The proposed CS can be computed as a solution to a finite number of linear and convex quadratic programs, which leads to a substantial decrease in *computation time* and *guarantee of global optimum*. As a result, the method provides uniformly valid inference in applications with the dimension of the parameter space, d , and the number of inequalities, k , that were previously computationally unfeasible ($d, k > 100$). The proposed approach is then extended to construct polygon-shaped joint CS for multiple components of θ . Inference for coefficients in the linear IV regression model with interval outcome is used as an illustrative example.

Key Words: Affine moment inequalities; Asymptotic linear representation; Delta-Method; Interval data; Intersection bounds; Partial identification; Regularization; Strong approximation; Stochastic Programming; Subvector inference; Uniform inference.

*The earlier version of the paper was previously circulated under title "Inference on scalar parameters in set-identified affine models" and was a chapter in my PhD dissertation, Gafarov (2017). The first draft date: November 10th, 2015.

[†]University of California, Davis, Department of Agricultural and Resource Economics. E-mail: bgafarov@ucdavis.edu; I am extremely grateful to Joris Pinkse and Patrik Guggenberger for their very helpful and detailed comments on this paper. I would like to thank Donald Andrews, Andres Aradillas-Lopez, Christian Bontemp, Ivan Canay, Joachim Fryberger, Ronald Gallant, Michael Gechter, Marc Henry, Keisuke Hirano, Sung Jae Jun, Nail Kashaev, Francesca Molinari, Demian Pouzo, Adam Rosen, Thomas Russell, Xiaoxia Shi, Jing Tao, Alexander Torgovitsky, and Fang Zhang (in alphabetical order) for the comments and suggestions on this project.

1 Introduction

Strong econometric assumptions can lead to poor estimates. Moment inequalities occasionally provide alternative estimates under weaker assumptions. Linear models with interval-valued data are a good example.¹ It is common practice to replace the income bracket data with the corresponding midpoints when estimating the returns to schooling (Trostel et al. (2002)). The conventional approach is applicable only under strong assumptions on the distribution of the residual term.² The affine moment inequality approach to interval-valued data proposed by Manski and Tamer (2002) can set-identify the return to schooling without such strong assumptions.

There are multiple methods that can be used to construct confidence sets for parameters defined by moment inequalities. The pioneering procedures of Chernozhukov et al. (2007) and Andrews and Soares (2010, AS) and their subsequent refinements by Bugni et al. (2016) (BCS) and Kaido et al. (2015) (KMS) are powerful procedures that solve this inference problem in the small-dimensional case. Some applications, such as panel or semiparametric regression models with interval measured outcome variables, have a large dimension of the parameter space (for example, Trostel et al. (2002) consider a panel regression with more than 60 variables including country fixed effects, time effects, exogenous demographic control variables, and their interactions) which poses a computational challenge for the existing procedures.

I propose confidence intervals (CIs) for an element θ_1 of an unknown parameter vector $\theta \in \mathbb{R}^d$ in models defined by affine moment equalities and inequalities. In the returns to schooling example, θ_1 corresponds to the returns to schooling and $\theta \in \mathbb{R}^d$ to the full vector of the regression coefficients that can include many control variables. I estimate the lower and upper extremes of the identified set for θ_1 , which is an interval, using an estimator of the regularized support function. This estimator has a closed-form asymptotic Gaussian distribution which I use to construct both point-wise valid and uniform CIs for θ_1 . The latter asymptotically controls the coverage probability uniformly over a class of data generating processes (DGP) (as it was pointed out in Imbens and Manski (2004), the uniformity in DGP is desirable as it controls coverage probability in finite sample properties better than point-wise CIs).

Procedure. The regularized support function proposed in this paper is a solution to a convex quadratic program that minimizes the sum of θ_1 and a penalty $\mu_n \|\theta\|^2$ with $\mu_n \rightarrow 0$, subject to the sample moment restrictions. If the set of optima for $\mu = 0$ is not a singleton, this additional convex term selects the optimum with the minimal norm as n increases. The standard errors are computed using the sample variance of the weighted moment conditions at the unique optima. To correct the asymptotic bias resulting from the regularization exactly, I suggest using the argmin of the regularized program with a larger tuning parameter $\kappa_n \rightarrow 0$. If $\kappa_n/\mu_n \rightarrow \infty$ as $n \rightarrow \infty$, then the bias correction does not affect the asymptotic distribution of the estimator. To achieve a uniformly valid CI, I replace the exact correction with an upper bound on the maximum of $\mu_n \|\theta\|^2$ over the argmin set of the non-regularized program.

The proposed CIs have several attractive statistical and computational properties which make them viable in high dimensional affine moment inequality models.

¹Other examples of affine moment inequalities include monotone instrumental variables (Manski and Pepper (2000), Freyberger and Horowitz (2015)) and models with missing data (Manski (2003)).

²Another common approach is to assume Gaussian distribution for the residuals and apply Maximum Likelihood method (Stewart (1983)).

Asymptotically linear representation. The estimator of the regularized support function has a Asymptotically linear (or Bahadur) representation that provides easy to compute asymptotic standard error and enable the multiplier bootstrap. This resampling procedure avoids the necessity of solving mathematical programs for every bootstrap draw present in the existing uniform methods.

This paper is the first to propose a closed-form estimator of the bounds on θ_1 in affine moment inequality models with asymptotic Gaussian distribution. In contrast, the estimator of the ordinary support function used in the existing literature (Beresteanu and Molinari (2008), Kaido and Santos (2014), Freyberger and Horowitz (2015, FH), Gafarov et al. (2018), among others) has non-Gaussian asymptotic distribution, which complicates inference.

Computational properties. The proposed approach requires only a fraction of the computational time of the existing pointwise and uniform procedures, in particular if θ has a large dimension. The computational cost is low since it involves only four quadratic programs, it does not require any resampling and it depends on covariance of the moment conditions at two points.

The computation time for my procedure increases only slowly in the dimension of $\theta \in \mathbb{R}^d$ and takes 1 sec only for $d = 15$ and $k = 30$ moment inequalities and 20 sec for $d = 100$ and $k = 100$. As a result, the proposed method can address the parameters with a large dimension and a large number of moment conditions. In contrast, the existing uniform inference methods for moment inequalities proposed by AS, KMS, and BCS are based on costly non-convex optimization.

The new estimators, which are based on strictly convex programs, also avoid the problem of distinguishing between local and global optima, which is present for the existing uniform procedures even in the affine moment inequality setup. The low computational cost together with applicability of the fast multiplier bootstrap allows one to perform joint inference on the components of θ using the support function representation of a convex set.

I provide an example of an affine moment inequality model illustrating that the number of local optimal solutions in the existing uniform procedures (AS,BCS, and KMS) can grow exponentially with the dimension d . As a result, the procedures take more computational time and can produce misleadingly short CIs if the optimization routine fails to find the global optimum. It takes 630 sec to compute the CI of AS in an affine model with $d = 15$ and 30 moment inequalities.³ In my numerical experiments the computational time for the AS procedure increases by 30% with every additional dimension d while my procedure is barely affected by changes in the dimension.

Length comparison. The proposed uniform CIs have length properties that are not worse than these of the existing methods as suggested by Monte Carlo experiments. The proposed *uniform* CI is has length within simulation error from the projection CI of AS in the MC design considered in the paper.

Assumptions. The uniform CI is applicable in a situation where the existing uniform procedures are inapplicable. I show that a linear model with an interval-valued outcome can have a moment inequality with zero variance which violates the assumptions in AS, Kaido et al. (2015, KMS) and Bugni et al. (2014, BCS).⁴

³I use the implementation of the AS procedure provided by KMS. This algorithm uses smooth interpolation of the critical values by the kriging method which allows one to use a Newton-type solver. This approach reduces the computational cost of the AS procedure.

⁴AS, KMS, and BCS procedures can be applied in some setups where my procedure is not applicable. I compare setups in Section 2.5.

The class of DGPs over which I prove uniform coverage properties is not nested in the classes considered in BCS and KMS. I impose a rank condition on the affine constraints typical for the support function approach (Beresteanu and Molinari (2008), Kaido and Santos (2014), FH, Gafarov et al. (2018)). These conditions rule out over-identification of the solutions to the regularized programs. In particular, they rule out the possibility of point-identification by moment inequalities of the components of θ , which can be addressed using BCS and KMS procedures. This complication can be alleviated if one can split the full set of the inequality constraints into (possibly overlapping) subsets that meet the rank condition. Within this framework, my procedure covers θ_1 for any sequences of DGP that drift to a DGP with a moment condition orthogonal to θ_1 , which is not the case in BCS.⁵ As mentioned earlier, my CIs remain valid if some moment inequalities have zero variance, which violates assumptions in both KMS and BCS. I expect poor coverage of the existing procedures as the variance becomes very small (but still positive).

The number of maintained assumptions for the large sample inference in the present paper is kept to a minimum. All the assumptions are testable.

Examples of affine inequality models. Identified sets defined by affine inequalities appear in various economic applications, in particular, those dealing with discrete variables and shape restrictions. Linear models with interval outcome, that were originally studied in Manski and Tamer (2002) and Haile and Tamer (2003), is just one example of affine inequalities. Other examples include bounds on marginal effects in panel dynamic discrete choice models (Honoré and Tamer (2006), Torgovitsky (2016, 2018)), bounds on average treatment effects (Kasy (2016), Laffers (2018), Russell (2017)), non-parametric instrumental variable models with shape restrictions (Manski and Pepper (2000), Freyberger and Horowitz (2015)), errors in variables (Molinari (2008)), intersection bounds (Honoré and Lleras-Muney (2006)), revealed preference restrictions (Kline and Tartari (2016), Shi et al. (2018)), game-theoretic models (Pakes et al. (2015), Syrgkanis et al. (2017)).

Inference on non-differentiable functions and regularized estimators. Bounds on components of a parameter characterized by a linear moment conditions considered in this paper is an example of a non-differentiable (*non-regular*) function of a parameter (the expectation of the data) that has an asymptotically normal estimator (the sample mean). Recently a number of papers have studied inference for such non-regular functions. The problem was considered first by Shapiro (1991) and Dümbgen (1993). Hirano and Porter (2012) proved that it is impossible to have a locally unbiased estimator of non-regular parameters. In particular, it implies that it is impossible perform non-conservative inference on non-regular parameters. Fang and Santos (2018) shown that the standard bootstrap inference in this setup is inconsistent in general provide a procedure based on a consistent estimator for the directional derivative of the non-regular function. Hong and Li (2018) provide a general way to estimate the directional derivative by developing the numerical delta-method based on rescaled bootstrap of Dümbgen (1993). In general, confidence sets based on this method are only point-wise consistent.

In this paper I propose uniformly valid CS for θ_1 using a different approach. I propose regular lower and upper bounds that converge to the non-regular parameter of interest as sample size grows. Since the bounds are regular parameters themselves, the standard delta-method and bootstrap can be used to conduct one or two-sided inference on the bounds. In the regular case these bounds collapse and coincide with the original parameter of interest, which results in \sqrt{n} -consistent and

⁵See KMS for a discussion of the assumptions in BCS

asymptotically normal estimator. In the non-regular case, the bounds converge at a slower rate which results in a locally biased estimator.

Inference for extrema of finitely many parameters known as intersection bounds problem (Hall and Miller (2010) and Chernozhukov et al. (2013)) can be framed as a value of a linear program. The regularized support function estimator can also be used for uniform delta-method CS in this setting. The approach considered here is expected to have similar statistical properties to Chernozhukov et al. (2013) but has additional advantage of closed form standard errors and critical values, which correspond to the standard normal distribution.

Finally, the regularization principle was recently considered in Jansson and Pouzo (2017) who studied general conditions for consistency and asymptotic linear representation for regularized estimators. That work was concerned with estimation and point-wise inference, whereas I also study uniform asymptotic linear representation and inference in a special case of affine moment inequality models.

Structure. The paper is structured as follows. Section 2 describes the setup and the main result. Section 3 outlines the extension to the general subvector inference and show how one can deal with the violation of the main regularity assumptions on the moment conditions. Section 4 compares the maintained assumptions and the computational properties of the proposed procedure with the existing alternatives. Section 5 provides the results of the Monte Carlo experiments. Section 6 concludes.

Notation. I use \triangleq to denote definitions. I write $\mathbb{E}_P[\cdot]$ to denote expectation with respect to a probability distribution P . I use uppercase English letters to denote random variables (scalar, vector, or matrix valued) and lower case letters to denote the corresponding realizations, W and w_i . I use \mathbb{P}_n for the sample distribution. I use $f(0^+)$ for $\lim_{x \downarrow 0} f(x)$. The vector $e_j \triangleq (0, \dots, 1, \dots, 0)'$ is the j -th coordinate vector, where the one occurs at position j . e_j is the projector on the j -th coordinate. I use the symbol \mathcal{J} for a finite set of indices $\mathcal{J} \triangleq \{i_1, \dots, i_\ell\} \subset \mathbb{N}$ and $\mathbb{J} \triangleq (e_{i_1}, \dots, e_{i_\ell})'$ as a coordinate projection matrix in the corresponding Euclidean space. I use $|\mathcal{J}|$ to denote the cardinality of the set \mathcal{J} . The acronym u.h.c. stands for upper hemi-continuous correspondence. I will use symbol $\text{sVar}(x)$ to denote the sample variance, $\text{sVar}(x) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2$.

2 Setup and main results

2.1 Affine moment conditions

I consider a parameter vector $\theta \in \Theta \subset \mathbb{R}^d$ where Θ is the set defined by

$$-\infty < a_\ell \leq \theta_\ell \leq b_\ell < \infty \quad (1)$$

for $\ell = 1, \dots, d$. The inequalities (1) can be written as a subsystem of the following system of unconditional moment equalities/inequalities

$$\begin{cases} \mathbb{E}_P g_j(W, \theta) = 0, & j \in \mathcal{J}^{eq}, \\ \mathbb{E}_P g_j(W, \theta) \leq 0, & j \in \mathcal{J}^{ineq}, \end{cases} \quad (2)$$

where $g_j(W, \theta) \triangleq \sum_{\ell=1}^d W_{j\ell} \theta_\ell - W_{j(d+1)}$, $|\mathcal{J}^{eq}| = p$, $0 \leq p \leq d$, $|\mathcal{J}^{ineq}| = k - p \geq 2d$, $k < \infty$, the random matrix W has probability measure P with the sample space $\mathbb{R}^{k \times (d+1)}$. Correspondingly, $|\mathcal{J}^{eq} \cup \mathcal{J}^{ineq}| = k$. A solution to (2) may not be unique. Let the identified set $\Theta(P) \subset \Theta \subset \mathbb{R}^d$ be the set of parameter values θ that satisfy (2) for a given data generating process (DGP) parametrized by P . The stochastic programming approach described below allows me to deal with both random and deterministic (in)equalities in (2) symmetrically.

The identified set $\Theta(P)$ is a polytope or an empty set. The convexity of $\Theta(P)$ provides characterization using support functions. Such support functions, for example, can provide bounds on coordinate projections of $\Theta(P)$ or any subvectors of $\theta \in \Theta(P)$. Without loss of generality, I will consider first a special case of $\theta_1 = e_1' \theta$, the value of the first component of $\theta \in \Theta(P)$. Any other support function can be obtained by a corresponding rotation of W and θ .

Definition 1. *The marginal identified set for θ_1 is the set*

$$\mathcal{S}(P) = \{e_1' \theta | \theta \in \Theta(P)\}. \quad \square$$

The first assumption I make rules out a possibility of specification and makes boundaries of $\mathcal{S}(P)$ well defined (specification testing is discussed later in Remark 1 in Section 3.2).

Assumption 1. *$\Theta(P)$ is nonempty for the probability measure P .*

The following example illustrates the general setup.

Example 1 (*Linear IV model with interval valued outcome*). Consider a linear IV model

$$\mathbb{E}_P [Y - \theta' X | Z] = 0,$$

where Y is unobserved. One can only observe bounds \underline{Y} and \overline{Y} such that $Y \in [\underline{Y}, \overline{Y}]$ a.s. Suppose that Z , the random vector of instruments, has a finite support $S_Z = \{z_1, \dots, z_K\} \subset \mathbb{R}^d$.⁶ In this case the model can be equivalently characterized by a finite number of conditional moments:

$$\mathbb{E}_P [\overline{Y} | Z = z_j] \geq \theta' \mathbb{E}_P [X | Z = z_j] \geq \mathbb{E}_P [\underline{Y} | Z = z_j], \quad j = 1, \dots, K.$$

The identified set $\Theta(P)$ is defined by the set of unconditional moment inequalities,

$$\begin{cases} \mathbb{E}_P [\underline{Y} 1\{Z = z_j\}] \leq \theta' \mathbb{E}_P [X 1\{Z = z_j\}], & j = 1, \dots, K, \\ \mathbb{E}_P [\overline{Y} 1\{Z = z_{j-K}\}] \geq \theta' \mathbb{E}_P [X 1\{Z = z_{j-K}\}], & j = K + 1, \dots, 2K. \end{cases} \quad (3)$$

These inequalities can be converted to the form (2) with $p = 0$, $k = 2K$ and

$$\begin{aligned} W_{j\ell} &\triangleq \begin{cases} -X_{\ell} 1\{Z = z_j\}, & \text{for } j = 1, \dots, K, \\ X_{\ell} 1\{Z = z_{j-K}\}, & \text{for } j = K + 1, \dots, 2K, \end{cases} \\ W_{j(d+1)} &\triangleq \begin{cases} \underline{Y} 1\{Z = z_j\}, & \text{for } j = 1, \dots, K, \\ -\overline{Y} 1\{Z = z_{j-K}\}, & \text{for } j = K + 1, \dots, 2K. \end{cases} \end{aligned}$$

⁶If S_Z is infinite, one can estimate an enlargement of $\mathcal{S}(P)$ using a finite number of unconditional moment inequalities. See Chernozhukov et al. (2007) for details. Andrews and Shi (2013) provide conditions for sharp characterization of the identified set by a finite number of unconditional moment functions. I leave the case of infinite number of moment inequalities for future extensions.

If it is known *a priori* that for some support points j

$$\mathbb{E}_P [\underline{Y}1\{Z = z_j\}] = \mathbb{E}_P [\overline{Y}1\{Z = z_{j-K}\}],$$

then one should replace the corresponding pair of inequalities with a single equality,

$$\mathbb{E}_P \left[\frac{\underline{Y} + \overline{Y}}{2} 1\{Z = z_j\} \right] = \theta' \mathbb{E}_P X 1\{Z = z_j\}. \quad (4)$$

In this case p is equal to the number of such support points.

One can incorporate additional information such as sign restrictions on θ in the form of linear inequalities to get a smaller identified set. \square

The following example illustrate one particular dataset and economic application of the interval-outcome IV regression.

Example 2 (The returns to schooling). Trostel et al. (2002) study economic returns to schooling for 28 countries using International Social Survey Programme data (ISSP), 1985–1995. They estimate a conventional Mincer (1974) model of earnings (the human capital earnings function), which has log wage determined by years of schooling, age, experience, and other explanatory variables:

$$\mathbb{E}[Y|Z] = \theta'Z, \quad (5)$$

where Y is the log of hourly wages, Z_1 is years of schooling and the other components of Z is a vector of observed exogenous explanatory variables including, where appropriate, country and year fixed effects. The component θ_1 is interpreted as the rate of returns to schooling; it is equal to the percentage change in wages due to an additional year of schooling. Their explanatory variables Z include year dummies, union status, marital status, age and age squared and, in the case of the aggregate equation, country-year dummies. Exact measures of Y are not available for some countries (including the USA); only income bracket data $[\underline{Y}, \overline{Y}]$ is available for those countries. Trostel et al. (2002) use a conventional technique to deal with this problem – they replace the interval data with the corresponding midpoints and estimate (5) using OLS. This technique is valid only under the unreasonably strong condition

$$\mathbb{E}_P [(Y - 0.5(\underline{Y} + \overline{Y})) Z] = 0. \quad (6)$$

If condition (6) is violated then the OLS estimator for the effect of schooling is inconsistent.

The interval outcome model from Example 1 can provide estimates of the marginal identified set for returns to schooling without assumption (6). The conventional estimates based on the midpoint approach converge to one of the elements in $\mathcal{S}(P)$. All the explanatory variables are discrete in this example, so the existing approach of Bontemps et al. (2012) is not applicable.⁷ The number d of the explanatory variables Z including country and time effects is larger than 60 which makes the existing (uniformly valid) approaches to moment inequalities computationally challenging. \square

Since in our setup all the moment conditions are affine, the identified set $\Theta(P)$ is a polytope

⁷ Bontemps et al. (2012) provide a sharp characterization of the identified set and the corresponding confidence intervals in a class of linear models with interval-valued outcome if all of the regressors have continuous support. This example has discrete-valued regressors.

and the marginal identified set is an interval, $\mathcal{S}(P) = [\underline{v}(P), \bar{v}(P)]$, where

$$\underline{v}(P) = \min_{\theta \in \Theta(P)} e_1' \theta \quad \text{and} \quad \bar{v}(P) = \max_{\theta \in \Theta(P)} e_1' \theta. \quad (7)$$

It is possible that $\mathcal{S}(P)$ is a singleton. The value functions $\underline{v}(P)$ and $-\bar{v}(P)$ are the support functions for e_1 and $-e_1$, respectively. The analysis for the upper bound is analogous to that for the lower bound, so from here on I focus on the lower bound. I will use the following example to illustrate ideas throughout the paper.

Example 3 (Running example). Consider the linear model with discrete IV from Example 1. Suppose that $\theta \in \mathbb{R}^2$, $X = Z$, and suppose that z_1, z_2 take values in $\{0, 1\}$ with equal probability ($K = 4$). As in equation (3), the identified set for can be characterized using $2K = 8$ inequality constraints,

$$\mathbb{E}[\underline{Y}\psi_z(Z)] \leq \mathbb{E}[Z_1\psi_z(Z)]\theta_1 + \mathbb{E}[Z_2\psi_z(Z)]\theta_2 \leq \mathbb{E}[\bar{Y}\psi_z(Z)], \quad (8)$$

where indicator functions $\psi_z(Z_1, Z_2) = 1\{Z = z\}$ correspond to all combinations of $z \in \{0, 1\}^2$. For illustrative purposes, consider the following subsystem of four inequalities

$$\mathbb{E}[\underline{Y}Z_1] \leq \frac{1}{2}\theta_1 + \theta_2\mathbb{E}[Z_1Z_2] \leq \mathbb{E}[\bar{Y}_iZ_1], \mathbb{E}[\underline{Y}(1 - Z_1)] \leq \theta_2\mathbb{E}[(1 - Z_1)Z_2] \leq \mathbb{E}[\bar{Y}(1 - Z_1)]. \quad (9)$$

Suppose that the identified set for θ is given by (9). These bounds are not sharp in general, but they are easy to study analytically. Suppose further that the a.s. bounds on the outcome variable Y satisfy $\mathbb{E}_P[\bar{Y}|Z_1 = i] = -\mathbb{E}_P[\underline{Y}|Z_1 = i] = \frac{1}{2}\Delta_i \geq 0$ for $i \in \{0, 1\}$ which implies Δ_i is the average length of the outcome interval depending on Z_1 . Let $\rho = \mathbb{E}(z_1z_2)$. Then the marginal identified set can be written in explicit form $\mathcal{S}(P) = [-\Delta_1 - 2|\rho|\Delta_0, \Delta_1 + 2|\rho|\Delta_0]$.

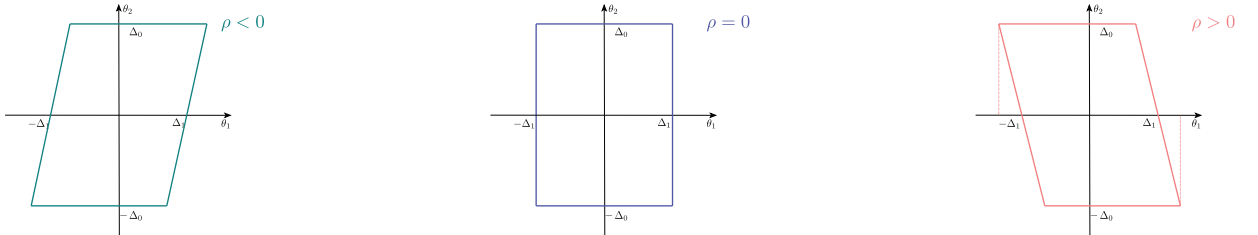


Figure 1: The identified sets corresponding to system (9) in Example 3 for various values of ρ .

Figure 1 shows that the shape of the full identified set $\Theta(P)$ depends on the sign of ρ . The borderline case $\rho = 0$ will play crucial role throughout the paper as since the coordinate of the argmin of Program (7), $\underline{\theta}_2$, is not uniquely defined. I will refer to such cases as non-regular. \square

2.2 Bounds based on the regularized support function.

The Delta-method framework is a natural way to do inference on $\underline{v}(P)$. In order to use this approach we need to study the behavior of Program (7) when we replace the coefficients $\mathbb{E}_P W$ with their consistent estimators. Namely, we need to consider a (directional) derivative of $\underline{v}(P)$ with respect to $(A_P, b_P) \triangleq \mathbb{E}_P W$ (correspondingly, $\mathbb{E}_P g(W, \theta) = A_P \theta - b_P$). The derivative can be obtained by the envelope theorem applied to the min/max representation for Program (7) which

is valid under Assumption 1,

$$\underline{v}(P) = \min_{\theta \in \mathbb{R}^d} \max_{\lambda \in \mathbb{R}^p \times \mathbb{R}_+^{k-p}} \{\theta_1 + \lambda'(A_P \theta - b_P)\}. \quad (10)$$

This representation shows that the derivative of $\underline{v}(P)$ with respect to (A_P, b_P) depends not only on the optimal solution of Program (7) but also on the solution $\underline{\lambda}(P)$ to the corresponding dual program which is defined below.

The dual program takes form

$$\begin{aligned} \underline{v}(P) = \max_{\lambda \in \mathbb{R}^p \times \mathbb{R}_+^{k-p}} \{-\lambda' b_P\} \\ \text{s.t. } \lambda' A_P = e'_1. \end{aligned} \quad (11)$$

If solutions to both (7) and (11) are unique, the envelope theorem suggests

$$\frac{\partial \underline{v}(P)}{\partial (A_P)_{ij}} = \underline{\lambda}_i \underline{\theta}_j \text{ and } \frac{\partial \underline{v}(P)}{\partial (b_P)_i} = -\underline{\lambda}_i. \quad (12)$$

Shapiro (1993) shows that under Assumption 1 and some additional regularity conditions that if we use $\frac{1}{n} \sum_{i=1}^n w_i$ as a consistent estimator for $\mathbb{E}_P W$, the value of

$$\hat{\underline{v}}_n = \min_{\theta \in \mathbb{R}^d} e'_1 \theta \quad (13)$$

$$\text{s.t. } \begin{cases} \frac{1}{n} \sum_{i=1}^n g_j(w_i, \theta) = 0, & j \in \mathcal{J}^{eq}, \\ \frac{1}{n} \sum_{i=1}^n g_j(w_i, \theta) \leq 0, & j \in \mathcal{J}^{ineq}, \end{cases} \quad (14)$$

is a consistent estimator of $\underline{v}(P)$ with a non-Gaussian asymptotic distribution that depends on both $\underline{\theta}$ and $\underline{\lambda}$. These parameters are not uniquely defined in general which makes them nuisance parameters. I suggest imposing a constraint qualification on $\mathbb{E}_P g(W, \theta)$ to ensure a unique $\underline{\lambda}$ and regularizing Program (7) to select a unique point in $\underline{\theta}$.⁸

To introduce the constraint qualification, I use the following notation. For any $\mathcal{J} \subset \mathcal{J}^{ineq}$ let the projection matrix $\mathbb{J}^a = (e_{i_1}, \dots, e_{i_\ell})'$ correspond to $\mathcal{J}^a \triangleq \mathcal{J}^{eq} \cup \mathcal{J} = \{i_1, \dots, i_\ell\}$, a set of active constraints. Let $\eta_1(\cdot)$ be the smallest left singular value function of a matrix, i.e. $\eta_1(A) \triangleq \sqrt{\min_u (u' A A' u / u' u)}$. The constraint qualification can now be formulated as the following two assumptions.

Assumption 2. *Measure P satisfies*

$$\eta(P) \triangleq \min_{\mathcal{J} \subset \mathcal{J}^{ineq}; |\mathcal{J}|=d-p} \eta_1(\mathbb{J}^a(A_P, b_P)) > 0. \quad (15)$$

In geometric terms, Assumption 2 restricts angles between the gradients of any d intersecting moment restrictions to be away from zero. Moreover, the gradients need to have the norm bounded away from zero.⁹

⁸Constraint qualifications are common in the literature on set identified models. In particular, BM, KS, FH and Gafarov et al. (2018) impose them in various forms.

⁹Potentially, one could normalize the inequality constraints by $\|e'_j \mathbb{E}_P W\|$. This way, it may be easier to meet Assumption 2 for inequalities j with smaller value of the norm. Moreover, the normalization will re-scale the corresponding Lagrange multiplier $\underline{\lambda}_j$ by the same factor which can reduce the variance of the resulting estimator

Assumption 3. *Measure P satisfies*

$$s(P) \triangleq \min_{\mathcal{J} \subset \mathcal{J}^{ineq}; |\mathcal{J}|=d-p+1; \theta \in \Theta(P)} \|\mathbb{J}^a(A_P \theta - b_P)\| > 0. \quad (16)$$

Example 4 (Example 3 continued). To meet Assumption 2 the moments (9) have to satisfy

$$\Delta_0 > 0, \Delta_1 > 0, \quad (17)$$

$$\mathbb{E}[(1 - Z_1) Z_2] > 0. \quad (18)$$

Inequality (17) implies that the upper and lower bounds on Y are different conditional on Z_1 , while inequality (18) implies that the instruments Z_2 with values in $\{0, 1\}$ is not perfectly correlated with Z_1 . Both conditions seem reasonable. Indeed, in case if the bounds \underline{Y} and \bar{Y} coincide w.p.1, one can replace the corresponding pair of moment inequalities with an equality, as noticed in Example 1. The case $\mathbb{E}[(1 - Z_1) Z_2] = 0$ corresponds to the multicollinearity problem in the conventional linear regression setup. Inequalities (17) also imply that Assumption 3 is satisfied.

Suppose that we add one more moment condition corresponding to instrument Z_2 ,

$$\mathbb{E}[\underline{Y} Z_2] \leq \theta_1 \mathbb{E}[Z_1 Z_2] + \frac{1}{2} \theta_2.$$

Assumption 3 would be violated if this additional inequality is binding at the corner points $\theta = (-\Delta_1 \mp 2\rho\Delta_0, \pm\Delta_0)$. If in addition $\rho = 1/2$, then Assumption 6 would also be violated (because of the collinearity of corresponding gradients). \square

Together Assumptions 2-3 imply that at any point in $\theta \in \Theta(P)$ any *binding (active)* moment conditions have linearly independent gradients, a condition called *Linear Independence Constraint Qualification (LICQ)* in the optimization theory.¹⁰ LICQ is a necessary and sufficient condition for uniqueness of the Lagrange multipliers $\underline{\lambda}$.¹¹ In particular, Assumptions 2-3 are sufficient to bound the Lagrange multipliers in Program 7.¹² The bounded Lagrange multipliers are necessary and sufficient condition for stability of solutions to the linear program.¹³ Correspondingly, the explicit bound on Lagrange multipliers, that depends on $\eta(P)$ and size of the box Θ , provides a bound on the variance of the corresponding estimator. Both these assumptions also play role in guaranteeing that the sample analog of the identified set is non-empty with probability approaching 1.¹⁴

The following *regularized* program has a unique solution and approximates Program (7) from above,

$$\underline{v}(\mu, P) = \min_{\theta \in \Theta(P)} \{e_1' \theta + \mu \|\theta\|^2\}. \quad (19)$$

Clearly, for any positive μ the value of the regularized program is larger than $\underline{v}(P)$ by at least

of $\underline{\theta}_1$. At the same time, the program with normalized constraints may result in a smaller value of the cut-off tuning parameter $\bar{\mu}(P)$ described in Theorem 1 and thus larger worst-case bias. Heuristically, the benefits of the normalization will be larger for inequalities j with large value of the ratio $\text{Var} \|e_j' \mathbb{E}_P W\| / \|e_j' \mathbb{E}_P W\|$. I leave the problem of optimal normalization for future work.

¹⁰See Assumption 6 and Lemma 2 in Appendix.

¹¹See Wachsmuth (2013).

¹²The assumptions are sufficient, but not necessary for bounded Lagrange multipliers. Proposition 5.45 on p.439 in Bonnans and Shapiro (2000) provide necessary and sufficient conditions.

¹³This is true since we restrict θ to a compact set Θ . See Robinson (1977) for details.

¹⁴See Lemma 10

$\mu \|\underline{\theta}(\mu, P)\|^2$. So it is reasonable to consider the following two tighter bounds on $\underline{v}(P)$,

$$\underline{v}^{in}(\mu, \kappa, P) \triangleq \underline{v}(\mu, P) - \mu \|\underline{\theta}(\kappa, P)\|^2, \quad (20)$$

$$\underline{v}^{out}(\mu, P) \triangleq \underline{v}(\mu, P) - \mu \|\theta^*\|^2, \quad (21)$$

where θ^* is any point in $\underline{\theta}(P)$. These bounds continuously shrink towards $\underline{v}(P)$ from both sides as μ and κ go to zero as the following theorem shows.

Theorem 1. *For any P satisfying Assumption 1 and any $\kappa \geq \mu \geq 0$, the following bounds hold*

$$\underline{v}^{out}(\mu, P) \leq \underline{v}(P) \leq \underline{v}^{in}(\mu, \kappa, P) \quad (22)$$

If in addition P satisfies Assumptions 2-3, then there exist $\bar{\mu}(P) > 0$ such that $\underline{v}^{in}(\mu, \kappa, P) = \underline{v}(P)$ for any $\mu < \kappa < \bar{\mu}(P)$. Further, if $\underline{\theta}(P)$ is a singleton, then $\underline{v}^{out}(\mu, P) = \underline{v}(P)$ for any $\mu < \bar{\mu}(P)$.

Proof. See Appendix 8.3. □

Theorem 1 is crucial for inference on $\underline{v}(P)$. Under LICQ, the value of Program (19) for $\mu > 0$ is differentiable in (A_P, b_P) so that the sample analog, $\underline{v}(\mu_n, \mathbb{P}_n)$, has asymptotic Gaussian distribution.¹⁵ As a result, the analog estimators of \underline{v}^{out} and \underline{v}^{in} are half median unbiased estimators for \underline{v} and can be used for uniform one-sided inference (for both sides) in the subsequent sections.

Assumptions 2-3 are rather strong but they considerably simplify the inference. There is a purely computational reason to ensure LICQ. If it is violated, Newton-type algorithms, which typically guarantee quadratic rate of convergence to a stationary point, have linear rate of convergence or do not converge at all.¹⁶ The thresholds η and s can be consistently estimated so it is possible to test their positiveness. Assumption 3, in particular, can be violated in applications with many moment inequality conditions if the identified set is very tight as a result.¹⁷ In Section 3 I show how one can use a representation of $\underline{v}(P)$ as a maximum of sub-problems that meet this requirement even if Assumption 3 is violated.

2.3 Large sample properties of the regularized support function.

In this section I will use versions of the limiting theorems that are uniform in the underlying DGP parametrized by $P \in \mathcal{P}$. This level of generality is necessary for to study uniform inference in Section 2.5. To work with uniform limiting theorems, it is convenient to extend $O_p(1)$ and $o_p(1)$ notation.¹⁸ Let $\zeta_n(P)$ be a sequence of random vectors with measures that depend on a parameter $P \in \mathcal{P}$. Probability measures of $\zeta_n(P)$ are called uniformly tight if for any $\epsilon > 0$ there exist $R > 0$ such that for all n $\sup_{P \in \mathcal{P}} P(\|\zeta_n\| \geq R) \leq \epsilon$. I will denote random vectors $\zeta_n(P)$ with uniformly tight measures (i.e. bounded in probability) as $O_{\mathcal{P}}(1)$. Analogously, I will denote $\zeta_n(P)$ as $o_{\mathcal{P}}(1)$ if $\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P(\|\zeta_n(P)\| \geq \epsilon) = 0$.

Consider the analog estimator of $\underline{v}(\mu_n, P)$,

$$\underline{v}(\mu_n, \mathbb{P}_n) = \min_{\theta \in \Theta(\mathbb{P}_n)} \{e_1' \theta + \mu_n \|\theta\|^2\}. \quad (23)$$

¹⁵See Lemma 7 in Appendix.

¹⁶See, for example, Golishnikov and Izmailov (2006).

¹⁷One interesting recent example is Shi et al. (2018).

¹⁸Stochastic $O_p(1)$ and $o_p(1)$ notation is introduced, for example, in Van der Vaart (2000) on p.12.

In order to prove consistency and study the asymptotic distribution of this estimator, I make the following two assumptions.

Assumption 4. $\{w_i \in \mathbb{R}^{k \times (d+1)} \mid i = 1, \dots, n\}$ is an i.i.d. sample with probability measure P .

Assumption 5. There exist an $\varepsilon > 0$ such that

$$\mathbb{E}_P \|W\|^{2+\varepsilon} < \infty. \quad (24)$$

For uniform inference I will consider the class of all measures $\mathcal{P} = \mathcal{P}(\underline{\eta}, \underline{s}, \varepsilon, \bar{M})$ that satisfy Assumptions 1-5 with some uniform positive constants $\underline{\eta}, \underline{s}, \varepsilon, \bar{M}$, i.e. every $P \in \mathcal{P}$ satisfies $\Theta(P) \neq \emptyset$, $\eta(P) > \underline{\eta}$, $s(P) > \underline{s}$, and $\mathbb{E}_P \|W\|^{2+\varepsilon} < \bar{M}^{2+\varepsilon}$. Within this class, Assumptions 4-5 are sufficient to guarantee a uniform law of large numbers (LLN) for first and second moments of W and a central limit theorem (CLT) for the estimated coefficients in Program (23). In particular, Assumption 5 provides an explicit bound on the CLT approximation error of the coefficients.¹⁹

Before we can study the properties of the estimator in Program (23), we need to show that it is well defined. Lemma 10 in Appendix shows that Program (23) has feasible points and satisfies LICQ with probability approaching 1 as $n \rightarrow \infty$ uniformly in $P \in \mathcal{P}$. This property guarantees that there exists a sample argmin and a unique vector of Lagrange multipliers in large enough samples with probability approaching 1.

The envelope theorem suggests the asymptotic linear (Bahadur) representation of the value function in Program (23),²⁰

$$\underline{v}(\mu_n, \mathbb{P}_n) = \underline{v}(\mu_n, P) + \frac{1}{n} \sum_{i=1}^n \underline{\lambda}(\mu_n, P)' g(w_i, \underline{\theta}(\mu_n, P)) + O_{\mathcal{P}}\left(\frac{1}{\mu_n n}\right). \quad (25)$$

This representation, in turn, suggests a coupling of $\underline{v}(\mu_n, \mathbb{P}_n)$ with a Gaussian process. Since the residual term $O_{\mathcal{P}}(1)$ is uniformly tight over $P \in \mathcal{P}$, the asymptotic linear representation (25) makes this coupling a *strong approximation* result.²¹ The asymptotically vanishing distance between the corresponding measures is denoted as

$$\rho_n(P) \triangleq \pi(\sqrt{n}(\underline{v}(\mu_n, \mathbb{P}_n) - \underline{v}(\mu_n, P)), N(0, \underline{\sigma}^2(\mu_n, P))),$$

where $\pi(\cdot)$ is the Levy-Prohorov metric. Representation (25) suggests an analog estimator of the asymptotic variance,

$$\underline{\sigma}^2(\mu_n, \mathbb{P}_n) = \frac{1}{n} \sum_{i=1}^n (\underline{\lambda}'(\mu_n, \mathbb{P}_n) g(w_i, \underline{\theta}(\mu_n, \mathbb{P}_n)))^2,$$

where $\underline{\theta}(\mu_n, \mathbb{P}_n)$ and $\underline{\lambda}(\mu_n, \mathbb{P}_n)$ are, respectively, the optimum and the vector of Lagrange multipliers of (23). In addition, it suggests an estimator of the influence functions that can be used in the multiplier bootstrap inference.

The following theorem summarizes the large sample results for regularized program (23).

¹⁹This bound is based on the generalization of the theorem of Yurinskii (1978) by van der Vaart and Wellner (1996).

²⁰See Lemma 11 in Appendix

²¹ See Definition 4 in Appendix A of Chernozhukov et al. (2013).

Theorem 2. Consider any sequence μ_n such that $\mu_n \rightarrow 0$ and $\mu_n\sqrt{n} \rightarrow \infty$. Then with probability approaching 1 uniformly in $P \in \mathcal{P}$

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \rho_n(P) = 0, \quad (26)$$

$$\underline{\theta}(\mu_n, \mathbb{P}_n) = \underline{\theta}(\mu_n, P) + O_{\mathcal{P}}\left(\frac{1}{\mu_n\sqrt{n}}\right), \quad (27)$$

$$\underline{\lambda}(\mu_n, \mathbb{P}_n) = \underline{\lambda}(\mu_n, P) + O_{\mathcal{P}}\left(\frac{1}{\sqrt{n}}\right), \quad (28)$$

$$\underline{\sigma}(\mu_n, \mathbb{P}_n) = \underline{\sigma}(\mu_n, P) + o_{\mathcal{P}}(1). \quad (29)$$

Proof. See Appendix 8.4. □

2.4 Point-wise valid confidence sets.

Now we can combine Theorems 1 and 2 to provide the following consistent and asymptotic normal estimator of $\underline{v}(P)$,

$$\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n) \triangleq \underline{v}(\mu_n, \mathbb{P}_n) - \mu_n \|\underline{\theta}(\kappa_n, \mathbb{P}_n)\|^2. \quad (30)$$

Theorem 1 guarantees that the corresponding parameter $\underline{v}^{in}(\mu, \kappa, P)$ bound becomes tight if κ_n and μ_n are sufficiently small. Theorem 2 implies that if κ_n converges to zero slower than μ_n and both converge slower than $1/\sqrt{n}$, then $\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n)$ is asymptotically normal estimator with variance $\underline{\sigma}(\mu_n, P)$. Using the bias corrected estimator $\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n)$ and its analog for the upper bound, $\bar{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n)$, I construct the following Delta-method confidence sets:

$$\begin{cases} \text{CB}_{\alpha,n} &= [\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n) - z_{1-\alpha}n^{-1/2}\underline{\sigma}(\mu_n, \mathbb{P}_n), \infty), \\ \text{CI}_{\alpha,n}^{\theta_1} &= [\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n) - z_{1-\alpha}n^{-1/2}\underline{\sigma}(\mu_n, \mathbb{P}_n); \bar{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n) + z_{1-\alpha}n^{-1/2}\bar{\sigma}(\mu_n; \mathbb{P}_n)], \\ \text{CI}_{\alpha,n}^{\mathcal{S}} &= \text{CI}_{\alpha/2,n}^{\theta_1}, \end{cases} \quad (31)$$

where $z_{1-\alpha}$ is $1 - \alpha$ quantile of the standard Gaussian distribution. Here $\text{CB}_{\alpha,n}$ is a one-sided confidence band for θ_1 , $\text{CI}_{\alpha,n}^{\theta_1}$ is a two-sided confidence interval that covers any θ_1 in the identified set, and $\text{CI}_{\alpha,n}^{\mathcal{S}}$ is a two-sided confidence interval that covers the entire identified set $\mathcal{S}(P)$ based on the Bonferroni inequality.²²

Theorem 3. Suppose that Assumptions 1–5 hold and that in addition $0 < \alpha < 1/2$, μ_n and κ_n are such that $\kappa_n \rightarrow 0$, $\mu_n/\kappa_n \rightarrow 0$ and $\mu_n\sqrt{n} \rightarrow \infty$. Moreover, suppose that $\lim_{n \rightarrow \infty} \underline{\sigma}^2(\mu_n, P) > 0$ and $\lim_{n \rightarrow \infty} \bar{\sigma}^2(\mu_n, P) > 0$. Then,

$$\begin{aligned} \lim_{n \rightarrow \infty} P(\mathcal{S}(P) \subset \text{CB}_{\alpha,n}) &= \liminf_{n \rightarrow \infty} \inf_{\theta \in \Theta(P)} P(\theta_1 \in \text{CB}_{\alpha,n}) = 1 - \alpha, \\ \lim_{n \rightarrow \infty} P(\mathcal{S}(P) \subset \text{CI}_{\alpha,n}^{\mathcal{S}}) &\geq 1 - \alpha, \quad \liminf_{n \rightarrow \infty} \inf_{\theta \in \Theta(P)} P(\theta_1 \in \text{CI}_{\alpha,n}^{\mathcal{S}}) \geq 1 - \alpha. \end{aligned}$$

If the model has no equality constraints, i.e. if $p = 0$, then

$$\liminf_{n \rightarrow \infty} \inf_{\theta \in \Theta(P)} P(\theta_1 \in \text{CI}_{\alpha,n}^{\theta_1}) = 1 - \alpha. \quad (32)$$

²²One can develop tighter confidence sets using the joint asymptotic normality of the estimators $\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n)$ and $\bar{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n)$. I leave this exercise for future work.

Proof. See Appendix 8.5. □

The confidence band $\text{CB}_{\alpha,n}$ and interval $\text{CI}_{\alpha,n}^{\theta_1}$ are asymptotically non-conservative, at least for a fixed DGP, i.e. they have coverage of exactly $1 - \alpha$. If $p > 0$, θ_1 can be point identified if there are equality constraints in the model that are orthogonal to e_1 . So in this case I recommend the Bonferroni-type confidence set $\text{CI}_{\alpha,n}^S$ which remains valid under point identification. The shorter $\text{CI}_{\alpha,n}^{\theta_1}$ proposed by Imbens and Manski (2004) is valid only if θ_1 is not point-identified.

The theory for optimal choice of tuning parameter is beyond the scope of this paper. The following considerations, however, can provide some guidance for the optimal choice. Theorem 1 suggests that the tuning parameters should be smaller than $\bar{\mu}(P)$ to avoid the bias in the first order asymptotic distribution. This choice is infeasible since $\bar{\mu}(P)$ is unknown (moreover, it is a discontinuous function of $\mathbb{E}_P W$ and hence cannot be consistently estimated) so one has to let κ_n and μ_n go to zero. The optimal rates of κ_n and μ_n should balance the higher order variance and the worst case bias. A reasonable choice is to set $\mu_n = \kappa_n^{1/2} n^{-1/4}$ and let κ_n go to zero at a slow rate, say $n^{-1/4}$. A specific choice of the tuning parameters is further discussed in Section 5.

Finally, note that if the limiting variance $\lim_{n \rightarrow \infty} \sigma^2(\mu_n, P) = 0$, then the Gaussian limiting distribution does not provide a the coverage probability which in this case is governed by the distribution of the higher order terms. In the next subsection I will discuss how to relax this requirement.

2.5 Uniform confidence sets

Theorem 3 provides asymptotic coverage probability for a given DGP with measure P . The size of the sample required to achieve the nominal coverage of $1 - \alpha$ with a given precision in this result can depend on P through $\bar{\mu}(P)$ defined in Theorem 1. This cut-off $\bar{\mu}(P)$ can be arbitrarily close to zero so it is possible to construct an example where sequence of measures P_n that meet the assumptions of Theorem 3 but such that

$$\sqrt{n}\mu_n(\|\underline{\theta}(\mu_n, P_n)\| - \|\underline{\theta}(\kappa_n, P_n)\|) \rightarrow +\infty.$$

In other words, there are examples of DGP with a measure P and some $\epsilon > 0$ such that for any n it is possible to find a measure Q in a neighborhood of P with

$$Q(\mathcal{S}(P) \subset \text{CB}_{\alpha,n}) < 1 - \alpha - \epsilon.$$

In practical terms it means that the large sample theory with a fixed P may provide a poor approximation for the true coverage probability.

This feature of the confidence sets from the Theorem 3 should not come as a surprise. Parameter of interest, $\underline{v}(P)$ is a non-differentiable function of $\mathbb{E}_P W$ which in turn is a parameter of a locally asymptotically normal model. By the impossibility theorem of Hirano and Porter (2012) $\underline{v}(P)$ does not have a locally unbiased estimator. This impossibility result justifies use of half-median unbiased estimators used in Chernozhukov et al. (2013), i.e. such that have median value smaller or equal (larger or equal) than the parameter of interest with probability. Such estimators can be used for uniformly valid one-sided inference on θ_1 . In fact, Theorems 1 and 2 imply that $\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n)$ has asymptotic distribution with median non-smaller than $\underline{v}(P)$ which makes it half-median unbiased, but in the direction that results in the worst-case coverage probability below $1 - \alpha$. The outer

bound provides another estimator,

$$\underline{v}^{out}(\mu_n, \mathbb{P}_n) \triangleq \underline{v}(\mu_n, \mathbb{P}_n) - \mu_n \|\theta^*(\mathbb{P}_n)\|^2, \quad (33)$$

that is also half-median unbiased in the opposite direction if $\theta^*(\mathbb{P}_n)$ is appropriately chosen. To achieve this goal, the estimator $\theta^*(\mathbb{P}_n)$ should have larger norm than the point with the minimal norm in $\underline{\theta}(P)$ with probability approaching 1. We can use the analog estimator $\theta^*(\mathbb{P}_n)$ that converges to the following point,

$$\theta_i^*(P) \triangleq \max\{\theta_i^+(P), \theta_i^-(P)\}, \quad (34)$$

where

$$\theta_i^\pm(P) \triangleq \left| \min_{\theta \in \Theta(P), \theta_1 \leq \underline{v}(P) + \mu_n} \{\pm \theta_i\} \right|. \quad (35)$$

By definition, $\|\theta^*\| \geq \|\theta\|$ for any $\theta \in \underline{\theta}(P)$.²³ This bound on $\|\theta^*\|$ has two attractive properties. First, it can be (uniformly) consistently estimated using only $2k$ linear programs, so that it can be computed in models with a very large dimension and a large number of inequalities using interior point numerical optimization methods. Second, if $\underline{\theta}(P)$ is a singleton, then $\|\theta^*\| \rightarrow \|\underline{\theta}(P)\|$. So for any such fixed P by Theorem 1 we have $\underline{v}^{out}(\mu_n, P) = \underline{v}(P)$ for sufficiently small μ_n , i.e. the corresponding confidence intervals will have the correct coverage.

Before I introduce the uniformly valid confidence sets, I would like to address the difficulty resulting from the degenerate Gaussian distribution mentioned in the end of the previous section. To do so, I consider a regularized estimator of the asymptotic variance,

$$\hat{\sigma}_n^{reg} \triangleq \max\{\underline{\sigma}(\mu_n, \mathbb{P}_n), \sigma_0\},$$

where σ_0 is some small positive number.

Let $CB_{\alpha, n, \mathcal{P}}$ and $CI_{\alpha, n, \mathcal{P}}^S$ be the confidence sets defined in (31) with $\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n)$ and $\underline{\sigma}(\mu_n, \mathbb{P}_n)$ being replaced by $\underline{v}^{out}(\mu_n, \mathbb{P}_n)$ and $\hat{\sigma}_n^{reg}$, correspondingly. As before, let \mathcal{P} contain all measures P that satisfy Assumptions 1-5 with some uniform positive constants $\underline{\eta}, \underline{\sigma}, \varepsilon, \bar{M}$.

Theorem 4. *Suppose that Assumption 4 holds. In addition, suppose that $0 < \alpha < 1/2$, $\mu_n \rightarrow 0$ and $\mu_n \sqrt{n} \rightarrow \infty$. Then the following results hold,*

$$\begin{aligned} \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}} P(\mathcal{S}(P) \subset CB_{\alpha, n, \mathcal{P}}) &= \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}} \inf_{\theta \in \Theta(P)} P(\theta_1 \in CB_{\alpha, n, \mathcal{P}}) \geq 1 - \alpha, \\ \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}} P(\mathcal{S}(P) \subset CI_{\alpha, n, \mathcal{P}}^S) &\geq 1 - \alpha, \quad \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}} \inf_{\theta \in \Theta(P)} P(\theta_1 \in CI_{\alpha, n, \mathcal{P}}^S) \geq 1 - \alpha. \end{aligned}$$

Proof. See Appendix 8.4. □

Note that the worst case asymptotic coverage probability of $CB_{\alpha, n, \mathcal{P}}$ is exactly equal to $1 - \alpha$ for a fixed regular DGP, i.e. such that $\underline{\theta}(P)$ is a singleton and $\lim_{n \rightarrow \infty} \bar{\sigma}^2(\mu_n, P) > 0$.

3 Extensions

In this section I outline how to use the regularized support function approach to construct joint confidence sets for multiple components of θ and tackle the intersection bound problem. The

²³See Lemma 12 in Appendix.

latter problem appears once we represent the value of program (7) as a maximum over values of its subproblems. Such subproblems may meet Assumptions 2-3 even if the full problem does not.

3.1 Joint confidences sets for multiple parameters

It is trivial to extend the analysis to

$$\underline{v}(P; a) = \min_{\theta \in \Theta(P)} a' \theta$$

for any $a \in R^d$ with $\|a\| = 1$. Indeed, Assumptions 1-5 are invariant with respect to orthogonal transformations of the coordinates, i.e. they are satisfied for the following program (with $\tilde{\theta} = U' \theta$, $\tilde{A}_P = A_P U$ and $a' = e'_1 U$ for any orthogonal matrix U)

$$\underline{v}(P; a) = \min_{\theta \in \mathbb{R}^d} e'_1 \tilde{\theta} \tag{36}$$

$$\text{s.t. } \begin{cases} e'_j \tilde{A}_P \tilde{\theta} = e'_j b_P, & j \in \mathcal{J}^{eq}, \\ e'_j \tilde{A}_P \tilde{\theta} \leq e'_j b_P, & j \in \mathcal{J}^{ineq}, \end{cases} \tag{37}$$

One can think about \tilde{A}_P as a coefficient matrix under a different measure \tilde{P} , $A_{\tilde{P}}$. The set of measures \mathcal{P} from Section 2.5 includes \tilde{P} corresponding to all orthogonal transformations of A_P .

The identified set $\Theta(P)$ is convex so any projection of it can be characterized using support functions. One can construct a joint confidence set for $\Theta(P)$ as follows. For any set of directions $\mathcal{A} \subset R^d$ take

$$CS_{\alpha, n}^{\mathcal{A}} = \{\theta | a \in \mathcal{A}, a' \theta \leq -\underline{v}^{out}(\mu_n, \mathbb{P}_n; -a) + c_{1-\alpha} n^{-1/2} \max\{\underline{\sigma}(\mu_n, \mathbb{P}_n; -a), \sigma_0\}\},$$

where $c_{1-\alpha}$ is $1 - \alpha$ quantile of the maximum of the corresponding Gaussian process that can be estimated using multiplier bootstrap enabled by the asymptotic linear representation, (25).

By appropriately choosing the set of directions \mathcal{A} we can construct joint confidence sets for projections of $\Theta(P)$ on any subvectors θ . If \mathcal{A} has finitely many elements, $CS_{\alpha, n}^{\mathcal{A}}$ is a polygon. So we can plot it directly without performing test inversion as in the one-dimensional case. The confidence set $\tilde{CI}_{\alpha, n}^B$ is a particular case of $CS_{\alpha, n}^{\mathcal{A}}$ corresponding to $\mathcal{A} = \{e_1, -e_1\}$ and the Bonferroni estimate of $c_{1-\alpha}$. The following example provides another interesting set of directions.

Example 5 (Natural joint confidence set). It seems natural to construct a joint confidence set for θ based on directions corresponding to the normal vectors of the moment conditions. For simplicity assume that $p = 0$. The original system (2) may have some inequalities that are slack for any point $\theta \in \Theta(P)$. We can characterize the identified set $\Theta(P)$ as solution to a tight system of inequalities

$$e'_j A_P \theta \leq \underline{b}_j, j \in \mathcal{J}^{ineq}. \tag{38}$$

where

$$\underline{b}_j = \max_{(\vartheta, \theta) \in \mathbb{R}^{d+1}} \vartheta \tag{39}$$

$$\text{s.t. } \begin{cases} \vartheta & = e'_j A_P \theta, \\ e'_\ell A_P \theta & \leq e'_\ell b_P, \ell \in \mathcal{J}^{ineq}. \end{cases} \tag{40}$$

Every inequality in System (38) is active at least in one point in $\Theta(P)$ (any point in the argmax of (39)). Programs (39) meet Assumptions 1-5 and so the outer estimators \hat{b}_j^{out} are half-median unbiased with corresponding standard error estimators $\hat{\sigma}_j$. Then the following polyhedron confidence set will cover any point $\theta \in \Theta(P)$ with asymptotic probability at least $1 - \alpha$ uniformly over $P \in \mathcal{P}$,

$$CS_{\alpha,n}^{\mathcal{N}} = \{\theta | e_j' \hat{A}_P \theta \leq \hat{b}_j^{\text{out}} + c_{1-\alpha} n^{-1/2} \max\{\hat{\sigma}_j, \sigma_0\}, j \in \mathcal{J}^{\text{ineq}}\}.$$

The generalization to the case $p \neq 0$ is straightforward. \square

3.2 Dealing with overidentification

Another interesting extension concerns the case when Assumptions 1-3 are violated for $\Theta(P)$, but it can be represented as an intersection of sets $s = 1, \dots, L$, $\Theta^s(P)$ which satisfy these assumptions. Then

$$\underline{v}(P) = \min_{\theta \in \Theta(P)} e_1' \theta \geq \max_{s=1, \dots, L} \left(\min_{\theta \in \Theta^s(P)} e_1' \theta \right) \geq \max_{s=1, \dots, L} \underline{v}^{s, \text{out}}(\mu_n, P) \quad (41)$$

If we define $\underline{v}(P) = +\infty$ in the case $\Theta(P) = \emptyset$, then the bounds (41) are trivially valid too.

It is possible to make the first bound in (41) sharp if we consider all subsets of $d - p$ inequality restrictions (besides the ones that define Θ) as the following example shows.

Example 6 (Example 3 continued). Consider all $L = C(K, d) = C(4, 2) = 6$ subsystems $\Theta^s(P)$ of (8) corresponding to any two support points $z, z' \in \{0, 1\}^2$. We can get a sharp bound on $\underline{v}(P)$ in form of a maximum of the values of the subproblems,

$$\underline{v}(P) = \min_{\theta \in \Theta(P)} e_1' \theta = \max_{s=1, \dots, L} \left(\min_{\theta \in \Theta^s(P)} e_1' \theta \right).$$

Indeed, value $\underline{v}(P)$ is attained at some basic solution $\theta^{(s)}$. By definition, at such $\theta^{(s)}$ at least two inequalities are active. There exist a subproblem $\Theta^s(P)$ that contains the same two inequalities. By construction, $\theta_1^{(s)} = \min_{\theta \in \Theta^s(P)} e_1' \theta \square$.

Let $\underline{v}^{(s)}$ denote the solution to the subproblem s . Since every subproblem satisfies Assumptions 1-3, we get

$$\sqrt{n} \begin{pmatrix} \underline{v}^{(1)}(\mu_n, \mathbb{P}_n) - \underline{v}^{(1)}(\mu_n, P) \\ \dots \\ \underline{v}^{(L)}(\mu_n, \mathbb{P}_n) - \underline{v}^{(L)}(\mu_n, P) \end{pmatrix} \rightsquigarrow N(0, \Omega).$$

As before,

$$\Omega_{ij} = \mathbb{E}_P[\lambda^{(i)} g(W, \underline{\theta}^{(i)}) g(W, \underline{\theta}^{(j)})' \lambda^{(j)}].$$

Inference on (41) can be done using methods developed in Hall and Miller (2010) or Chernozhukov et al. (2013), which are based on bootstrap. Alternatively, we can use the following representation to reduce it to a linear program,

$$\max_{s=1, \dots, L} \underline{v}^{s, \text{out}} = \max_{\sum_{s=1}^L \gamma_s = 1, \gamma_s \geq 0} \gamma_s \underline{v}^{s, \text{out}}. \quad (42)$$

The solution γ to (42) is not unique in general. Program (42) satisfies the Assumptions 1-3. In

order to restore the asymptotic normality of the corresponding estimator we need to regularize it,

$$\begin{aligned}
& - \min_{\gamma} \quad - \sum_{s=1}^L \gamma_j \underline{v}^{(j),out} + \mu_n \|\gamma\|^2 \\
& \text{s.t.} \quad \sum_{s=1}^L \gamma_j = 1, \gamma_j \geq 0.
\end{aligned} \tag{43}$$

The estimator of the value of this program also has a Bahadur representation which can be used to derive an estimator for the variance,

$$\text{sVar}\left(\sum_{s=1}^L \gamma_s \lambda^{(s)} g(w_i, \underline{\theta}^{(s)})\right).$$

The inner bias correction analogous to \underline{v}^{in} defined in Section 2.2 to the plug-in estimator of (43) would in this case make the corresponding confidence sets longer and preserve asymptotic coverage probability of at least $1 - \alpha$ uniformly in $P \in \cap_{s=1,L} \mathcal{P}^s$. The case with no-overidentification would result in zero bias uniformly. Case with over-identification can result in overcoverage along some DGP sequences but is not conservative for a fixed DGP if bound (41) is sharp.

Remark 1 (Specification testing). The subproblem representation (41) can provide a test for model specification, i.e. the test of Assumption 1. Under the misspecification, the case $\Theta(P) = \emptyset$, the lower bound on the minimum $\underline{v}(P)$ is larger than the analogous upper bound on the maximum $\bar{v}(P)$. It should always be possible to find subproblems with non-empty domains and finite values in the sample, $\underline{v}^s(\mathbb{P}_n)$ and $\bar{v}^{s'}(\mathbb{P}_n)$. The t-test of the hypothesis

$$\underline{v}^{s,out}(\mu_n, P) \leq \bar{v}^{s',out}(\mu_n, P)$$

can be interpreted as a specification test. I leave this extension for future work.

4 Discussion

4.1 Scope

There are at least two recent papers, BCS and KMS, to construct confidence sets that provide explicit classes of DGP with uniform asymptotic coverage probability. Both methods are applicable in non-linear moment inequality models. They however are more restrictive in some other dimensions and there are examples of affine moment inequality models that fall outside of their scope.

Both methods are using standardized moment conditions and hence restrict variance of the moment conditions to be strictly larger than some small constant (Definition 4.2.ii in BCS and Assumption 4.1 b (iii) in KMS). Moreover, the standardized moment conditions has to be differentiable (Assumption A.3.c in BCS and 4.4.i in KMS). Both Assumptions are also present in the pioneering AS paper. The following example illustrates violation of these assumptions in an affine models.

Example 7 (Example 1 continued). Suppose that for some support point z_0 with a positive mass the lower bound on the outcome is deterministic, i.e. $\mathbb{E}_P(\underline{Y}|Z = z_0) = y_0$ and $\text{Var}_P(\underline{Y}|Z = z_0) = 0$.

The corresponding moment inequality condition is

$$\mathbb{E}_P g(Y, Z, \theta) = \mathbb{E}_P [(Y - z'_0 \theta) 1 \{Z = z_0\}] \leq 0. \quad (44)$$

The standard deviation of the moment condition is $|z'_0 \theta - y_0| \sqrt{p_z (1 - p_z)}$, where $p_z = \mathbb{E}_P 1 \{z = z_0\}$. It is equal to zero for any solution of the linear equation $z'_0 \theta = y_0$. These values of θ make the constraint (44) binding, which is exactly the case when this constraint determines $\underline{v}(P)$. This moment condition is also non-differentiable at these points.

$$\partial_\theta g(y_0, z_0, \theta) = \text{sign}(y_0 - z'_0 \theta) \sqrt{p_z / (1 - p_z)}. \square$$

I also avoid imposing any restrictions on the correlation matrix of the moment conditions which is present in KMS (Assumption 4.3) and the polynomial minorant condition (Assumption A.3.a in BCS, present in CHT, avoided in KMS).

I do impose an explicit assumptions that guarantee the LICQ which is not present in KMS or BCS, but I only require them to hold in subproblems, as noted in Section 3.2. I would like to note here that the M-step in KMS uses standard Newton optimization routines that can find all the stationary (KKT) points. If there is no constraint qualification, however, KKT conditions are no longer providing the necessary conditions for the optimum, which in this case is a John-Fritz conditions.²⁴ As a result, E-A-M algorithm considered in KMS is not guaranteed to converge to the global optimum without some constraint qualification.

4.2 Computation properties

4.2.1 Fast convergence to a minimum

The existing uniform methods of AS, BCS and KMS are based on standardized moment conditions that are non-convex even if the original inequalities are affine in θ . Example 2 in the previous section illustrate this feature. The estimator $\underline{\theta}(\mu_n, \mathbb{P}_n)$ is a solution to a strictly convex quadratic program for any affine moment inequality model. For convex programs the set Karush–Kuhn–Tucker (KKT) conditions²⁵ provide necessary and sufficient conditions for the global optimum. Moreover, convex quadratic programs can be solved using an interior point algorithm with a polynomial rate of convergence.²⁶ This strict convexity gives a dramatically faster rate of obtaining the optimum than the ones used in BCS and KMS. These methods are based on non-convex constraint optimization problems, which are NP-hard. Section 5 compares computational time in specific examples.

4.2.2 Uniqueness of a global optimum

KKT system for strictly convex optimization problems has a unique solution. The number of KKT points of the optimization problems in the KMS, BCS and AS procedures in affine moment inequality models can be large and typically grows exponentially with the dimension d and number of inequalities k . The following example illustrates this point.

²⁴See Section 5.2.2 in BS(2000)

²⁵See Lemma 3 in Appendix

²⁶See, for example, Ye and Tse (1989).

Example 8. Consider a set of moment inequalities with coefficients that have expectation

$$\mathbb{E}_P W = \begin{pmatrix} -I_d & -\iota \\ I_d & -\iota \end{pmatrix}.$$

Suppose that components of W are independent and have the same variance s^2 . $\Theta(P)$ is a box $[-1, 1]^d$. The standardized moment conditions take the form

$$\frac{\pm\theta_j + 1}{s\sqrt{1 + \|\theta\|^2}} \leq 0, \quad j = 1, \dots, d. \quad (45)$$

The KMS procedure adds slack $c(\theta)$ to the right hand side of every standardized moment inequality. Consider , for example, $j = 1$,

$$\theta_1 \geq 1 - c(\theta) s\sqrt{1 + \|\theta\|^2}. \quad (46)$$

The slack function $c(\theta)$ is computed using a resampling on a grid of points. Assume, for simplicity, that $c(\theta)$ is a constant, for example, provided by the Bonferroni approach. Figure 2 shows the identified set and the corresponding expansion with $c(\theta) = \text{const}$. The optimization domain of the E-A-M algorithm in KMS is similar the non-convex set on right of Figure 2. Every vertex of the $[-1, 1]^d$ with $\theta_1 = -1$ corresponds to an isolated local minimum of the optimization procedure in KMS. Correspondingly, the number of local minima grows exponentially with the dimension d . For example, the number of local minima for $d = 10$ is 512. The growth in the number of local optima is even faster in models with more than 2 inequalities per coordinate. \square

Multiplicity of KKT points makes the procedures of KMS, AS and BCS both computationally costly does not provide guarantees of convergence to a global optimum for large d .

4.2.3 Multiplier bootstrap

The proposed estimators of the regularized support functions have a Bahadur representation with explicit influence functions. One can use this property to justify multiplier bootstrap for inference on support of the identified set. The main advantage of this approach is that it allows one to solve the mathematical programs only once. For example, FH and KMS solve mathematical programs repeatedly for every bootstrap sample. The multiplier bootstrap approach is particularly appealing in subvector inference on more that one component as discussed in Section 3.1, since one has repeat computations for various directions a .

4.2.4 Implementation

The point-wise CIs in (31) can be computed using any Newton type optimization software that provides accurate Lagrange multipliers. I use *fmincon* function of MATLAB software. I recommend using the *'active set'* or *'SQP'* options since the *'interior point'* solver does not provides accurate Lagrange multipliers. The estimator $\|\theta^*(\mathbb{P}_n)\|^2$ which enters the uniformly valid CS described in Theorem 4 is based on $2d$ linear programs. Linear programs typically scale very well.²⁷

²⁷A bench-marking of results for the state-of-art commercial LP solvers can be found, for example, at <http://plato.asu.edu/bench.html>. The commercial solvers can tackle LP with tens thousands of constraints and variables in a matter of minutes. Matlab's linprog solver tends to underperform in the time comparison.

The extended procedure outlined in Section 3.2 requires finding subproblems that make the bound (41) sharp. This can be achieved if Assumption 2 is satisfied for the original program and one considers all subproblems with exactly d moment conditions. Such an exhaustive approach may require considerable computational resources if d is large. Potentially, one could use a moment selection procedure that would restrict attention to combinations of moments inequalities that are close to be binding at point $\underline{\theta}(\mu_n, \mathbb{P}_n)$. Such an extension is beyond the scope of this paper.

5 Monte Carlo

5.1 Designs

I examine two-dimensional and multi-dimensional designs. The two-dimensional design is based on the identified set defined by four moment inequality conditions with the following coefficients:

$$\mathbb{E}_P W = \begin{pmatrix} -\cos\left(\frac{\omega\pi}{180}\right) & -\sin\left(\frac{\omega\pi}{180}\right) & \cos\left(\frac{\omega\pi}{180}\right) + \sin\left(\frac{\omega\pi}{180}\right) \\ \cos\left(\frac{\omega\pi}{180}\right) & \sin\left(\frac{\omega\pi}{180}\right) & \cos\left(\frac{\omega\pi}{180}\right) + \sin\left(\frac{\omega\pi}{180}\right) \\ 0 & -1 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

The shape of this set is a parallelogram analogous to the one on Figure 1, case $\rho \geq 0$. The parameter $\omega \in [0^\circ, 36^\circ]$ defines the angle between the normal vectors of the rear sides of the parallelogram and the horizontal axis. The value $\omega = 0^\circ$ corresponds to a square-shaped identified set. In vicinity of $\omega = 0^\circ$ $\text{CB}_{\alpha, N}$ may have coverage probability below the nominal level because it lacks uniform validity of $\text{CB}_{\alpha, N, \mathcal{P}}$. The expectations $\mathbb{E}_P W$ are parametrized to guarantee $\underline{\theta}_1 = \underline{\theta}_2 = -1$ and $\bar{\theta}_1 = \bar{\theta}_2 = 1$ for all values of ω . The components of W_i are independent Gaussian random variables with variance $s_2^2 = 0.01$. For each value of ω , I compute the frequency of coverage and the average length in excess of the identified set for $\text{CB}_{\alpha, N}$ and $\text{CB}_{\alpha, N, \mathcal{P}}$ based on sample sizes $N \in \{100, 1000, 10000\}$. Number of MC simulations is 1000 for every combination of N and ω .

The d -dimensional design (values of d under consideration are $d = 2, \dots, 16$, $d = 50$ and $d = 100$) is based on the value of $\mathbb{E}_P W$ from Example 8. As before, random matrices W_i consist of independent Gaussian random variables, but variance of each component now changes with the dimension of the problem as $s_d^2 = 0.09/1 + 4d$. It is done to keep lengths of confidence bounds comparable across different dimensions. I use samples of $N = 1000$ and compute frequency of coverage, average length in excess of 2 and computational time for $\text{CB}_{\alpha, N}$, $\text{CB}_{\alpha, N, \mathcal{P}}$. As a benchmark, I use $\text{CB}_{\alpha, N, AS}$, one-sided confidence bounds for $\underline{\theta}_1$ based on Andrews and Soares (2010) with Bonferroni critical values implemented using the fast E-A-M algorithm of Kaido et al. (2015). This choice of the benchmark is the fastest available uniformly valid procedure in the literature.²⁸ I use 100 MC simulations for $\text{CB}_{\alpha, N}$ and $\text{CB}_{\alpha, N, \mathcal{P}}$ and 20 simulations for $\text{CB}_{\alpha, N, AS}$ (due to its higher computational cost).²⁹

²⁸The two alternative approaches, Bugni et al. (2016) and Kaido et al. (2015), can potentially provide uniformly-valid confidence sets with shorter average length. Both of them, however are expected to be considerably slower because they add a profiling or calibration step.

²⁹The code is available on <https://molinari.economics.cornell.edu/programs.html>. Note that this implementation of AS procedure requires additional constraint qualification assumption, which was not made explicitly in KMS.

5.2 Choice of the tuning parameters

The theory for optimal choice of the tuning parameter is beyond the scope of this paper. In the Monte Carlo exercise the following tuning parameters performed particularly well, $\mu_n = \hat{\mu}_1 \sqrt{\frac{\log n}{n}}$ and $\kappa_n = \hat{\mu}_1 \sqrt{\frac{\log \log n}{n}}$ with $\hat{\mu}_1 = \hat{\mu}^{in} \triangleq \sqrt{\text{tr}(s\text{Var}(\underline{\lambda}'(0, \mathbb{P}_n)w_i))}$ for $\text{CB}_{\alpha, N}$ and $\hat{\mu}_1 = \hat{\mu}^{out} \triangleq \hat{\mu}^{in} / \max \{ \|\theta^*(\mathbb{P}_n)\|^2, 1 \}$ for $\text{CB}_{\alpha, N, \mathcal{P}}$. The size of $\hat{\mu}^{in}$ captures the sample variation of the relevant data: the noisier data requires stronger regularization. Division by $\max \{ \|\theta^*(\mathbb{P}_n)\|^2, 1 \}$ makes the worst case bias of $\underline{v}^{out}(\mu_n, \mathbb{P}_n)$ independent of the scale of the identified set. This re-scaling considerably reduced the average length of the confidence set in my simulations.

5.3 Results

Figures 3 and 4 for the two dimensional design illustrate the difference in the coverage for the point-wise and uniformly-valid confidence bounds.

First note that average length and coverage frequency for both $\text{CB}_{\alpha, N}$ and $\text{CB}_{\alpha, N, \mathcal{P}}$ are almost indistinguishable for $\omega > 3^\circ$ even for $N = 100$. This is due to the fact that both confidence bounds are asymptotically equivalent and have correct coverage of 95% in the regular case (see Section 2.5).

Second, for values $\omega < 3^\circ$ the average length for $\text{CB}_{\alpha, N}$ and $\text{CB}_{\alpha, N, \mathcal{P}}$ are apparently different. In particular, length of $\text{CB}_{\alpha, N}$ for $\omega = 0.9^\circ$ and 1.8° is shorter than for either $\omega = 0^\circ$ or $\omega = 2.7^\circ$, which results in the coverage frequency below the nominal level of 95%. This reflects the positive bias in $\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n)$ that makes the corresponding confidence bound unreliable for small but positive ω . In contrast, the average length of $\text{CB}_{\alpha, N, \mathcal{P}}$ spikes at $\omega = 0^\circ$ which results in coverage close to 100%. In most practical situations the true value of ω is unknown so $\underline{v}^{out}(\mu_n, \mathbb{P}_n)$ is the preferred option for confidence bounds despite its larger length, in particular in small samples.³⁰

Third, the point-wise valid bound $\text{CB}_{\alpha, N}$ performs reasonably well even for small samples. The coverage frequency for the considered grid values of ω is only moderately below the nominal level (82% instead of 95%). Moreover as sample size grows, the converge frequency approaches 95%. So the size of the problematic neighborhood of $\omega = 0$ shrinks with as the number of data grows.

The results for d -dimensional design are given on Figures 5-7. The first observation is that the coverage frequency for $\text{CB}_{\alpha, N}$ is reliably close to 95% for all even for large dimensions. The coverage frequency for $\text{CB}_{\alpha, N, \mathcal{P}}$ and $\text{CB}_{\alpha, N, AS}$ was equal to 100% for all d . Figure 5 shows that the negative bias of $\underline{v}^{out}(\mu_n, \mathbb{P}_n)$ grows with the dimension of the problem. The length is comparable to that of $\text{CB}_{\alpha, N, AS}$, which grows as $\log(d)$ due to the Bonferroni-correction of the critical values. Figure 5 suggests that in very high-dimensional cases the uniform validity comes at big costs in terms of the length (for $d = 50$ the average excess length of $\text{CB}_{\alpha, N}$ is almost 14 times shorter than that of $\text{CB}_{\alpha, N, \mathcal{P}}$). The substantially shorter average length may justify use of the point-wise valid $\text{CB}_{\alpha, N}$ in such cases.

Finally, Figure 7 shows that the computational time of $\text{CB}_{\alpha, N, \mathcal{P}}$ grows at a very slow rate. It takes only a second to compute $\text{CB}_{\alpha, N, \mathcal{P}}$ for $d = 15$ (the average computational time for $d = 50$ and 100 is 4.3 and 20 seconds, correspondingly). The computational time for the AS procedure increases by approximately 30% with every additional dimension and takes 630 seconds to compute the CI for $d = 15$. With estimated growth rate of 30% per dimension the KMS procedure with

³⁰An exception to this rule is the natural joint confidence polygon in Example 5. There the objective function is orthogonal to the moment inequalities by construction which makes $\underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n)$ preferable to $\underline{v}^{out}(\mu_n, \mathbb{P}_n)$.

precompiled code would take more than hour to compute a CS for $d = 15$, while $d > 30$ would be practically infeasible.

6 Conclusion

This paper shows that the regularization approach provides a fast way to construct both point-wise and uniform confidence sets for θ_1 that has comparable or shorter length to those in the existing literature. Moreover, the confidence remain valid in some situations where the existing procedures cannot be used. Monte Carlo simulations show that the proposed confidence sets have good finite sample coverage properties. The computational benefits of the new approach are particularly prominent if the dimension of θ is large. The general framework can be extended in the number of ways to allow for overidentification and joint inference. My approach is attractive in applications like linear model with interval-valued outcome variable and a large number of regressors.

References

- ANDREWS, D. W. AND X. SHI (2013): “Inference based on conditional moment inequalities,” *Econometrica*, 81, 609–666.
- ANDREWS, D. W. 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*, 763–814.
- BERGE, C. (1963): *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*, Courier Corporation.
- BONNANS, J. F. AND A. SHAPIRO (2000): *Perturbation Analysis of Optimization Problems*, Springer Science & Business Media.
- BONTEMPS, C., T. MAGNAC, AND E. MAURIN (2012): “Set identified linear models,” *Econometrica*, 80, 1129–1155.
- BOYD, S. AND L. VANDENBERGHE (2004): *Convex optimization*, Cambridge university press.
- BUGNI, F., I. CANAY, AND X. SHI (2016): “Inference for functions of partially identified parameters in moment inequality models,” Tech. rep., cemmap working paper, Centre for Microdata Methods and Practice.
- CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): “Estimation and confidence regions for parameter sets in econometric models,” *Econometrica*, 75, 1243–1284.
- CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): “Intersection Bounds: Estimation and Inference,” *Econometrica*, 81, 667–737.
- DÜMBGEN, L. (1993): “On nondifferentiable functions and the bootstrap,” *Probability Theory and Related Fields*, 95, 125–140.

- FANG, Z. AND A. SANTOS (2018): “Inference on Directionally Differentiable Functions,” *The Review of Economic Studies*, rdy049.
- FISCHER, A. (1992): “A special Newton-type optimization method,” *Optimization*, 24, 269–284.
- FREYBERGER, J. AND J. L. HOROWITZ (2015): “Identification and shape restrictions in nonparametric instrumental variables estimation,” *Journal of Econometrics*, 189, 41 – 53.
- GAFAROV, B. (2017): “Essays on Partially Identified Models,” Ph.D. thesis, Pennsylvania State University.
- GAFAROV, B., M. MEIER, AND J. L. M. OLEA (2018): “Delta-Method inference for a class of set-identified SVARs,” *Journal of Econometrics*, 203, 316–327.
- GOLISHNIKOV, M. AND A. F. IZMAILOV (2006): “Newton-type methods for constrained optimization with nonregular constraints,” *Computational Mathematics and Mathematical Physics*, 46, 1299–1319.
- HAILE, P. A. AND E. TAMER (2003): “Inference with an incomplete model of English auctions,” *Journal of Political Economy*, 111, 1–51.
- HALL, P. AND H. MILLER (2010): “Bootstrap confidence intervals and hypothesis tests for extrema of parameters,” *Biometrika*, 97, 881–892.
- HIRANO, K. AND J. R. PORTER (2012): “Impossibility results for nondifferentiable functionals,” *Econometrica*, 1769–1790.
- HONG, H. AND J. LI (2018): “The numerical delta method,” *Journal of Econometrics*, 206, 379–394.
- HONORÉ, B. E. AND A. LLERAS-MUNEY (2006): “Bounds in competing risks models and the war on cancer,” *Econometrica*, 74, 1675–1698.
- HONORÉ, B. E. AND E. TAMER (2006): “Bounds on parameters in panel dynamic discrete choice models,” *Econometrica*, 74, 611–629.
- IMBENS, G. W. AND C. F. MANSKI (2004): “Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 72, 1845–1857.
- JANSSON, M. AND D. POUZO (2017): “Some Large Sample Results for the Method of Regularized Estimators,” *arXiv preprint arXiv:1712.07248*.
- KAIDO, H., F. MOLINARI, AND J. STOYE (2015): “Inference for projections of identified sets,” *manuscript*.
- KAIDO, H. AND A. SANTOS (2014): “Asymptotically Efficient Estimation of Models Defined by Convex Moment Inequalities,” *Econometrica*, 82, 387–413.
- KASY, M. (2016): “Partial identification, distributional preferences, and the welfare ranking of policies,” *Review of Economics and Statistics*, 98, 111–131.
- KLINE, P. AND M. TARTARI (2016): “Bounding the labor supply responses to a randomized welfare experiment: A revealed preference approach,” *American Economic Review*, 106, 972–1014.

- LAFFÉRS, L. (2018): “Bounding average treatment effects using linear programming,” *Empirical Economics*.
- MANSKI, C. F. (2003): *Partial identification of probability distributions*, Springer Science & Business Media.
- MANSKI, C. F. AND J. V. PEPPER (2000): “Monotone instrumental variables: With an application to the returns to schooling,” *Econometrica*, 68, 997–1010.
- MANSKI, C. F. AND E. TAMER (2002): “Inference on regressions with interval data on a regressor or outcome,” *Econometrica*, 70, 519–546.
- MINCER, J. (1974): “Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2.” .
- MOLINARI, F. (2008): “Partial identification of probability distributions with misclassified data,” *Journal of Econometrics*, 144, 81 – 117.
- OK, E. A. (2007): *Real analysis with economic applications*, vol. 10, Princeton University Press.
- PAKES, A., J. PORTER, K. HO, AND J. ISHII (2015): “Moment inequalities and their application,” *Econometrica*, 83, 315–334.
- ROBINSON, S. M. (1977): “A characterization of stability in linear programming,” *Operations Research*, 25, 435–447.
- RUSSELL, T. M. (2017): “Sharp Bounds on Functionals of the Joint Distribution in the Analysis of Treatment Effects,” .
- SHAPIRO, A. (1991): “Asymptotic analysis of stochastic programs,” *Annals of Operations Research*, 30, 169–186.
- (1993): “Asymptotic behavior of optimal solutions in stochastic programming,” *Mathematics of Operations Research*, 18, 829–845.
- SHAPIRO, A., D. DENTCHEVA, AND A. RUSZCZYNSKI (2014): *Lectures on stochastic programming: modeling and theory*, vol. 16, SIAM.
- SHI, X., M. SHUM, AND W. SONG (2018): “Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity,” *Econometrica*, 86, 737–761.
- STEWART, M. B. (1983): “On least squares estimation when the dependent variable is grouped,” *The Review of Economic Studies*, 50, 737–753.
- SYRGKANIS, V., E. TAMER, AND J. ZIANI (2017): “Inference on Auctions with Weak Assumptions on Information,” *arXiv preprint arXiv:1710.03830*.
- TORGOVITSKY, A. (2016): “Nonparametric inference on state dependence with applications to employment dynamics,” .
- (2018): “Partial identification by extending subdistributions,” .

- TROSTEL, P., I. WALKER, AND P. WOOLLEY (2002): “Estimates of the economic return to schooling for 28 countries,” *Labour economics*, 9, 1–16.
- VAN DER VAART, A. AND J. WELLNER (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Science & Business Media.
- VAN DER VAART, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.
- WACHSMUTH, G. (2013): “On LICQ and the uniqueness of Lagrange multipliers,” *Operations Research Letters*, 41, 78–80.
- YE, Y. AND E. TSE (1989): “An extension of Karmarkar’s projective algorithm for convex quadratic programming,” *Mathematical programming*, 44, 157–179.
- YURINSKII, V. V. (1978): “On the error of the Gaussian approximation for convolutions,” *Theory of Probability & Its Applications*, 22, 236–247.

Appendix

7 Figures

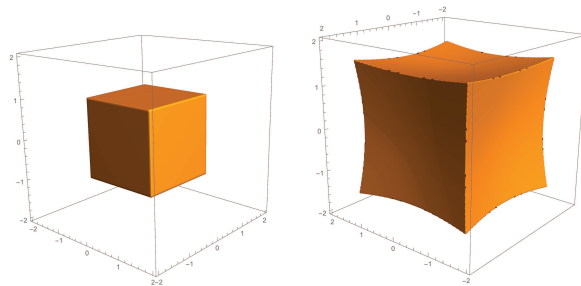


Figure 2: The identified set and the corresponding optimization domain of KMS procedure for $d = 3$ in Example 8.

Figure 3: Average excess length in the two-dimensional design for $CB_{\alpha,N}$ and $CB_{\alpha,N,\mathcal{P}}$ for various sample sizes as a function of angle ω

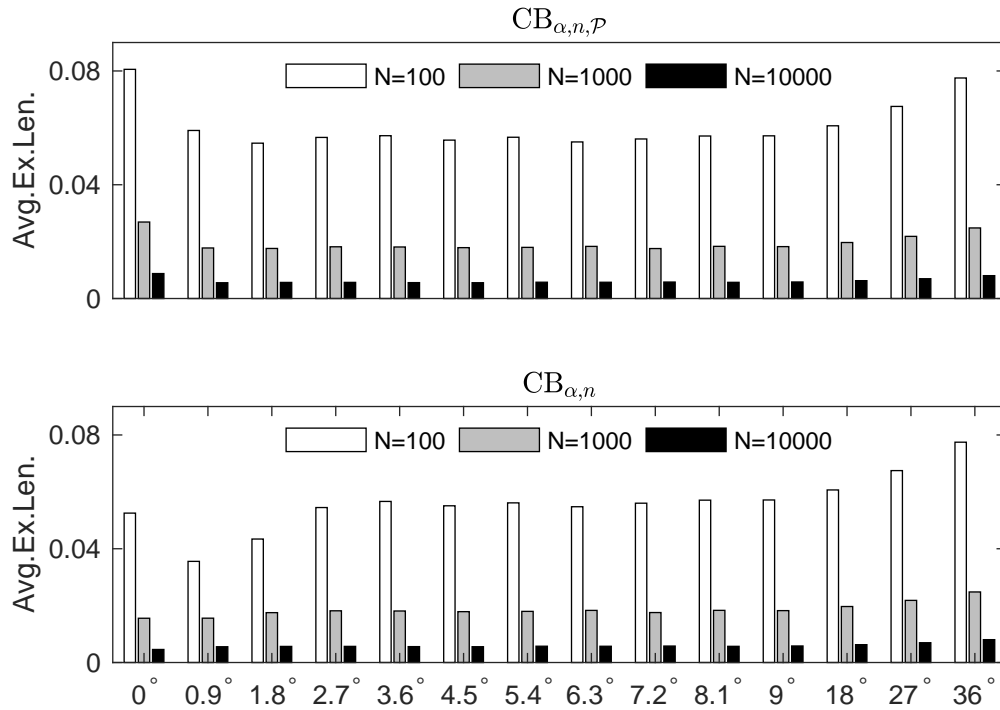
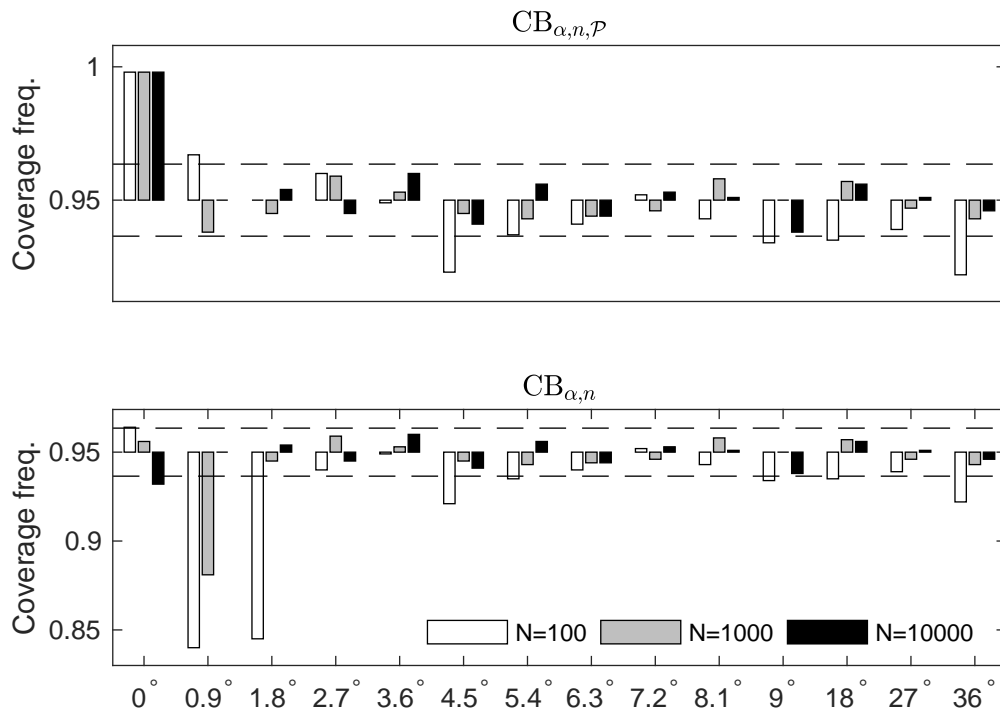
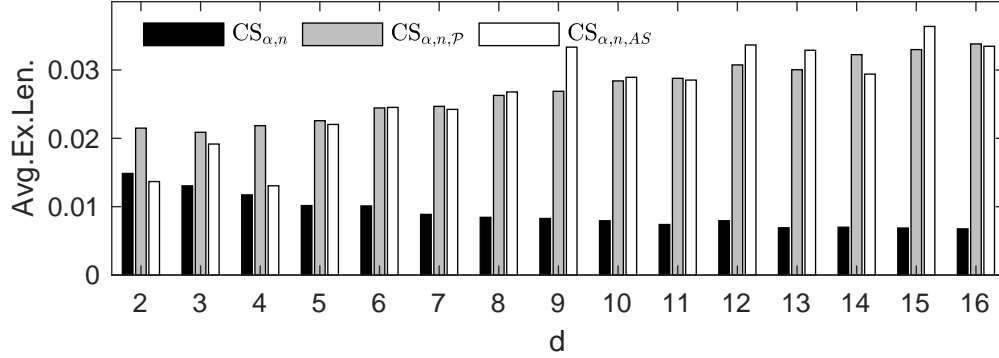


Figure 4: Coverage frequency in the two-dimensional design for $CB_{\alpha,N}$ and $CB_{\alpha,N,\mathcal{P}}$ for various sample sizes as a function of angle ω



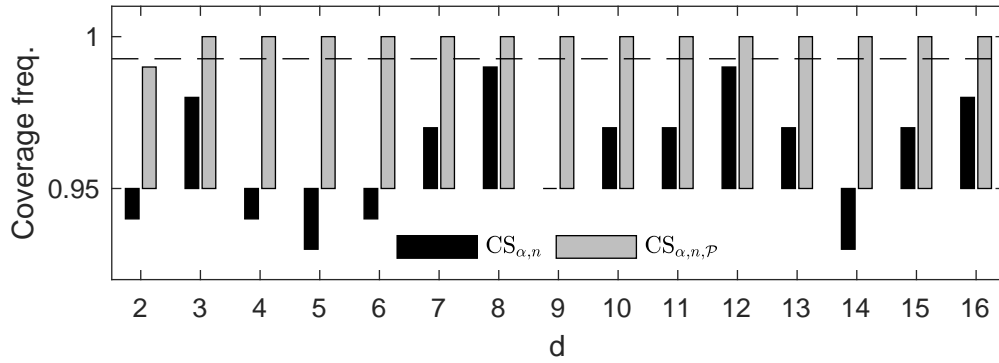
Note: the dashed line corresponds to the asymptotic 95% confidence interval for the parameter $p = 0.95$ of Bernoulli random variable based on a random sample of 1000 observations.

Figure 5: Average excess length in the d -dimensional design for $CB_{\alpha,N}, CB_{\alpha,N,\mathcal{P}}$ and $CB_{\alpha,N,AS}$ for sample size $N = 1000$ as a function of the dimension d .



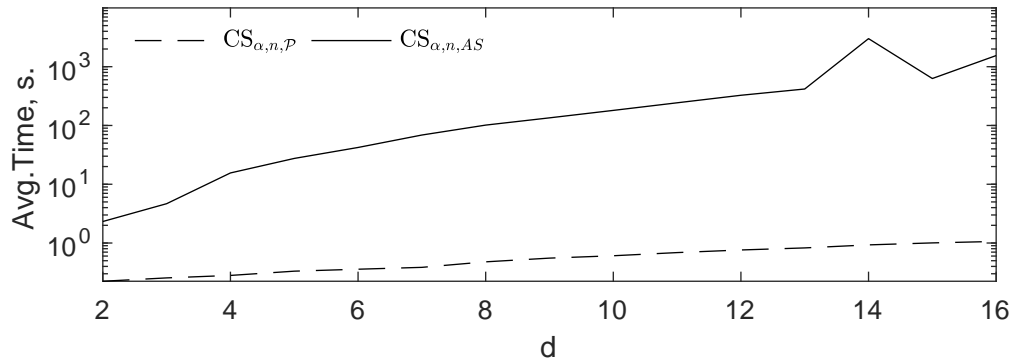
Note: MC sample size for $CI_{\alpha,n}^{AS}$ is 20, MC sample size for $\widetilde{CI}_{\alpha,n}$ is 100.

Figure 6: Coverage frequency in the d -dimensional design for $CB_{\alpha,N}, CB_{\alpha,N,\mathcal{P}}$ and $CB_{\alpha,N,AS}$ for sample size $N = 1000$ as a function of the dimension d .



Note: the dashed line corresponds to the asymptotic 95% confidence interval for the parameter $p = 0.95$ of Bernoulli random variable based on a random sample of 100 observations.

Figure 7: Average computation time in the d -dimensional design for $CB_{\alpha,N,\mathcal{P}}$ and $CB_{\alpha,N,AS}$ for sample size $N = 1000$ as a function of the dimension d .



Note: The solid line corresponds to $CI_{\alpha,n}^{AS}$, the dashed line corresponds to $\widetilde{CI}_{\alpha,n}$.

8 Proofs

8.1 Topological properties of optimal solutions

Consider any distribution P with support on $\mathbb{R}^{(k-2d) \times (d+1)}$ such that $(A_P, b_P) \triangleq \mathbb{E}_P W$ exist. Let $\mathcal{J}^a(\theta; P) \subset \{1, \dots, k\}$ be the set of indices of moment equality and inequality constraints active at θ , i.e. all j s.t. $m_j(\theta, P) \triangleq \mathbb{E}_P g_j(W, \theta) = 0$. $\mathcal{J}^a(\theta; P)$ can be empty.

Lemma 1 (Characterization of the optimal solution). *Under Assumption 1 for any $\mu \geq 0$ any minimizer θ for Program (19) is a solution to the corresponding Karush–Kuhn–Tucker (KKT) optimality conditions for some finite $\lambda \in \mathbb{R}^k$,*

$$\begin{cases} (e_1 + 2\mu\theta)' = -\lambda' A_P, & (47) \\ m_j(\theta, P) = 0 & j \in \mathcal{J}^{eq}, & (48) \\ m_j(\theta, P) \leq 0, \lambda_j \geq 0, \lambda_j m_j(\theta, P) = 0 & j \in \mathcal{J}^{ineq}. & (49) \end{cases}$$

Proof. By Assumption 1, $\Theta(P) \subset \Theta$ is non-empty and closed, so the global optima for Program (19) exist. Program (19) is convex for any $\mu \geq 0$, i.e. the objective function is convex, the constraints are affine. Assumption 1 implies Slater's condition. Since the Program (19) is convex, any global optimum $\underline{\theta}(\mu, P)$ of Program 19 satisfies (47)-(49) for some finite vector of Lagrange multipliers λ (maybe non-unique) (see p.244 in Boyd and Vandenberghe (2004)). \square

If we introduce notation $\underline{\mathcal{L}}(\lambda, \theta; \mu, P) \triangleq \theta_1 + \mu \|\theta\|^2 + \lambda' m(\theta, P)$, (47) becomes

$$\partial_{\theta} \underline{\mathcal{L}}(\lambda, \theta; \mu, P) = 0$$

Let $\underline{\xi}(\mu, P) \triangleq (\underline{\theta}(\mu, P), \underline{\lambda}(\mu, P))$ be a set of solutions to (47)-(49). In order to have a unique solution λ Program (19) need to meet a stronger constraint qualification condition, Assumption 6 defined below.

As before, I use symbols $\mathcal{J}^a(\theta; P)$, $\mathcal{J}^a(\mu, P)$ etc to denote the projectors on the coordinates with the corresponding indices. Let $\mathbb{J}_{d+1}^d \triangleq (e_1, \dots, e_d)$.

Assumption 6 (Linear Independence Constraint Qualification (LICQ)). *The matrix $\mathbb{J}^a(\theta; P) A_P$ has full row rank for any $\theta \in \Theta(P)$.*

Lemma 2 (Sufficient condition for LICQ). *Assumptions 2-3 imply Assumption 6.*

Proof. Assumption 3 implies that $\mathcal{J}^a(\theta; P)$ has at most d elements at any $\theta \in \Theta(P)$. Consider any point $\theta \in \Theta(P)$. The set \mathcal{J}^{ineq} includes (1), so $k \geq 2d \geq d+1$. It implies that there exists a set \mathcal{J} with $|\mathcal{J}| = d$ such that $\mathcal{J}^a(\theta; P) \subset \mathcal{J}$. By Assumption 2, $\text{rk}[\mathbb{J}\mathbb{E}_P W] = d$, so $M \triangleq \mathbb{J}^a(\theta; P)(A_P, b_P)$ has full row rank which is equal to $|\mathcal{J}^a|$. By definition, $\mathbb{J}^a(\theta; P) A_P \theta = \mathbb{J}^a b_P$. It implies by the Rouché–Capelli theorem that the matrices $M_{\theta} \triangleq \mathbb{J}^a(\theta; P) A_P$ and M have the same rank. This result implies Assumption 6. \square

The inverse implication does not hold in general as the following remark shows.

Remark 2. Assumption 6 implies Assumption 3 and that for any $\theta \in \Theta(P)$

$$\text{rk}[\mathbb{J}^a(\theta; P) \mathbb{E}_P W] = |\mathcal{J}^a(\theta; P)|. \quad (50)$$

Indeed, suppose that Assumption 6 holds. It immediately implies Assumption 3. To see (50) consider any point $\theta \in \Theta(P)$ such that $|\mathcal{J}^a(\theta; P)| \leq d$. By the Rouché–Capelli theorem and the full row rank property of M_{θ} correspondingly,

$$\text{rk}(M) = \text{rk}(M_{\theta}) = |\mathcal{J}^a(\theta; P)|.$$

Lemma 3 (Uniqueness of the optimal solutions). *Suppose that both Assumptions 1 and 6 are satisfied. Then for any $\mu \geq 0$ the set of multipliers $\underline{\lambda}(\mu, P)$ is a singleton. Moreover if $\mu > 0$, then $\underline{\xi}(\mu, P)$ is a singleton.*

Proof. By definition of $\mathbb{J}^a(\theta; P)$, any θ and λ satisfying (47) satisfy

$$\lambda'(\mathbb{J}^a(\theta; P))' \mathbb{J}^a(\theta; P) = \lambda'. \quad (51)$$

So (47) becomes

$$(e_1 + 2\mu\theta)' = -\gamma' \mathbb{J}^a(\theta; P) A_P, \quad (52)$$

where $\gamma' \triangleq \lambda'(\mathbb{J}^a(\theta; P))' \in \mathbb{R}^{|\mathcal{J}^a(\theta; P)|}$. By Assumption 6, for any $\theta \in \Theta(P)$ the matrix $A \triangleq \mathbb{J}^a(\theta; P) A_P$ has full rank. Hence for any θ there can be at most one $\gamma^* \in \mathbb{R}^{|\mathcal{J}^a(\theta; P)|}$ satisfying (52). If $e_1 + 2\mu\theta = 0$, then trivially λ is a zero vector. Otherwise it is given by

$$\gamma^* = -(AA')^{-1} A' (e_1 + 2\mu\theta). \quad (53)$$

Then $(\underline{\lambda}(\mu, P))' \triangleq (\gamma^*)' \mathbb{J}^a(\theta; P)$ is the unique solution to (47)-(49) for any solution θ .

Now consider the case $\mu > 0$. The second order derivative matrix of $\underline{\mathcal{L}}(\lambda, \theta; \mu, P)$ with respect to θ at any solution $\underline{\xi}(\mu, P)$ is $2\mu I_d$. It is positive definite for any $\mu > 0$, so the Second Order Sufficient Condition (SOSC) is satisfied at any point. By Theorem 3.63 from Bonnans and Shapiro (2000) the second order growth condition holds at $\underline{\theta}(\mu, P)$, i.e. $\exists \varepsilon > 0$ and $c > 0$ s.t. for $\forall \theta \in \Theta(P)$ s.t. $\|\theta - \underline{\theta}(\mu, P)\| < \varepsilon$ the following inequality holds

$$\theta_1 + \mu \|\theta\|^2 \geq e_1' \underline{\theta}(\mu, P) + \mu \|\underline{\theta}(\mu, P)\|^2 + c \|\theta - \underline{\theta}(\mu, P)\|^2.$$

So the value of the objective function at $\underline{\theta}(\mu, P)$ is strictly smaller than the value at any other point in a neighborhood of $\underline{\theta}(\mu, P)$. Since for the convex program the set of global optima is convex and connected, it implies that $\underline{\theta}(\mu, P)$ is the unique global minimizer. \square

Lemma 4. *Suppose that Assumptions 1-3 are satisfied and $\mu \leq 1/2$. Then*

$$\|\underline{\lambda}(\mu, P)\|^2 \leq C_\Lambda^2 \triangleq \frac{C_\Theta^3}{\eta^2} < \infty, \quad (54)$$

where $C_\Theta \triangleq (1 + \max_{\theta \in \Theta} \|\theta\|)$.

Proof. Consider any point $\theta \in \underline{\theta}(\mu, P)$ and the corresponding $\mathcal{J}^a(\theta; P)$. Let $A \triangleq \mathbb{J}^a(\theta; P) A_P$ and $b \triangleq \mathbb{J}^a(\theta; P) b_P$. Let $\eta_A^2 \triangleq \text{eig}(AA')$ so that equation (53) implies

$$\|\underline{\lambda}(\mu, P)\| \leq \eta_A^{-1} \|e_1 + 2\mu\theta\|. \quad (55)$$

By the variational property of eigenvalues,

$$\eta_A^2 = \min_{v \in \mathbb{R}^\ell} \frac{v' AA' v}{v' v}. \quad (56)$$

By Assumption 2

$$\eta^2 \leq \text{eig}((A, b)(A, b)') \triangleq \min_{v \in \mathbb{R}^\ell} \frac{v'(AA' + bb')v}{v' v}.$$

Let v_A be any minimizer of the r.h.s. of (56) such that $v_A' v_A = 1$. Then

$$\eta^2 \leq v_A'(AA' + bb')v_A = (A'v_A)'(I_d + \theta\theta')(A'v_A) \quad (57)$$

where the last equality holds since by definition $b = A\theta$. Finally,

$$\frac{\eta^2}{\eta_A^2} \leq \frac{(A'v_A)'(I_d + \theta\theta')(A'v_A)}{(A'v_A)'(A'v_A)} \leq \|I_d + \theta\theta'\| \leq C_\Theta. \quad (58)$$

Result (54) then follows from (55) and (58) for any $\mu \leq 1/2$. \square

Remark 3. Equation (58) provides bound for, A , a matrix with gradients of active moment conditions at any point $\theta \in \Theta$,

$$\|(AA')^{-1}\| \leq C_\Theta \eta^{-2}. \quad (59)$$

The function $\phi(a, b) \triangleq \sqrt{a^2 + b^2} + a - b$, considered in Fischer (1992), has the following property.

Proposition 1.

$$\phi(a, b) = 0 \text{ if and only if } a \leq 0, b \geq 0, ab = 0. \quad (60)$$

It can be used to replace (49) with an equivalent equality so that the KKT system becomes a system of equations. This result can be used to establish the continuity of the solutions in μ as the following lemma shows.

Lemma 5. *Under Assumptions 1- 3 $\underline{\xi}(\mu, P)$ is u.h.c. in μ ; $\underline{v}(\mu, P)$ is continuous in μ for $\mu \geq 0$.*

Proof. By Proposition 1 equation (49) is equivalent to

$$\phi(m_j(\theta, P), \lambda_j) = 0 \text{ for } j \in \mathcal{J}^{ineq}. \quad (61)$$

Solutions to (47),(48),(61) coincide with solutions to

$$\Psi(\theta, \lambda; \mu, P) \triangleq \|\partial_\theta \underline{\mathcal{L}}(\lambda, \theta; \mu, P)\|_2^2 + \sum_{j \in \mathcal{J}^{eq}} (m_j(\theta, P))^2 + \sum_{j \in \mathcal{J}^{ineq}} (\phi(m_j(\theta, P), \lambda_j))^2 = 0. \quad (62)$$

Lemmas 3-4 imply that $\underline{\lambda}(\mu, P)$ is unique and satisfies (54) for any $\mu \in [0, 1/2]$. So the solution to (62) coincides with solutions of

$$\begin{aligned} \min_{\theta, \lambda} \quad & \Psi(\theta, \lambda; \mu, P) \\ \text{s.t.} \quad & \theta \in \Theta, \lambda \in \mathbb{R}^k, \|\lambda\| \leq C_\Lambda. \end{aligned} \quad (63)$$

The objective function of this program is continuous in μ and the domain is a compact valued continuous correspondence in μ . By the Maximum Theorem (see Ok (2007)) $\underline{\xi}(\mu, P)$ is u.h.c. function of $\mu \geq 0$.

Function $\underline{v}(\mu, P) = e_1' \underline{\theta}(\mu, P) + \mu \|\underline{\theta}(\mu, P)\|^2$ is a composition of u.h.c. functions and hence, by Theorem VI.2.1' from Berge (1963), is u.h.c. in $\mu \in \mathbb{R}_+$. Since by definition $\underline{v}(\mu, P)$ is a single-valued function, u.h.c. implies continuity in $\mu \geq 0$ for any fixed P . \square

8.2 Smoothness properties

In this section we will study the directional derivatives of the value and the optimal solutions of Program (19). We will pursue this goal by taking a limit of the perturbed program defined below as the size of the perturbation goes to zero. Consider a perturbation in parameters $\mathbb{E}_P W = (A_P, b_P)$ and μ in a direction $h' = (\text{vec}(h_W)', h_\mu) \in \mathbb{R}^{k(d+1)+1}$, where $h_W \triangleq (h_A, h_b) \in \mathbb{R}^{k \times (d+1)}$ and $h_\mu \in \mathbb{R}$. The corresponding perturbation in the constraints is $\dot{m}_h(\theta) \triangleq h_A \theta - h_b$. Given these directions, for any $t \geq 0$, $\mu > 0$ we can define a perturbed program,

$$\min_{\theta \in \Theta} \quad e_1' \theta + (\mu + t h_\mu) \|\theta\|^2, \quad (64)$$

$$\text{s.t.} \quad \begin{cases} m_j(\theta, P) + t\dot{m}_{h;j}(\theta) = 0 & \text{for } j \in \mathcal{J}^{eq}, \\ m_j(\theta, P) + t\dot{m}_{h;j}(\theta) \leq 0 & \text{for } j \in \mathcal{J}^{ineq}. \end{cases}$$

Perturbations of inequality constrained programs can lead to changes in the set of the constraints active at the optimum in response to an arbitrarily small perturbations. It is instructive to consider the following sets of constraints for the unperturbed program, i.e. with $h_A = 0, h_b = 0, h_\mu = 0$,

$$\begin{aligned} \mathcal{J}^+(\mu, P) &\triangleq \{j \in \mathcal{J}^{ineq} | \underline{\lambda}_j(\mu, P) > 0\} \cup \mathcal{J}^{eq}, \\ \mathcal{J}^-(\mu, P) &\triangleq \{j \in \mathcal{J}^{ineq} | m_j(\underline{\theta}(\mu, P), P) > 0\}, \\ \mathcal{J}^0(\mu, P) &\triangleq \{j \in \mathcal{J}^{ineq} | \underline{\lambda}_j(\mu, P) = 0, m_j(\underline{\theta}(\mu, P), P) = 0\}, \\ \mathcal{J}^a(\mu, P) &\triangleq \mathcal{J}^0(\mu, P) \cup \mathcal{J}^+(\mu, P). \end{aligned}$$

Set \mathcal{J}^+ contains active inequality constraints with positive Lagrange multipliers and the equality constraints. These constraints will remain active for small enough perturbations in any directions h_A, h_b (by continuity of the Lagrange multipliers). Set \mathcal{J}^- contains slack constraints. They will remain slack in response to sufficiently small perturbations (by continuity of the optimal solution and the constraints functions). Set \mathcal{J}^0 contains active inequality constraints with zero Lagrange multipliers. If we drop these constraints the optimal solution will not change, but they play an important role in the perturbed program. Constraints in \mathcal{J}^0 become inactive in response to perturbations in some directions no matter how small the perturbation is or remain active and acquire positive Lagrange multipliers for other directions. The optimal solution ξ would be fully differentiable iff \mathcal{J}^0 is empty as will be evident from the explicit formula for its directional derivative. Finally, \mathcal{J}^a contains all active constraints at the optimal solution.

Suppose that the perturbation size $t > 0$ is small enough such that Program (64) satisfies Assumptions 1-3. Then it has a unique solution $\underline{\xi}_h(t)$ for $0 \leq t < T$ which can be represented as $\underline{\xi}_h(t) = \underline{\xi} + t\underline{\dot{\xi}}_h$, as the following lemma shows. The directional derivative $\underline{\dot{\xi}}_h'(\mu, P) \triangleq (\underline{\dot{\theta}}'(\mu, P), \underline{\dot{\lambda}}'(\mu, P))$ will depend on the following objects,

$$\begin{aligned} \mathcal{J}^h(\mu, P) &\triangleq \{j \in \mathcal{J}^0(\mu, P) | \dot{\lambda}_{h;j}(\mu, P) > 0\} \cup \mathcal{J}^+(\mu, P), \\ A_h(\mu, P) &\triangleq \mathbb{J}^h(\mu, P) A_P, \\ Q_h &\triangleq I_d - A_h'(A_h A_h')^{-1} A_h, \\ A^\dagger &\triangleq A'(A A')^{-1}. \end{aligned}$$

I suppress the argument (μ, P) from now on.

Lemma 6 (Local linear representation). *Suppose that Assumptions 1 -3 hold for P . There is a neighborhood $[0, T(\mu, h, P)]$ in which Program (64) has a unique solution $\underline{\xi}_h(t) = \underline{\xi} + t\underline{\dot{\xi}}_h$ with*

$$\underline{\dot{\xi}}_h = - \begin{pmatrix} (2\mu)^{-1} Q_h & A_h^\dagger \\ (\mathbb{J}^h)'(A_h^\dagger)' & -2\mu(\mathbb{J}^h)'(A_h A_h')^{-1} \end{pmatrix} \begin{pmatrix} (h_A)'\underline{\lambda} + 2h_\mu\underline{\theta} \\ \mathbb{J}^h(h_A\underline{\theta} - h_b) \end{pmatrix}. \quad (65)$$

Proof. By Lemma 3, if $t = 0$ Program (64) has a unique solution $\underline{\xi}$. Since this solution satisfies LICQ (Assumption 6) and SOSC, it is strongly regular by Proposition 5.38 from BS(2000). The remaining argument follows the proof of Theorem 5.60 from BS(2000), which uses an implicit function theorem for generalized equations (Theorem 5.13 in the same book) at a strongly regular solution. We are going to

apply it to the KKT conditions for Program (64) at the strongly regular solution $\underline{\xi}$,

$$\begin{cases} (e_1 + 2(\mu + h_\mu t)\theta)' = -\lambda'(A_P + th_A), & (66) \\ m_j(\theta, P) + t\dot{m}_{h;j}(\theta) = 0 & j \in \mathcal{J}^{eq}, & (67) \\ \phi(m_j(\theta, P) + t\dot{m}_{h;j}(\theta), \lambda_j) = 0 & j \in \mathcal{J}^{ineq}. & (68) \end{cases}$$

By Theorem 5.60 from BS(2000), $\underline{\xi}_h(t)$ is analytic in t in some neighborhood $[0, T(\mu, h, P)]$, i.e. it can be represented as power series. First, let us compute the linear term. By the strong regularity and Theorem 5.13 in BS(2000), there exist a unique solution $(\underline{\theta}, \underline{\lambda})$ to the following system of equations (this system is the gradient of (66)-(68) with respect to t at point $t = 0$)

$$\begin{cases} 2\mu\underline{\theta}'I_d + \underline{\lambda}'A_P = -\underline{\lambda}'h_A - 2h_\mu\underline{\theta}', & (69) \\ e_j'A_P\underline{\theta} + \dot{m}_{h;j}(\underline{\theta}) = 0 & j \in \mathcal{J}^+(\mu, P), & (70) \\ \phi(e_j'A_P\underline{\theta} + \dot{m}_{h;j}(\underline{\theta}), \underline{\lambda}_j) = 0 & j \in \mathcal{J}^0(\mu, P), & (71) \\ \dot{\lambda}_{h;j} = 0 & j \in \mathcal{J}^-(\mu, P). & (72) \end{cases}$$

This unique solution determines the set \mathcal{J}^h . System (69)-(72) can be represented in a matrix form:³¹

$$\begin{pmatrix} 2\mu I_d & A_h' \\ A_h & 0 \end{pmatrix} \begin{pmatrix} \underline{\theta} \\ \mathbb{J}^h \underline{\lambda} \end{pmatrix} = - \begin{pmatrix} (h_A)'\underline{\lambda} + 2h_\mu\underline{\theta} \\ \mathbb{J}^h(h_A\underline{\theta} - h_b) \end{pmatrix}. \quad (73)$$

In addition to that, $\underline{\lambda} = (\mathbb{J}^h)'\mathbb{J}^h \underline{\lambda}$. One can check by direct computation that

$$\begin{pmatrix} 2\mu I_d & A_h' \\ A_h & 0 \end{pmatrix}^{-1} = \begin{pmatrix} (2\mu)^{-1}Q_h & A_h^\dagger \\ (A_h^\dagger)' & -2\mu(A_h A_h')^{-1} \end{pmatrix}.$$

Since the higher order derivatives of every constraint function and the objective function of Program (64) with respect to t are zero, the higher order directional derivatives of $\underline{\xi}_h$ are equal to zero at $t = 0$. Thus the power series expansion of $\underline{\xi}_h$ has only constant and linear terms. \square

Now we can rewrite Program (64) in an explicit form assuming $h_\mu = 0$,

$$\begin{aligned} \min_{\theta \in \Theta} \quad & e_1'\theta + \mu \|\theta\|^2, & (74) \\ \text{s.t.} \quad & \begin{cases} e_j'(A_P + th_A)\theta = b_P + th_b & \text{for } j \in \mathcal{J}^{eq}, \\ e_j'(A_P + th_A)\theta \leq b_P + th_b & \text{for } j \in \mathcal{J}^{ineq}. \end{cases} \end{aligned}$$

Lemma 7. *Suppose that Assumptions 1-3 hold for P . There exist such $\delta > 0$ such that for any $h = (h_A, h_b) \in \mathbb{R}^{k \times (d+1)}$ with norm $\|h\| < \delta$ and any $\mu \in (0, 1/2]$ and $t \in [0, 1]$ the solution of Program (74) satisfies*

$$\|\underline{\theta}_h(t) - \underline{\theta}\| \leq L_\theta \frac{\|th\|}{\mu} \quad (75)$$

$$\|\underline{\lambda}_h(t) - \underline{\lambda}\| \leq L_\lambda \|th\| \quad (76)$$

$$\|\underline{v}_h(t) - \underline{v} - t\underline{\lambda}'(h_A\underline{\theta} - h_b)\| \leq L_v \frac{\|th\|^2}{\mu}, \quad (77)$$

³¹Compare Equation (73) with Equation (5.81) on page 186 in Shapiro et al. (2014).

where $L_\theta = \frac{\sqrt{2C_\Theta^3}}{(\underline{\eta}/2)}$, $L_\lambda = \frac{(\|\mathbb{E}_P W\| + \delta)C_\Theta^4 + \eta^2 C_\Theta^2}{(\underline{\eta}/2)^4}$, and $L_v = \frac{2C_\Theta^3}{(\underline{\eta}/2)^2}$.

Proof. First consider $t \in [0, T(\mu, h, P)]$ with $T(\mu, h, P)$ defined in Lemma 6. By Lemma 6 the value function $\underline{v}_h(t) \triangleq e_1' \underline{\theta}_h(t) + \mu \|\underline{\theta}_h(t)\|^2$ can be represented as

$$\underline{v}_h(t) = \underline{v} + t(e_1 + 2\mu\underline{\theta})' \dot{\underline{\theta}} + \mu t^2 \|\dot{\underline{\theta}}\|^2. \quad (78)$$

First, consider the second term. Since by definition $\mathcal{J}^+ \subseteq \mathcal{J}^h$, we have $\underline{\lambda}' = \underline{\lambda}'(\mathbb{J}^h)' \mathbb{J}^h$. Correspondingly, $\underline{\lambda}' A_P = \underline{\lambda}'(\mathbb{J}^h)' A_h$. By Lemma 3, $(e_1 + 2\mu\underline{\theta})' = -\underline{\lambda}' A_P$. So

$$(e_1 + 2\mu\underline{\theta})' Q_h = -\underline{\lambda}'(\mathbb{J}^h)'(A_h Q_h) = 0, \quad (79)$$

$$(e_1 + 2\mu\underline{\theta})' A_h^\dagger = -\underline{\lambda}'(\mathbb{J}^h)'(A_h A_h^\dagger) = -\underline{\lambda}'(\mathbb{J}^h)'. \quad (80)$$

Equations (79) and (80) imply that

$$(e_1 + 2\mu\underline{\theta})' \dot{\underline{\theta}} = -\underline{\lambda}' \dot{m}_h(\underline{\theta}). \quad (81)$$

Second, by Lemma 4 and Remark 3 for any $\mu \leq 1/2$,

$$\|(A_h A_h')^{-1}\| \leq C_\Theta \eta^{-2}(P) \text{ and } \|\underline{\lambda}\|^2 \leq C_\Theta^3 \eta^{-2}(P). \quad (82)$$

Then by the triangular inequality and inequalities (82) (and the fact that $C_\Theta \geq 1$)

$$\|\dot{\underline{\theta}}\|^2 = \frac{1}{(2\mu)^2} (\underline{\lambda}' h_A) Q_h (\underline{\lambda}' h_A)' + \dot{m}_h(\underline{\theta})' (\mathbb{J}^h)' (A_h A_h')^{-1} \mathbb{J}^h \dot{m}_h(\underline{\theta}) \quad (83)$$

$$\leq \|h_W\|^2 \frac{C_\Theta^3}{\eta^2(P)} \left(\frac{1}{\mu^2} + 1 \right), \quad (84)$$

which implies (75). Equation (76) can be proven similarly,

$$\|\dot{\underline{\lambda}}\| = \|(\mathbb{J}^h)'(A_h^\dagger)'(\underline{\lambda}' h_A) - 2\mu(\mathbb{J}^h)'(A_h A_h')^{-1} \mathbb{J}^h \dot{m}_h(\underline{\theta})\| \quad (85)$$

$$\leq \frac{\|\mathbb{E}_P W\| C_\Theta^4 + \eta^2(P) C_\Theta^2}{\eta^4(P)} \|h_W\| \quad (86)$$

Finally, the bound in (77) follows from equations (75), (78), and (81).

To extend the argument to the entire interval $t \in [0, 1]$, notice that $\eta(P)$ and $s(P)$ are Lipschitz-continuous. So there exist $\delta > 0$ such that $\eta(\mathbb{E}_P W + th) > \underline{\eta}/2$ and $s(\mathbb{E}_P W + th) > \underline{s}/2$ for any $h = (h_A, h_b) \in \mathbb{R}^{k \times (d+1)}$ with norm $\|h\| < \delta$ and any $t \in [0, 1]$. Such δ can be chosen uniformly for $P \in \mathcal{P}$. Now we can replace $\|\mathbb{E}_P W\|$ with $(\|\mathbb{E}_P W\| + \delta)$ and $\eta(P)$ with $\underline{\eta}/2$ to obtain uniform constants L_θ , L_λ , and L_v . □

8.3 Proof of Theorem 1

Lemma 8. *Suppose that Assumptions 1-3 hold. There exist some $\bar{\mu}(P) > 0$ such that for any $\mu < \bar{\mu}(P)$ the solution to Program (19), $\underline{\theta}(\mu, P)$ is constant.*

Proof. Consider a direction $h' = (\text{vec}(h_W)', h_\mu) \in \mathbb{R}^{k(d+1)+1}$ satisfying $h_W = 0$, $h_\mu = 1$ and any $\mu_0 > 0$ in

a neighborhood of 0. By Lemma 6 we get

$$\dot{\underline{\theta}}(\mu_0, P) = -\frac{1}{\mu_0} Q_h(\mu_0, P) \underline{\theta}(\mu_0, P). \quad (87)$$

Substitute difference in eq.(47) (Lemma 1) between $\mu = 0$ and μ_0 for $\underline{\theta}$ in (87),

$$\dot{\underline{\theta}}(\mu_0, P) = (2\mu_0^2)^{-1} Q_h(\mu_0, P) A'_P (\underline{\lambda}(\mu_0, P) - \underline{\lambda}(0, P)). \quad (88)$$

Now we need to show that $\dot{\underline{\theta}}(\mu_0, P) = 0$. To establish this result, we need to study the behavior of the set of inequalities with positive Lagrange multipliers.

Consider any $j \in \mathcal{J}^{ineq}$. We know by Lemma 5 that $\underline{\lambda}_j(\mu, P)$ is continuous in μ . If $\underline{\lambda}_j(0, P) > 0$ then by continuity $\underline{\lambda}_j(\mu, P) > 0$ in some neighborhood $(0, \bar{\mu}_j(P)]$. If $\underline{\lambda}_j(0, P) = 0$ set $\bar{\mu}_j = 1$. Take $\bar{\mu}(P) \triangleq \min_{j \in \mathcal{J}^{ineq}} \bar{\mu}_j(P)$. WLOG suppose that $\mu_0 \in [0, \bar{\mu}(P)]$, so we get the inclusion

$$\mathcal{J}^+(0, P) \subseteq \mathcal{J}^+(\mu_0, P). \quad (89)$$

By definition of \mathcal{J}^h

$$\mathcal{J}^+(\mu_0, P) \subseteq \mathcal{J}^h(\mu_0, P). \quad (90)$$

By definition of the index matrices, inclusions (89) and (90) imply that

$$\underline{\lambda}(0, P) = (\mathbb{J}^h(\mu_0, P))' \mathbb{J}^h(\mu_0, P) \underline{\lambda}(0, P), \quad (91)$$

$$\underline{\lambda}(\mu_0, P) = (\mathbb{J}^h(\mu_0, P))' \mathbb{J}^h(\mu_0, P) \underline{\lambda}(\mu_0, P), \quad (92)$$

so

$$A'_P (\underline{\lambda}(\mu_0, P) - \underline{\lambda}(0, P)) = A'_h(\mu_0, P) (\mathbb{J}^h(\mu_0, P)) (\underline{\lambda}(\mu_0, P) - \underline{\lambda}(0, P)). \quad (93)$$

Since by definition $Q_h(\mu_0, P) A_h(\mu_0, P) = 0$ and Q_h is a symmetric matrix, equation (88) implies $\dot{\underline{\theta}}(\mu_0, P) = 0$. By Lemma 5 the single valued function $\underline{\theta}(\mu, P)$ is continuous for $\mu > 0$. So the r.h.s. directional derivative being equal to zero implies that $\underline{\theta}(\mu_0, P) = \underline{\theta}(\bar{\mu}(P), P)$ for any $\mu_0 \in (0, \bar{\mu}(P)]$. \square

Remark 4. Equation (47) from Lemma 1 with $\mu = 0$ and $\mu = \mu_0$ also implies

$$\underline{\lambda}(\mu_0, P) = \underline{\lambda}(0, P) - 2\mu_0 \underline{\theta}'(\bar{\mu}(P), P) A_h^\dagger(\mu_0, P) \mathbb{J}^h(\mu_0, P).$$

This implies that $\underline{\lambda}(\mu, P)$ is Lipschitz at $\mu = 0$. By Lemma 4 the Lipschitz constant can be taken equal to $2C_\Lambda$. So for any $j \in \mathcal{J}^{ineq}$ with $\underline{\lambda}_j(0, P) > 0$, we can take $\bar{\mu}_j(P) = \underline{\lambda}_j(0, P) / 2C_\Lambda$. On a top of that, Lemma 6 implies that $\underline{\lambda}(\mu, P)$ is Lipschitz in μ with the same constant for any $\mu \in [0, 1/2]$

Proof of Theorem 1. Take any $\theta^* \in \underline{\theta}(0, P)$. Since θ^* is a feasible point of Program (19),

$$\underline{\theta}_1(\mu, P) + \mu \|\underline{\theta}(\mu, P)\|^2 \leq \theta_1^* + \mu \|\theta^*\|^2. \quad (94)$$

By definition $\underline{v}(P) = \theta_1^*$ which immediately implies the l.h.s. inequality in (22).

The first coordinate $\underline{\theta}_1(\mu, P)$ is increasing in μ while the norm $\|\underline{\theta}(\mu, P)\|^2$ is decreasing in μ . Since $\kappa \geq \mu \geq 0$, we get

$$\|\underline{\theta}(\mu, P)\|^2 \geq \|\underline{\theta}(\kappa, P)\|^2, \quad (95)$$

$$\underline{v}(P) \leq \underline{\theta}_1(\mu, P). \quad (96)$$

These two inequalities imply the r.h.s. inequality in (22),

$$\underline{v}(P) \leq \underline{\theta}_1(\mu, P) + \mu (\|\underline{\theta}(\mu, P)\|^2 - \|\underline{\theta}(\kappa, P)\|^2) = \underline{v}^{in}(\mu, \kappa, P). \quad (97)$$

The remaining part of the Theorem's assertion follows from Lemma 8. \square

8.4 Proof of Theorem 2

Proposition 2. *Suppose that $\mathbb{E}_P |\xi|^{1+\epsilon} \leq \infty$ for some $\epsilon > 0$. Then for any $r > 0$*

$$\mathbb{E}_P[|\xi| I\{|\xi| \geq r\}] \leq \mathbb{E}_P |\xi|^{1+\epsilon} / r^\epsilon.$$

Proof. The result follows from the monotonicity of integrals. \square

Let $\pi(P, Q)$ denote the Prohorov distance between probability laws on P and Q , which induces the weak topology (see p. 456 in van der Vaart and Wellner (1996)). Let

$$G_n(P) \triangleq \sqrt{n} \left(\text{vec} \left(\frac{1}{n} \sum_{i=1}^n w_i \right) - \text{vec}(\mathbb{E}_P W) \right).$$

Lemma 9. *Consider \mathcal{P} , a class of distributions satisfying Assumptions 4-5, and any $\epsilon > 0$. Then*

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left(\sup_{m \geq n} \left\| \frac{1}{m} \sum_{i=1}^m w_i - \mathbb{E}_P W \right\| \geq \epsilon \right) = 0, \quad (98)$$

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left(\sup_{m \geq n} \left\| \frac{1}{m} \sum_{i=1}^m w_i \otimes w_i - \mathbb{E}_P[W \otimes W] \right\| \geq \epsilon \right) = 0, \quad (99)$$

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \pi(G_n(P), N(0, \Omega_P)) = 0, \quad (100)$$

where $\Omega_P = \text{Cov}_P(\text{vec}(W))$.

Proof. Consider any combination of indices r, ℓ, j, m . Assumption 5 together with the Schwarz inequality implies,

$$\mathbb{E}_P |W_{r,\ell} W_{j,m}|^{1+\epsilon/2} \leq \left(\mathbb{E}_P |W_{r,\ell}|^{2+\epsilon} \mathbb{E}_P |W_{j,m}|^{2+\epsilon} \right)^{1/2} \leq \bar{M}. \quad (101)$$

So the random variables $|W_{r,\ell}|$ and $|W_{r,\ell} W_{j,m}|$ have correspondingly finite $1 + \epsilon/2$ and $2 + \epsilon$ moments. The bound (101) on the moments is independent of $P \in \mathcal{P}$, so these random variables are uniformly integrable on \mathcal{P} by Proposition 2. The limits (98) and (99) follow immediately from Proposition A.5.1 in van der Vaart and Wellner (1996). The result (100) follows from Proposition A.5.2 in the same book. \square

Lemma 10. *Suppose that $P \in \mathcal{P}$. Then \mathbb{P}_n satisfies Assumptions 1 and 6 with probability approaching 1 uniformly in $P \in \mathcal{P}$.*

Proof. Consider any $P \in \mathcal{P}$. Since such a P satisfies Assumption 1, there exists a set of constraints \mathcal{J} with $|\mathcal{J}| = d$ and containing \mathcal{J}^{eq} such that

$$\theta_P^{\mathcal{J}} \triangleq ((A_P^{\mathcal{J}})' A_P^{\mathcal{J}})^{-1} (A_P^{\mathcal{J}})' b_P^{\mathcal{J}} \in \Theta(P), \quad (102)$$

where $A_P^{\mathcal{J}} \triangleq \mathbb{J} A_P$ and $b_P^{\mathcal{J}} \triangleq \mathbb{J} b_P$. By Assumption 3 for all $j \in \mathcal{J}^{ineq} \setminus \mathcal{J}$ we have

$$e_j (A_P \theta_P^{\mathcal{J}} - b_P) \geq \underline{\epsilon}. \quad (103)$$

The function $(A_P, b_P) = \mathbb{E}_P W$ is uniformly continuous on \mathcal{P} since this class of measures is uniformly integrable, as was shown in the proof of Lemma 9. The function $\theta_P^{\mathcal{J}}$ is uniformly continuous on \mathcal{P} since

the product matrix $(A_P^{\mathcal{J}})'A_P^{\mathcal{J}}$ has eigenvalues uniformly bounded from below by η^2 . By Lemma 9 and the continuous mapping theorem,

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left\{ \inf_{m \geq n; j \in \mathcal{J}^{ineq} \setminus \mathcal{J}} \{e_j(\hat{A}_m \hat{\theta}_m^{\mathcal{J}} - \hat{b}_m)\} < \underline{s}/2 \right\} = 0, \quad (104)$$

where \hat{A}_m and \hat{b}_m are sample analog estimators of A_P and b_P based on a sample of size m ; $\hat{\theta}_m^{\mathcal{J}} \triangleq ((\hat{A}_m^{\mathcal{J}})' \hat{A}_m^{\mathcal{J}})^{-1} (\hat{A}_m^{\mathcal{J}})' \hat{b}_m^{\mathcal{J}}$. This result implies that $\Theta(\mathbb{P}_n)$ contains at least one element, $\hat{\theta}_n^{\mathcal{J}}$, with probability approaching 1 uniformly in $P \in \mathcal{P}$ as $n \rightarrow \infty$.

Analogously, the continuous mapping theorem implies

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left\{ \inf_{m \geq n} \eta(\mathbb{P}_m) < \underline{\eta}/2, \inf_{m \geq n} s(\mathbb{P}_m) < \underline{s}/2 \right\} = 0. \quad (105)$$

Result follows from Lemma 2. □

Lemma 11. *Suppose that $P \in \mathcal{P}$ and that $\mu_n \rightarrow 0$. Then*

$$\underline{v}(\mu_n, \mathbb{P}_n) = \underline{v}(\mu_n, P) + \frac{1}{n} \sum_{i=1}^n \underline{\lambda}(\mu_n, P)' g(w_i, \underline{\theta}(\mu_n, P)) + O_{\mathcal{P}}\left(\frac{1}{\mu_n n}\right). \quad (106)$$

Proof. First note that by Lemma 10 the random variables $\underline{\theta}(\mu_n, \mathbb{P}_n)$ and $\underline{v}(\mu_n, \mathbb{P}_n)$ are well defined with probability approaching 1 uniformly in $P \in \mathcal{P}$. Consider $t = 1/\sqrt{n}$ and h such that $h_W = \sqrt{n}(\frac{1}{n} \sum_{i=1}^n w_i - \mathbb{E}_P W)$ and $h_{\mu} = 0$. The sequence of perturbations satisfies $\sqrt{n}(\frac{1}{n} \sum_{i=1}^n w_i - \mathbb{E}_P W) = O_{\mathcal{P}}(1)$ (i.e. it has uniformly tight measures) by (100) in Lemma 9 and the fact that $\|\Omega_P\| \leq \bar{M}$ (by Jensen's inequality). The assertion of the Lemma follows from equation (77) in Lemma 7. Indeed, by Lemma 9

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left(\sup_{m \geq n} \left\| \frac{1}{m} \sum_{i=1}^m w_i - \mathbb{E}_P W \right\| \geq \delta \right) = 0, \quad (107)$$

so the perturbation is small enough to preserve the results of Lemma 7, i.e. $\|th_W\| < \delta$, with probability approaching 1 uniformly as sample size grows. □

Let's introduce the following definitions

$$\begin{aligned} \Sigma(\theta) &\triangleq \mathbb{E}_P [g(W, \theta)g(W, \theta)'] - m(\theta, P)m(\theta, P)', \\ \hat{\Sigma}_n(\theta) &\triangleq \frac{1}{n} \sum_{i=1}^n g(w_i, \theta)g(w_i, \theta)' - \frac{1}{n} \sum_{i=1}^n g(w_i, \theta) \frac{1}{n} \sum_{i=1}^n g(w_i, \theta)'. \end{aligned}$$

Proof of Theorem 2. First note that Lemma 10 the random variables $\underline{\theta}(\mu_n, \mathbb{P}_n)$ and $\underline{\lambda}(\mu_n, \mathbb{P}_n)$ are well defined with probability approaching 1 uniformly in $P \in \mathcal{P}$. Consider t and h as in the proof of Lemma 11. Equation (26) follows from Lemma 9, Lemma 11, and Slutsky's theorem. The results (27) and (28) follow from Lemma 7. Finally, by the triangular inequality

$$\begin{aligned} \left| \underline{\sigma}(\mu, \mathbb{P}_n)^2 - \underline{\sigma}(\mu, P)^2 \right| &= \left| \underline{\lambda}(\mu, \mathbb{P}_n)' \hat{\Sigma}_n(\underline{\theta}(\mu, \mathbb{P}_n)) \underline{\lambda}(\mu, \mathbb{P}_n) - \underline{\lambda}(\mu, P)' \Sigma(\underline{\theta}(\mu, P)) \underline{\lambda}(\mu, P) \right| \leq \\ &\left| \underline{\lambda}(\mu, \mathbb{P}_n)' \hat{\Sigma}_n(\underline{\theta}(\mu, \mathbb{P}_n)) \underline{\lambda}(\mu, \mathbb{P}_n) - \underline{\lambda}(\mu, P)' \hat{\Sigma}_n(\underline{\theta}(\mu, P)) \underline{\lambda}(\mu, P) \right| \\ &+ \left| \underline{\lambda}(\mu, P)' \hat{\Sigma}_n(\underline{\theta}(\mu, P)) \underline{\lambda}(\mu, P) - \underline{\lambda}(\mu, P)' \Sigma(\underline{\theta}(\mu, P)) \underline{\lambda}(\mu, P) \right|. \quad (108) \end{aligned}$$

Together with (27), (28) and Lemmas 4 and 9 it implies (29).

□

8.5 Proof of Theorem 3

Proof of Theorem 3. STEP 1. Consider

$$\zeta_n \triangleq \frac{\sqrt{n} \underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n) - \underline{v}^{in}(\mu_n, \kappa_n, P) + \underline{v}^{in}(\mu_n, \kappa_n, P) - \underline{v}(P)}{\underline{\sigma}(\mu, \mathbb{P}_n)} \quad (109)$$

Since $\mu_n/\kappa_n \rightarrow 0$ and $\mu_n \rightarrow 0$, for all n large enough, such that $\mu_n \leq \kappa_n \leq \bar{\mu}(P)$, by Theorem 1 we get

$$\underline{v}^{in}(\mu_n, \kappa_n, P) = \underline{v}(P). \quad (110)$$

By Theorem 2 we get

$$\mu_n \sqrt{n} \left(\|\underline{\theta}(\kappa_n, \mathbb{P}_n)\|^2 - \|\underline{\theta}(\kappa_n, P)\|^2 \right) = \frac{\mu_n}{\kappa_n} O_p(1) = o_p(1). \quad (111)$$

By Lemma 5, $\underline{\theta}(\mu, P)$ and $\underline{\lambda}(\mu, P)$ are continuous for $\mu > 0$. The matrix function $\underline{\Sigma}(\theta, P)$ is continuous in θ and thus $\underline{\sigma}(\mu, P)$ is continuous in μ for $\mu > 0$. So the limit $\lim_{n \rightarrow \infty} \underline{\sigma}^2(\mu_n, P)$ exists and belongs to the set $\underline{\sigma}^2(0, P)$ which by assumptions implies $\lim_{n \rightarrow \infty} \underline{\sigma}^2(\mu_n, P) > 0$. Result (26) in Theorem 2 together with (110) and (111) imply by Slutsky's theorem that ζ_n converges in distribution to $N(0, 1)$.

STEP 2. Consider the one-sided confidence band $\text{CB}_{\alpha, n}$.

$$\begin{aligned} & \lim_{n \rightarrow \infty} P \{ \mathcal{S}(P) \subset \text{CB}_{\alpha, n} \} \\ &= \lim_{n \rightarrow \infty} P \left\{ \underline{v}(P) \geq \underline{v}^{in}(\mu_n, \kappa_n, \mathbb{P}_n) - \underline{\sigma}(\mu_n, \mathbb{P}_n) z_{1-\alpha} n^{-1/2} \right\} \\ &= \lim_{n \rightarrow \infty} P \left\{ \zeta_n \leq z_{1-\alpha} \right\} \\ &= \Phi(z_{1-\alpha}) = 1 - \alpha. \end{aligned}$$

Proof for $\text{CI}_{\alpha, n}^S$ follows immediately from the Bonferroni inequality. Finally, consider the case $p = 0$. Then by Lemma 2 and Assumption 1, $\underline{v}(0, P) < -\bar{v}(0, P)$. So

$$\begin{aligned} & \lim_{n \rightarrow \infty} \min_{\theta \in \Theta(P)} P \left(\theta \in \text{CI}_{\alpha, n}^{\theta_1} \right) = \\ & \min \left\{ \lim_{n \rightarrow \infty} P \left\{ \underline{v}(\mu_n, \mathbb{P}_n) - \mu_n \|\underline{\theta}(\kappa_n, \mathbb{P}_n)\|^2 - \underline{\sigma}(\mu_n, \mathbb{P}_n) z_{1-\alpha} n^{-1/2} \leq \underline{v}(P) \right\}, 1, \dots \right. \\ & \quad \left. \lim_{n \rightarrow \infty} P \left\{ -\bar{v}(\mu_n; \mathbb{P}_n) + \mu_n \|\bar{\theta}(\kappa_n, \mathbb{P}_n)\|^2 + z_{1-\alpha} \bar{\sigma}(\mu_n; \mathbb{P}_n) n^{-1/2} \geq \bar{v}(P) \right\} \right\} = \\ & \min \left\{ \lim_{n \rightarrow \infty} P \{ \mathcal{S}(P) \subset \text{CB}_{\alpha, n} \}, 1, \lim_{n \rightarrow \infty} P \{ \mathcal{S}(P) \subset \text{CB}_{\alpha, n}^R \} \right\} = \quad (112) \\ & \min \{ 1 - \alpha, 1, 1 - \alpha \} = 1 - \alpha. \quad (113) \end{aligned}$$

To understand the second equation, consider the following argument. Suppose that $\theta \in \Theta(P)$ is such that $\underline{v}(P) < \theta_1 < \bar{v}(P)$. Then such θ_1 will be covered with probability 1 since $\text{CI}_{\alpha, n}^{\theta}$ is the intersection of $\text{CB}_{\alpha, n}$ and $\text{CB}_{\alpha, n}^R$ which cover correspondingly $\underline{v}(P)$ and $\bar{v}(P)$. □

8.6 Proof of Theorem 4

Lemma 12. *Suppose that A_4 holds then for any $\epsilon > 0$ there exists $R \geq 0$ such that for any n the following uniform bound holds*

$$\sup_{P \in \mathcal{P}} P \left(\sqrt{n} \|\theta^*(\mathbb{P}_n) - \theta^*(P)\| \geq R \right) \leq \epsilon. \quad (114)$$

Proof. The proof is based on the delta method applied to $\theta^*(P) = \theta^*(A_P, b_P)$, a composition of directionally differentiable functions.³²

First note that $\underline{v}(P) = \underline{v}(A_P, b_P)$ is directionally differentiable function. To see it, consider the minimax representation, which is valid since P satisfies Assumption 1,

$$\underline{v}(A_P, b_P) = \min_{\theta \in \Theta} \max_{\lambda \in \mathbb{R}^p \times \mathbb{R}_+^{k-p}, \|\lambda\| \leq C_\Lambda} \{\theta_1 + \lambda'(A_P \theta - b_P)\}. \quad (115)$$

Here we also used Lemma 4 to bound λ and make the domain compact. By Theorem 7.28 from Shapiro et al. (2014) for any direction $(h_A, h_b) \in \mathbb{R}^{k \times (d+1)}$ we have

$$\dot{\underline{v}}(P|h_A, h_b) \triangleq \lim_{t \rightarrow 0^+} \frac{\underline{v}(A_P + th_A, b_P + th_b) - \underline{v}(A_P, b_P)}{t} = \min_{\theta \in \underline{\theta}(P)} \{\underline{\lambda}(P)'(h_A \theta - h_b)\}. \quad (116)$$

Similarly,

$$\theta_i^\pm(A_P, b_P) = \left| \min_{\theta \in \Theta} \max_{\gamma \in \mathbb{R}^p \times \mathbb{R}_+^{k-p}, \|\gamma\| \leq C_\Lambda, \nu \geq 0} \{\pm \theta_i + \gamma'(A_P \theta - b_P) + \nu(\theta_1 - \underline{v}(P) - \mu_n)\} \right|. \quad (117)$$

Once again, by Theorem 7.28 from Shapiro et al. (2014) for any direction $(h_A, h_b) \in \mathbb{R}^{k \times (d+1)}$ we have

$$\dot{\theta}_i^\pm(P|h_A, h_b) = \min_{\theta \in \arg\min(117)} \{\underline{\gamma}(P)'(h_A \theta - h_b)\} + \underline{v}(P) \min_{\vartheta \in \underline{\theta}(P)} \{\underline{\lambda}(P)'(h_A \vartheta - h_b)\}. \quad (118)$$

Here we used Proposition 2.47 from Bonnans and Shapiro (2000), the chain rule for directional derivatives. By the same proposition, the vector with maximal components $\theta^*(P)$ is directionally differentiable.

Delta method (Theorem 7.67 in Shapiro et al. (2014)) and Lemma 9 imply

$$\sqrt{n}(\theta^*(\mathbb{P}_n) - \theta^*(P)) = \dot{\theta}^*(P|G_n(P)) + o_p(1). \quad (119)$$

By compactness of \mathcal{P} , the vanishing term $o_p(1)$ is bounded in probability uniformly in $P \in \mathcal{P}$. It is a routine to compute a uniform bound on the directional derivative, $\left\| \dot{\theta}^*(P|h_A, h_b) \right\| \leq L_{\mathcal{P}} < \infty$. The constant $L_{\mathcal{P}}$ provides a uniform asymptotic bound

$$\sqrt{n} \|\theta^*(\mathbb{P}_n) - \theta^*(P)\| \leq L_{\mathcal{P}} \|G_n(P)\| + O_p(1). \quad (120)$$

By Lemma 9, this representation implies (114). □

Proof of Theorem 4. The proof is analogous for all CI. Consider, for example, $\text{CB}_{\alpha, n, P}$. Pick an arbitrary measure $P \in \mathcal{P}$. Consider any $\delta > 0$ such that $z_{1-\alpha} \sigma^0 > 2\delta > 0$. Then by Lemma 12 and Theorem 2 correspondingly there exist $n(\delta, \epsilon)$ such that for any $n > n(\delta, \epsilon)$

$$\inf_{P \in \mathcal{P}} P\{\mu_n \sqrt{n} \left| \|\theta^*(\mathbb{P}_n)\|^2 - \|\theta^*(P)\|^2 \right| \leq \delta\} \geq 1 - \epsilon, \quad (121)$$

$$\inf_{P \in \mathcal{P}} P\{z_{1-\alpha} |\underline{\sigma}(\mu_n, \mathbb{P}_n) - \underline{\sigma}(\mu_n, P)| \leq \delta\} \geq 1 - \epsilon, \quad (122)$$

$$\sup_{P \in \mathcal{P}} \rho_n(P) \leq \delta \quad (123)$$

By construction $\|\theta^*(P)\| \geq \min_{\theta \in \underline{\theta}(P)} \|\theta\|$, so by Theorem 1

$$\underline{\theta}_1(\mu_n, P) + \mu_n (\|\underline{\theta}(\mu_n, P)\|^2 - \|\theta^*(P)\|^2) \leq \underline{v}(P). \quad (124)$$

³²Since the space of A_P, b_P is finite dimensional, Gâteaux and Hadamard directional derivatives coincide.

Using this bound,

$$\begin{aligned}
& P \left\{ \underline{v}(\mu_n, \mathbb{P}_n) - \mu_n \|\theta^*(\mathbb{P}_n)\|^2 - \underline{\sigma}(\mu_n, \mathbb{P}_n) z_{1-\alpha} n^{-1/2} \leq \underline{v}(P) \right\} \geq \\
& P \left\{ \sqrt{n}(\underline{v}(\mu_n, \mathbb{P}_n) - \underline{v}(\mu_n, P)) - \mu_n \sqrt{n}(\|\theta^*(\mathbb{P}_n)\|^2 - \|\theta^*(P)\|^2) \leq \underline{\sigma}(\mu_n, \mathbb{P}_n) z_{1-\alpha} \right\} \geq \text{(by (75))} \\
& P \left\{ \sqrt{n}(\underline{v}(\mu_n, \mathbb{P}_n) - \underline{v}(\mu_n, P)) \leq \underline{\sigma}(\mu_n, P) z_{1-\alpha} - 2\delta \right\} (1 - \epsilon)^2 \geq \\
& \quad \left(\Phi(z_{1-\alpha} - \frac{2\delta}{\sigma_0}) - \delta \right) (1 - \epsilon)^2
\end{aligned}$$

Since P is arbitrary, for any $n > n(\delta, \epsilon)$

$$\inf_{P \in \mathcal{P}} \min_{\theta \in \Theta(P)} P(\theta_1 \in \text{CB}_{\alpha, n, \mathcal{P}}) \geq \left(\Phi(z_{1-\alpha} - \frac{2\delta}{\sigma_0}) - \delta \right) (1 - \epsilon)^2. \quad (125)$$

Hence,

$$\liminf_{n \rightarrow \infty} \inf_{P_n \in \mathcal{P}} \min_{\theta \in \Theta(P_n)} P_n(\theta_1 \in \text{CB}_{\alpha, n, \mathcal{P}}) \geq (1 - \alpha). \quad (126)$$

□