Inference on scalar parameters
in set–identified affine models
Job market paper*

Bulat Gafarov†

Monday 21st November, 2016

Abstract

This paper proposes both point-wise and uniform confidence intervals (CIs) for an element $\theta_1$ of a parameter vector $\theta \in \mathbb{R}^d$ that is partially identified by affine moment equality and inequality conditions. The CIs are based on an estimator of a regularized support function of the identified set and have a closed–form. I provide examples in which my CIs are shorter (with probability approaching 1) than those in the existing literature. Furthermore, unlike much of the existing literature, the proposed CIs can be computed as a solution to a convex optimization problem, which leads to a substantial decrease in computation time (relative to the existing uniform procedures). My approach can be used, for instance, to compute a CI for the returns to schooling using income bracket data without strong distributional assumptions.

Key Words: Affine moment inequalities; Delta–Method; Interval data; Partial identification; Regularization; Stochastic Programming; Subvector inference; Uniform inference.

*Please, check the latest version on the website http://www.personal.psu.edu/bzg134/index.html.
†Penn State University, Department of Economics. E-mail: bzg134@psu.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, Isaiah Andrews, Andres Aradillas-Lopez, Ronald Gallant, Michael Gechter, Marc Henry, Keisuke Hirano, Sung Jae Jun, Nail Kashaev, Francesca Molinari, Xiaoxia Shi, Jing Tao, Fang Zhang for the comments and suggestions on this project. The first draft was: November 10th, 2015.
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.

I propose confidence intervals (CIs) for an element \(\theta_1\) 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, \(\theta_1\) 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 \(\theta_1\), 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 and uniform\(^3\) CIs for \(\theta_1\).

The proposed CIs have several attractive statistical and computational properties. I prove that in large samples my point-wise CI is not longer than that of Freyberger and Horowitz (2015, FH) and provide an example where my CI is shorter. Monte Carlo experiments suggest that my uniform CI can be shorter than the projection CI of Andrews and Soares (2010, AS). My approach requires only a fraction of the computational time of the AS procedure if \(\theta\) 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. Finally, my 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, KMS).

---

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

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

\(^3\) Importance of uniformly valid CI in partially identified models was first pointed out in Imbens and Manski (2004). The authors show that CI which are not uniformly valid can have poor finite sample properties.
This paper is the first to propose a closed-form estimator of the bounds on \( \theta_1 \) 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. (2015), among others) has non-Gaussian asymptotic distribution, which complicates inference. The regularization introduces a bias in the estimator of the support function that can be corrected without a change in the asymptotic distribution. The correction provides a CI with an exact asymptotic coverage probability for a fixed data generating process (DGP). To control the coverage uniformly, I propose an alternative correction that makes the CIs longer than necessary for some DGP sequences.

The computation time for my procedure increases only slowly in the dimension of \( \theta \in \mathbb{R}^d \) and takes 1.5 sec only for \( d = 15 \) and 30 moment inequalities. 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. 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\(^5\) in an affine model with \( d = 15 \) and 30 moment inequalities which is 420 times slower than my procedure. 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.

The class of DGPs over which I prove uniform coverage properties is not nested in the classes considered in BCS and KMS. I impose rank conditions on the affine constraints typical for the support function approach (Beresteanu and Molinari (2008), Kaido and Santos (2014), FH, Gafarov

\(^4\)AS, KMS, and BCS procedures can be applied in some setups where my procedure is not applicable. I compare setups in Section 3.

\(^5\)I 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.
et al. (2015)). 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 $\theta$, which can be addressed using BCS and KMS procedures. Within this framework, my procedure covers $\theta_1$ for any sequences of DGP that drift to a DGP with a moment condition orthogonal to $\theta_1$, which is not the case in BCS. 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 regularized support function proposed in this paper is a solution to a convex quadratic program that minimizes the sum of $\theta_1$ and a penalty $\mu_n \|\theta\|^2$ with $\mu_n \to 0$, subject to 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. The standard errors are computed using the sample variance of the weighted moment conditions at the unique optima. To correct the regularization bias exactly, I suggest using the argmin of the regularized program with a larger tuning parameter $\kappa_n \to 0$. If $\kappa_n/\mu_n \to \infty$ as $n \to \infty$, 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 the maximum of $\mu_n \|\theta\|^2$ over the identified set.

The paper is structured as follows. Section 2 describes the setup and procedure for a fixed DGP. Section 3 studies the properties of the CI under drifting DGP. Section 4 discusses the computational properties of my procedure. Section 5 provides the results of the Monte Carlo experiments. Section 6 concludes and discusses the possible extensions of my research.

Notation I use $\triangleq$ to denote definitions. I write $\mathbb{E}_P[\cdot]$ to denote expectation with respect to a probability distributions $P$. I use $\mathbb{P}_n$ for the sample distribution, so that $\mathbb{E}_{\mathbb{P}_n} W \triangleq \mathbb{W}_n \triangleq \frac{1}{n} \sum_{i=1}^n W_i$. Boldface uppercase letters denote matrices, $\mathbb{W}$; boldface lowercase letters denote vectors, $\mathbf{w}$; scalar variables can be both lowercase and uppercase and are not boldface, $\bar{y}$ or $W_{r,t}$. I use $f(0+)$ for $\lim_{x \downarrow 0} f(x)$. The vector $e_j \triangleq (0, \ldots, 1, \ldots, 0)'$ is the $j$-th coordinate vector, where the one occurs at position $j$. is the projector on the $j$-th coordinate. I use the same symbol for a

---

6See KMS for a discussion of the assumptions in BCS
finite set of indices $I \subset \mathbb{N}$ and the corresponding coordinate projection matrix. I use $|I|$ to denote cardinality of the set $I$. I omit the observation index $i$ throughout the paper to avoid multiple subscripts. The acronym u.h.c. stands for for upper hemicontinuous correspondence. I will use symbol $s\text{Var}(x)$ to denote $\overline{x^2_n} - (\overline{x}_n)^2$, sample variance.

2 Asymptotically exact CI under a fixed DGP

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

$$-\infty < a_{\ell} \leq \theta_{\ell} \leq b_{\ell} < \infty$$

(1)

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

$$\begin{cases}
\mathbf{e}'_j (\mathbb{E}_P X \theta + \mathbb{E}_P w) = 0, & j \in I^{eq}, \\
\mathbf{e}'_j (\mathbb{E}_P X \theta + \mathbb{E}_P w) \leq 0, & j \in I^{ineq},
\end{cases}$$

(2)

where $|I^{eq}| = p, 0 \leq p \leq d$, $|I^{ineq}| = k - p \geq 2d$, $k < \infty$, $\theta \in \mathbb{R}^d$, the random matrix $W = (X, w)$ has probability measure $P$ with sample space $\mathbb{R}^{k \times (d+1)}$. Correspondingly, $|I^{eq} \cup I^{ineq}| = k$. 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 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.

Without loss of generality, the object of interest in this paper is $\theta_1 = e'_1 \theta$, the value of the first component of $\theta \in \Theta (P)$.

Definition 1. The marginal identified set for $\theta_1$ is the set

$$S(P) = \{e'_1 \theta | \theta \in \Theta (P)\}.$$  

Since all the moment conditions are affine, the identified set $\Theta (P)$ is a polytope and the marginal
identified set is an interval, \( S(P) = [v(P), \bar{v}(P)] \), where

\[
\begin{align*}
    v(P) &= \min_{\theta \in \Theta(P)} e_j' \theta, \\
    \bar{v}(P) &= -\min_{\theta \in \Theta(P)} \{-e_j' \theta\}.
\end{align*}
\]

(3)

It is possible that \( S(P) \) is a singleton. The value functions \( v(P) \) 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.

**Example 1** (Linear model with interval valued outcome). Consider a linear model \( \mathbb{E}_P [y|z] = \theta' z \), where \( y \) is unobserved. One can only observe bounds \( \underline{y}, \bar{y} \) (random or deterministic) on \( y \) such that \( y \in [\underline{y}, \bar{y}] \) a.s. Suppose that \( z \), the random vector of regressors, has a finite support \( S_Z = \{z_1, \ldots, z_K\} \subset \mathbb{R}^d \). In this case the model can be equivalently characterized\(^7\) by a finite number of conditional moments:

\[
\mathbb{E}_P [\underline{y} | z = z_j] \geq \theta' z_j \geq \mathbb{E}_P [\bar{y} | z = z_j], \quad j = 1, \ldots, K.
\]

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

\[
\begin{align*}
    \mathbb{E}_P [\underline{y} 1 \{z = z_j\}] &\leq \theta' z_j \mathbb{E}_P 1 \{z = z_j\}, & j = 1, \ldots, K, \\
    \mathbb{E}_P [\bar{y} 1 \{z = z_{j-K}\}] &\geq \theta' z_j \mathbb{E}_P 1 \{z = z_{j-K}\}, & j = K + 1, \ldots, 2K.
\end{align*}
\]

(4)

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

\[
e_j' X \triangleq \begin{cases} 
    -z_j' 1 \{z = z_j\}, & \text{for } j = 1, \ldots, K, \\
    z_j' 1 \{z = z_{j-K}\}, & \text{for } j = K + 1, \ldots, 2K,
\end{cases}
\]

\(^7\)If \( S_Z \) is infinite, one can estimate an enlargement of \( S(P) \) using as 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.
\[
e'_j w \triangleq \begin{cases} 
p_1 \{ z = z_j \}, & \text{for } j = 1, \ldots, K, \\
-p_1 \{ z = z_{j-K} \}, & \text{for } j = K + 1, \ldots, 2K.
\end{cases}
\]

One can incorporate additional information such as sign restrictions on \( \theta \) in the form of linear inequalities to get a smaller identified set. One can also study identified sets for \( \gamma \triangleq \theta' z^* \) by adding this equation to (4). \( \square \)

Example 1 has discrete-valued regressors which makes the existing approach of Bontemps et al. (2012) inapplicable. These authors 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. The following example illustrates that in applications regressors can be predominantly discrete.

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:

\[
E[y|z] = \theta' z, \tag{5}
\]

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 \( \theta_1 \) 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 \( y, \bar{y} \) 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

$$\mathbb{E} \left( y - 0.5(\bar{y} + \underline{y}) \right) z = 0. \quad (6)$$

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 $S(P)$. All the explanatory variables are discrete, so the existing approach of Bontemps et al. (2012) is not applicable. □

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

Assumption 2 (Bounds on the moments). There exist an $\varepsilon > 0$ such that

$$\mathbb{E}_P \|W\|^{2+\varepsilon} < \infty. \quad (7)$$

Assumptions 1 and 2 are standard in the literature.

Assumption 3. $\Theta(P)$ is nonempty for the probability measure $P$.

One can replace the domain $\Theta(P)$ in (3) by $\Theta(\mathbb{P}_n)$, defined as a set of solutions to

$$\begin{align*}
\left\{ 
\begin{array}{ll}
e_j' (\bar{X}_n \theta + \bar{w}_n) = 0, & j \in I^{eq}, \\
e_j' (\bar{X}_n \theta + \bar{w}_n) \leq 0, & j \in I^{ineq},
\end{array}
\right. \quad (8)
\end{align*}$$

to get a consistent estimator for $\nu(P)$. Shapiro (1993) shows that under Assumptions 1, 3 such an estimator has a non-Gaussian asymptotic distribution. The asymptotic distribution depends on the distribution of the vector $W(\theta', 1)'$ at every point $\theta$ in the argmin set for (3) which can be set–valued. This fact complicates inference in the existing literature (FH, BCS, KMS among

---

8 Assumption 1 can be relaxed to allow for dependent data if one assumes regularity conditions that are sufficient to guarantee a Functional Central Limiting theorem for the first moments of $W$ and the Uniform Law of Large numbers for the second moments of $W$. 
others), since one has to estimate a superset for the argmin set to compute the critical values. As I will discuss later in Example 5, that the argmin set can be large if $d$ is large. Instead, one can approximate the program (3) by its regularized counterpart with the unique solution,

$$
u(\mu_n; P) = \min_{\theta \in \Theta(P)} \theta'_1 + \mu_n \|\theta\|^2,$$  \hspace{1cm} (9)

The argmin to (9), $\vartheta(\mu_n; P)$, is a continuous and single–valued function of the tuning parameter $\mu_n$, that shrinks to zero as the sample size grows to infinity. The one–sided limit is the element with the minimal norm in to the argmin set of [3],

$$\vartheta(0+; P) \triangleq \lim_{\mu_n \to \infty} \vartheta(\mu_n; P) = \arg \min_{\theta \in \Theta(P), \theta_1 = \nu(P)} \|\theta\|.$$

If $\mu_n$ converges to zero sufficiently slowly, then

$$\nu(\mu_n; \mathbb{P}_n) = \min_{\theta \in \Theta(\mathbb{P}_n)} \theta'_1 + \mu_n \|\theta\|^2,$$ \hspace{1cm} (10)

is a consistent estimator of $\nu(P)$ with asymptotic distribution that depends only on the distribution of $W(\vartheta'(0+; P), 1)'$. The asymptotic analysis can be further simplified if moment conditions exactly identify $\vartheta(0+; P)$ and all linear functions of its components.

**Assumption 4. (No overidentifying moment conditions)**

1. For any $I \subset I^{ineq}$ with $|I| \geq d + 1 - p$ and any $\theta \in \Theta(P)$ \hspace{1cm} (10)

$$\left( \begin{array}{c} I^{eq} \\ I \end{array} \right) (\mathbb{E}_P X\theta + \mathbb{E}_P w) \neq 0.$$

\[\text{See Lemma 5 in Appendix.}\]
\[\text{I use the same symbol for a finite set of indices } I^{eq} \subset \mathbb{N} \text{ and the corresponding coordinate projection matrix.}\]
2. For any $\mathcal{I} \subset \mathcal{I}^{ineq}$ with $|\mathcal{I}| = d - p$, 

$$\text{rk} \left[ \begin{pmatrix} \mathbf{I}^{eq} \\
\mathcal{I} \end{pmatrix} \mathbb{E}_P (\mathbf{X}, -\mathbf{w}) \right] = d.$$ 

Assumption 4.1 implies that at any point $\theta$ in $\Theta$ at most $d$ moment conditions are binding, which is a sufficient condition to avoid over-identification of $\mathbf{\theta}' (0+; P)$. Assumption 4.2 implies that any system of $d$ active moment conditions has a unique solution, which rules out over-identification of $\mathbf{\zeta}' \mathbf{\theta} (0+; P)$ for any $\mathbf{\zeta} \in \mathbb{R}^d$. Together, these two conditions imply that at any point in $\theta \in \Theta (P)$ any binding (or active) moment conditions have linearly independent gradients, a condition called Linear Independence Constraint Qualification (LICQ) in the optimization theory. Under LICQ, the solution to (9) for $\mu_n > 0$ is strongly regular\footnote{A solution $\theta$ is strongly regular iff the LICQ and the strong of second order sufficient condition hold at $\theta$. See p.179 in \cite{Shapiro2014}.} so that its sample analog has asymptotic Gaussian distribution. Assumption 4 is sufficient but not necessary for LICQ since it implies rank restrictions on the gradients of moment conditions outside of $\Theta (P)$\footnote{LICQ (Assumption 6 in Appendix) is sufficient for Theorem 1.}.

Assumption 4 or LICQ are common in the literature on set identified models. In particular, BM, KS, FH and Gafarov et al. (2015) use it in various forms. Assumption 4 is formulated as a rank constraint for a finite number of matrices. LICQ, in contrast, imposes the rank on the gradients of active constraints at every point a set $\Theta (P)$, which makes it harder to check. Both conditions in Assumption 4 can, in principle, be tested using either a rank test for all combinations of $d$ moment conditions or a $J$–test for all combinations of $d + 1$ moment conditions. Both LICQ and Assumption 4 can be violated in applications with too many moment conditions that include Shi and Shum (2016).

The asymptotic variance of $\psi (\mu_n; P_n)$, $\sigma^2 (0+; P)$ can be represented in closed form using $\mathbf{\lambda} (P) \in \mathbb{R}^k$, which is a solution to the dual program for (3), i.e.

$$\psi (P) = \max_{\mathbf{\lambda} \in \mathbb{R}^p \times \mathbb{R}^{k-p}_+} \left\{ -\mathbf{\lambda}' \mathbb{E}_P \mathbf{w} \right\} \quad (12)$$

$$\text{s.t.} \mathbf{\lambda}' \mathbb{E}_P \mathbf{X} = e_1'.$$
Under Assumption 4, $\lambda(P)$ is unique, so

$$\sigma^2(0+; P) = \text{Var} \left[ \lambda'(P) W(\vartheta'(0+; P), 1) \right]' .$$

**Assumption 5.** The probability measure $P$ satisfies $\sigma^2(0+; P) > 0$ and $\sigma^2(0+; P) > 0$.

Assumption 5 implies that there is at least one active moment condition and it has positive variance at $\vartheta(0+; P)$. It also implies that the lower bound is not attained at the boundary of the domain, $v(P) > a_1$. It allows, however, for some equality or active inequality restrictions with zero variance if at least one moment condition has positive variance and a positive Lagrange multiplier. If Assumption 5 is violated, then $v(\mu_n; P)$ is a superconsistent estimator for $v(P)$. This case requires separate consideration in the Delta–method framework used in this paper.

I estimate the variance $\sigma^2(0+; P)$ by

$$\sigma^2(\mu_n; P) = \text{sVar} \left[ \lambda'(\mu_n; P) W(\vartheta'(\mu_n; P), 1) \right]' ,$$

where $\text{sVar}(\cdot)$ is the sample variance operator, $\vartheta(\mu_n; P)$ and $\lambda(\mu_n; P)$ are, respectively, the optimum and the vector of Lagrange multipliers of (10).

The estimators $v(\mu_n; P)$ and $\bar{v}(\mu_n; P)$ have $1/\sqrt{n}$-bias that can be corrected,

$$\hat{v}_n \triangleq v(\mu_n; P) - \mu_n \| \vartheta(\kappa_n; P) \|^2,$$

$$\hat{v}_n \triangleq \overline{v}(\mu_n; P) + \mu_n \| \vartheta(\kappa_n; P) \|^2 .$$

(13)

If $\kappa_n$ converges to zero slower than $\mu_n$, then $\hat{v}_n$ and $v(\mu_n; P)$ have the same asymptotic variance. Using the bias corrected estimators $\hat{v}_n$ and $\hat{v}_n$ and their variances, I construct Delta–method confidence sets:

---

13One can replace Assumption 5 with a one-sided version if one is interested only in $\text{CI}^L_{\alpha,n}$ or $\text{CI}^R_{\alpha,n}$.
\[
\begin{align*}
\text{CI}_L &= \hat{v} - z_{1-\alpha} n^{-1/2} \sigma(\mu_n; \mathbb{P}_n), \infty,
\text{CI}_R &= -\infty, \hat{v} + z_{1-\alpha} n^{-1/2} \sigma(\mu_n; \mathbb{P}_n),
\text{CI}_{\alpha,n} &= \hat{v} - z_{1-\alpha} n^{-1/2} \sigma(\mu_n; \mathbb{P}_n) + \hat{v} + z_{1-\alpha} n^{-1/2} \sigma(\mu_n; \mathbb{P}_n),
\text{CI}_B &= \text{CI}_{\alpha/2,n},
\end{align*}
\]

where \( z_{1-\alpha} \) is the \( 1 - \alpha \) quantile of the standard Gaussian distribution.

**Theorem 1.** Suppose that Assumptions 1–5 hold and that in addition \( 0 < \alpha < 1/2 \), \( \mu_n \) and \( \kappa_n \) are such that \( \kappa_n \to 0 \), \( \mu_n / \kappa_n \to 0 \) and \( \mu_n \sqrt{n} \to \infty \). Then,

\[
\lim_{n \to \infty} P \left( S(P) \subset \text{CI}_L_{\alpha,n} \right) = \lim_{n \to \infty} \inf_{\theta \in \Theta(P)} P \left( \theta_1 \in \text{CI}_L_{\alpha,n} \right) = 1 - \alpha,
\]

\[
\lim_{n \to \infty} P \left( S(P) \subset \text{CI}_R_{\alpha,n} \right) = \lim_{n \to \infty} \inf_{\theta \in \Theta(P)} P \left( \theta_1 \in \text{CI}_R_{\alpha,n} \right) = 1 - \alpha,
\]

\[
\lim_{n \to \infty} P \left( S(P) \subset \text{CI}_B_{\alpha,n} \right) \geq 1 - \alpha, \quad \lim_{n \to \infty} \inf_{\theta \in \Theta(P)} P \left( \theta_1 \in \text{CI}_B_{\alpha,n} \right) \geq 1 - \alpha.
\]

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

\[
\lim_{n \to \infty} \inf_{\theta \in \Theta(P)} P \left( \theta_1 \in \text{CI}_{\alpha,n} \right) = 1 - \alpha.
\]

**Proof.** See Appendix 7.3

The choice of \( \mu_n \) and \( \kappa_n \) is discussed in Section 5.

The component \( \theta_1 \) is point identified if there are equality constraints in the model that are orthogonal to \( e_1 \). If \( p > 0 \), I recommend the Bonferroni–type confidence set \( \text{CI}_{\alpha,n}^B \) which remains valid under point identification. The shorter \( \text{CI}_{\alpha,n} \) proposed by [imbens and Manski (2004)] is valid if \( \theta_1 \) is not point–identified. Under Assumption 4, \( p = 0 \) implies that \( S(P) \) is not a singleton.

FH provide an alternative CI for \( \theta_1 \) in the affine moment inequality model under set identification. To cover \( \theta_1 \) with probability at least \( 1 - \alpha \) FH proposed \( \text{CI}_{\alpha,n}^{FH} \), an intersection of one-sided bootstrap CIs, which is analogous to \( \text{CI}_{\alpha,n} \). In general, with probability approaching one \(|\text{CI}_{\alpha,n}|\),

---

14 Assumption 3 in FH is another sufficient condition for set-identification of \( \theta_1 \) for \( p \neq 0 \). The authors acknowledge that this assumption cannot be tested. In contrast, Assumption 4 is testable.

15 See Section 3.3 of Freyberger and Horowitz (2015). FH also proposed a CI for \( S(P) \). One can use the joint
the length of $\text{CI}_{\alpha,n}$ is not longer than $|\text{CI}_{\alpha,n}^{FH}|$, the length of $\text{CI}_{\alpha,n}^{FH}$, as the following proposition shows.

**Proposition 1.** Suppose that Assumptions 1–5 hold. For any sequences $\mu_n$ and $\kappa_n$ such that $\kappa_n \to 0$, $\mu_n/\kappa_n \to 0$ and $\mu_n \sqrt{n} \to \infty$,

$$\lim_{n \to \infty} P\left(|\text{CI}_{\alpha,n}^{FH}| \geq |\text{CI}_{\alpha,n}|\right) = 1.$$ 

**Proof.** See Appendix 7.4

The following example shows that $\text{CI}_{\alpha,n}$ can be shorter than $\text{CI}_{\alpha,n}^{FH}$ in large samples.

**Example 3.** Consider the system (2) with $I_{eq} = \emptyset$ and coefficients $W$ that have expectation

$$\mathbb{E}_P W = \begin{pmatrix} -I_2 & -\iota_2 \\ I_2 & -\iota_2 \end{pmatrix}.$$ 

Suppose that all components of $W$ are independent random variables with variance $s^2$. The identified set is $\Theta(P) = [-1, 1]^2$. The standard deviation of every moment condition is given by $s\sqrt{1 + \|	heta\|^2}$. The asymptotic behavior of $\text{CI}_{\alpha,n}^{FH}$ depends on the asymptotic distribution of the basic solutions to (3).16 The minimization problem (3) has two basic solutions, $(-1, 1)$ and $(-1, -1)$. The basic solutions for the maximization problem in (3) are $(1, 1)$ and $(1, -1)$. The critical value $c_{1-\alpha}^{FH}$ corresponds to the maximum of the $1 - \alpha$ quantiles of the four basic solutions17

$$c_{1-\alpha}^{FH} = z_{1-\alpha} s \max_{\theta_1 = \pm 1, \theta_2 = \pm 1} \left\{ \sqrt{1 + \|	heta\|^2} \right\} = z_{1-\alpha} s \sqrt{3}.$$ 

The solution to (10) and to the upper bound counterpart converge to the optima of (3) with the asymptotic Gaussian distribution of $\hat{\theta}_n$ and $\hat{v}_n$ to construct an analogous Delta-method CI for $S(P)$, which can be shorter than $\text{CI}_{\alpha,n}^{FH}$.

16See Section 3 in Freyberger and Horowitz (2015) for details.

17See Section 3.3 in Freyberger and Horowitz (2015) for details.
minimal norm\textsuperscript{18} i.e.
\[
\vartheta' (\mu_n; \mathbb{P}_n) \xrightarrow{P} (-1, 0),
\]
\[
\bar{\vartheta}' (\mu_n; \mathbb{P}_n) \xrightarrow{P} (1, 0).
\]
So the critical values for both sides of CI\textsubscript{α,n} are \(z_{1-\alpha} s\sqrt{2}\) which leads\textsuperscript{19} to
\[
\text{CI}_{\alpha,n}^{FH} - |\text{CI}_{\alpha,n}| = 2z_{1-\alpha} s \left(\sqrt{3} - \sqrt{2}\right) n^{-1/2} + o_p n^{-1/2}.
\]
As sample size grows, CI\textsubscript{α,n} becomes shorter than CI\textsubscript{α,n}FH with probability approaching 1,
\[
\lim_{n \to \infty} P \text{ CI}_{\alpha,n}^{FH} > |\text{CI}_{\alpha,n}| = 1.
\]
\[\square\]

3 Uniform confidence sets

The distribution of the estimator \(\underline{v}(\mu_n; \mathbb{P}_n)\) has a uniform asymptotic Gaussian approximation over the following class of DGPs.

**Definition 2.** The class of distributions \(\mathcal{P} \in \tilde{M}, \epsilon, \eta, \sigma_0\) is the set of all Borel measures \(P\) on \(\mathbb{R}^{k \times (d+1)}\) satisfying the following conditions

1. \(\mathbb{E}_P \|W\|^{2+\epsilon} \leq \tilde{M}\), where \(\tilde{M}\) is a finite constant.

2. \(\Theta(P)\) is nonempty.

3. For any \(\mathcal{I} \subset \mathcal{I}_{ineq}\) with \(|\mathcal{I}| \geq d + 1 - p\)

\[
\min_{\theta \in \Theta(P)} \left[ \left( \mathbb{E}_P X \theta + \mathbb{E}_P w \right)' \left( \begin{array}{c} \mathcal{I}^q \\ \mathcal{I} \end{array} \right) \right] \geq \epsilon > 0.
\] (16)

4. For any \(\mathcal{I} \subset \mathcal{I}_{ineq}\) with \(|\mathcal{I}| = d - p\) the following matrix has minimum eigenvalues bounded

\textsuperscript{18}See Lemma\textsuperscript{12}.
\textsuperscript{19}See Appendix\textsuperscript{7.4} for details.
by a positive number,

\[ \text{eig} \left[ \begin{pmatrix} I^q \\ \mathcal{I} \end{pmatrix} \mathbb{E}_P (\mathbf{X}, -\mathbf{w}) \mathbb{E}_P (\mathbf{X}, -\mathbf{w})' \begin{pmatrix} I^q \\ \mathcal{I} \end{pmatrix} \right] \geq \eta^2 > 0. \]

5. For every \( \theta \in \Theta (P) \) such that \( \theta_1 = v (P) \)

\[ \text{Var} \ \mathbf{X}' (P) \mathbf{W} (\theta', 1)' \geq \sigma_0^2; \]

for every \( \theta \in \Theta (P) \) such that \( \theta_1 = \bar{v} (P) \)

\[ \text{Var} \ \mathbf{X}' (P) \mathbf{W} (\theta', 1)' \geq \sigma_0^2. \]

\( \square \)

Definition 2.3 rules out the possibility of point identification by moment inequalities. This condition can be violated in applications, see e.g. Gafarov (2014) and Shi and Shum (2016). It is possible, however, to have \( \{P_n\}_{n=1}^\infty \subset \mathcal{P} \) \( \varepsilon, \bar{M}, \epsilon, \eta, \sigma_0 \) with a set-valued sequence \( S (P_n) \) converging to a singleton if the gradients of the equality moment conditions become linearly dependent with \( \mathbf{e}_1 \). Class \( \mathcal{P} \) \( \varepsilon, \bar{M}, \epsilon, \eta, \sigma_0 \) is compact in weak topology \( \text{20} \) so it contains the weak limit of any \( \{P_n\}_{n=1}^\infty \subset \mathcal{P} \) \( \varepsilon, \bar{M}, \epsilon, \eta, \sigma_0 \).

Under drifting DGP sequences it is possible that a sequence of regularized solutions \( \hat{\theta} (\mu_n; P_n) \) converges to any point in the argmin set of (3), where \( P_n \leadsto P \). So the property in Definition 2.5 requires positive variance for every point in the argmin set of (3).

The estimator \( \hat{v}_n \) defined in (13) has asymptotic Gaussian distribution and converges to a non-regular function \( v (P) \). By the impossibility theorem of Hirano and Porter (2012), such an estimator is biased for some sequences of DGPs. Indeed, for drifting DGPs it is possible that \( \lim_{n \to \infty} \| \hat{\theta} (\kappa_n; P_n) \|^2 \) and \( \lim_{n \to \infty} \| \hat{\theta} (\mu_n; P_n) \|^2 \) are different. As a result, there is a non-zero bias

\[ \lim_{n \to \infty} \mu_n \sqrt{n} \| \hat{\theta} (\mu_n; P_n) \|^2 - \| \hat{\theta} (\kappa_n; P_n) \|^2 \neq 0 \]

\( \text{20} \) See Lemma 6 in Appendix
in the asymptotic distribution of $\hat{v}_n$. Such bias can be positive which would reduce the asymptotic coverage probability of the CIs below the nominal level. A simple solution to this problem is to replace $\|\hat{\theta} (\kappa_n; P_n)\|^2$ in (13) to get

$$\hat{\nu}_n = \nu (\mu_n; P_n) - \mu_n \beta (P_n), \quad (17)$$

$$\hat{\nu}_n = \nu (\mu_n; P_n) + \mu_n \beta (P_n),$$

where

$$\beta (P_n) = \max_{\theta \in \Theta (P_n)} \|\theta\|^2. \quad (18)$$

The asymptotic bias in this case is $\mu_n \sqrt{n} \|\hat{\theta} (\mu_n; P_n)\|^2 - \beta (P_n) \leq 0$. It is bounded in absolute value by $\mu_n \sqrt{n} \beta (P_n)$ \(^{21}\). The bias can be equal to zero for some sequence $P_n$ if

$$\lim_{n \to \infty} \|\hat{\theta} (\mu_n; P_n)\|^2 = \lim_{n \to \infty} \beta (P_n),$$

in which case the asymptotic coverage probability of the resulting CI is equal to the nominal size.

Let $\tilde{C}_{\alpha,n}^L$, $\tilde{C}_{\alpha,n}^R$, $\tilde{C}_{\alpha,n}^B$, and $\tilde{C}_{\alpha,n}$ the confidence sets defined in (14) with $\hat{\nu}_n$ and $\hat{v}_n$ being replaced by $\hat{\nu}_n$ and $\hat{v}_n$, correspondingly.

**Theorem 2.** Consider any positive numbers $\varepsilon, M, \epsilon, \eta, \sigma_0$ and the corresponding $\mathcal{P} = \mathcal{P} \varepsilon, M, \epsilon, \eta, \sigma_0$. Suppose that Assumption \(^1\) holds. In addition, suppose that $0 < \alpha < 1/2$, $\mu_n \to 0$ and $\mu_n \sqrt{n} \to \infty$.

Then the following results hold,

$$\lim_{n \to \infty} \inf_{P \in \mathcal{P}} \inf_{\theta \in \Theta (P)} \mathbb{P} (\theta \in \tilde{C}_{\alpha,n}^L \geq 1 - \alpha), \quad \lim_{n \to \infty} \inf_{P \in \mathcal{P}} \inf_{\theta \in \Theta (P)} \mathbb{P} (\theta \in \tilde{C}_{\alpha,n}^R \geq 1 - \alpha),$$

$$\lim_{n \to \infty} \inf_{P \in \mathcal{P}} \min_{\theta \in \Theta (P)} \mathbb{P} (\theta \in \tilde{C}_{\alpha,n}^L \geq 1 - \alpha), \quad \lim_{n \to \infty} \inf_{P \in \mathcal{P}} \min_{\theta \in \Theta (P)} \mathbb{P} (\theta \in \tilde{C}_{\alpha,n}^R \geq 1 - \alpha),$$

$$\lim_{n \to \infty} \inf_{P \in \mathcal{P}} \mathbb{P} (\theta \in \tilde{C}_{\alpha,n}^B \geq 1 - \alpha), \quad \lim_{n \to \infty} \inf_{P \in \mathcal{P}} \inf_{\theta \in \Theta (P)} \mathbb{P} (\theta \in \tilde{C}_{\alpha,n}^B \geq 1 - \alpha).$$

\(^{21}\)Such correction is unnecessarily large. Notice that $\beta (P_n) = -\lim_{n \to \infty} \mu_n^{-1} \mu (\theta; P_n)$. A better option might be to use $\|\hat{\theta} (\mu_n; P_n)\|^2$ instead of $\beta (P_n)$ which could eliminate the bias for any DGP with unique optimum. The theory behind this option is under investigation.
If the model has no equality constraints, i.e. if $p = 0$, then

$$\lim_{n \to \infty} \inf_{P \in P} \inf_{\theta \in \Theta(P)} \min_{P} P \theta_1 \in \widetilde{CI}_{\alpha,n} = 1 - \alpha. \quad (19)$$

Proof. See Appendix 7.5. \qed

BCS and KMS provide alternative ways to construct a uniform CI for $\theta_1$. Both procedures can deal with non-linear moment conditions and overidentifying restrictions. My analysis is restricted to a smaller class of affine models without overidentifying conditions (in the sense of Assumption (4)). Within this smaller class of models, I am able to relax the following assumptions of BCS and KMS while still providing uniform coverage.

1. Assumption A3.a in BCS which rules out sequences of DGP that converge to a DGP with multiple points in the argmin sets in (3). KMS call these near-flat-face designs.

2. Assumptions 3.2.(iii) and 3.4 in KMS and Assumption A.3.(c) in BCS prohibit moment conditions with zero variance at any $\theta \in \Theta$.

The second point precludes the use of BCS and KMS in the linear model with interval outcome data described in Example 1.

Example 4 (Example 1 continued). Suppose that for some support point $z_0$ the lower bound on the outcome is deterministic, i.e. $\mathbb{E}_P y|z = z_0 = y_0$ and $\text{Var}_P y|z = z_0 = 0$. It can take place in Example 2 if the lower bound on the labour income for some particular social group is zero for all individuals in the population, i.e. $y_0 = 0$. The corresponding moment inequality condition is

$$\mathbb{E}_P y - z_0' \theta' \mathbb{1} \{z = z_0\} \leq 0. \quad (20)$$

The standard deviation of the moment condition is $|z_0' \theta - y_0| \sqrt{p_z (1 - p_z)}$, where $p_z = \mathbb{E}_P \mathbb{1} \{z = z_0\}$.

The standard deviation is equal to zero for any solution of the linear equation $z_0' \theta = y_0$. These

---

\textsuperscript{22}The implementation of the KMS procedure, however, uses a Newton-type solver which only finds local optima satisfying a constraint qualification condition. So implicitly KMS also assume some constraint qualification for the moment conditions.
values of $\theta$ make the constraint (20) binding. This feature makes procedures of AS, KMS, and BCS inapplicable, since all the three procedures use standardized moment conditions. The standardized version of the moment condition (20) also violates the smoothens assumptions in KMS and BCS,

$$\text{sign}(y_0 - z_0'\theta) \sqrt{p_z/(1 - p_z)} \leq 0.$$ 

Since this function is not differentiable in $\theta$, it violates Assumptions 3.2(iii), 3.4 in KMS and A3.a in BCS. □

4 Computation

The estimators $\vartheta(\mu_n; P_n)$ and $\tilde{\vartheta}(\mu_n; P_n)$ are based on strictly convex programs. For convex programs the set Karush–Kuhn–Tucker (KKT) conditions\(^{23}\) provide necessary and sufficient conditions for the global optimum. Moreover, the KKT system for strictly convex optimization problems has a unique solution. In contrast, for general smooth optimization programs used in the existing literature the KKT conditions under a constraint qualification\(^{24}\) are necessary but not sufficient conditions for optima. The numerical solvers like fmincon of MATLAB only find a single solution to the KKT conditions which might not be the global optimal point. In order to get an optimal solution to a non–convex optimization problem using fast Newton type solvers one needs globalization strategies like multistart from a large number of initial points. This feature can dramatically increase the computational costs. 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 $d$. The following example illustrates this point.

**Example 5** (Example 3 continued). Consider a set of moment inequalities with coefficients that

\(^{23}\)See Lemma 2 in Appendix

\(^{24}\)The KKT conditions are not necessarily satisfied for optima that violate constraint qualifications. Such optima necessarily satisfy more general Fritz John conditions. See Proposition 5.47 in Bonnans and Shapiro (2000). The standard software like fmincon of MATLAB only finds solutions to KKT system.
have expectation
\[
E_P W = \begin{pmatrix}
-I_d & -I \\
I_d & -I
\end{pmatrix}.
\]

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
\[
\pm \theta_j + 1 \leq 0, \quad j = 1, \ldots, d.
\]

(21)

The KMS procedure adds slack \(c(\theta)\) to the right hand side of every standardized moment inequality. Consider, for example, \(j = 1,\)
\[
\theta_1 \geq 1 - c(\theta) \frac{s}{1 + \|\theta\|^2}.
\]

(22)

The slack function \(c(\theta)\) is computed using a resampling on a grid of points. Assume for, simplicity, that \(c(\theta)\) is a constant. Figure 1 shows the identified set and the corresponding expansion with \(c(\theta) = \text{const}\). The optimization domain of in KMS is similar the non-convex set on right of Figure 1. 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.

The multiple local optima also affect BCS procedure and the KMS implementation of the AS procedure. KMS provide a fast numerical procedure to compute a projection of the joint AS confidence set using interpolated critical values. The domain of the corresponding optimization program is, in general, non–convex and has shape similar to Figure 1. BCS procedure minimizes the Modified Method of Moments statistics which is a function of the standardized moment conditions (21). Correspondingly this objective function is not convex and can have isolated local minima even though the domain is convex.

\[25\] Projection of the AS confidence set on \(\theta_1\) can also be constructed using a fine grid of points on \(\Theta \subset \mathbb{R}^d\). This approach is computationally intensive and would take even more time than the multistart method.
A large number of KKT points makes the procedures of KMS, AS and BCS both computationally costly and fragile for large $d$. In order to get a reliable result one need to run the optimization routine with multiple initial points in neighborhoods of all vertices of the sample identified set. As a result, the number of numerical operations grows exponentially with the dimension $d$.

Figure 3 shows the computational time for $\widetilde{CI}_{\alpha,n}$ and the AS procedure using the implementation of KMS\textsuperscript{26} in Example 5. The average computation time for $\widetilde{CI}_{\alpha,n}$ is almost insensitive to the dimension $d$. It takes 1.5 seconds to compute $\widetilde{CI}_{\alpha,n}$ for $d = 15$. The computational time for the AS procedure increases by approximately 30% with every additional dimension and takes 630 seconds to compute the CI for $d = 15$. The KMS procedure by construction is more computationally intensive than the AS procedure. The KMS procedure implemented in Matlab with precompiled solvers takes 560 sec for $d = 8$\textsuperscript{27} With estimated growth rate of 30% per dimension KMS procedure with precompiled code would take more than hour to compute a CI for $d = 15$.

The point-wise CIs in (14) can be computed using any Newton type optimization software that provides accurate Lagrange multipliers\textsuperscript{28} I use fmincon function of MATLAB software. I recommend using the 'active set' or 'SQP' options since the 'interior point' solver does not provide accurate Lagrange multipliers. The uniformly valid CIs described in Theorem 2 uses the solution to a quadratic non-convex program [18]. Even though this program is non-convex, the computational time of $\widetilde{CI}_{\alpha,n}$ compares favorably with the AS procedure. I recommend to use multiple initial points when using the fmincon function of MATLAB to make the solution more

\textsuperscript{26}The 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.

\textsuperscript{27}These numbers correspond to a different DGP. The available code of KMS achieves this speed using a precompiled optimization routine for a specific model. The build-in linear programming solvers in Matlab cannot achieve this speed.

\textsuperscript{28}The numerical properties of the point-wise FH CI are similar to my point-wise CIs.
5 Monte Carlo

5.1 Choice of the tuning parameters

The choice of the rate $\mu_n$ determines the rate of the higher order terms in the stochastic expansion of the estimators. On one hand, the bias is small if $\mu_n$ converge to 0 at fast rate. On the other hand, the regularization need to be stronger than the random sample perturbation to have the stabilizing effect on the asymptotic distribution. The square root of the iterated logarithm is the fastest rate, under which in large samples the regularization dominates the estimation noise almost surely. I use the tuning parameters $\mu_n = \hat{\mu}_1 \frac{\log \log n}{n}$ and $\kappa_n = \hat{\mu}_1 \frac{\log n}{n}$ with

$$\hat{\mu}_1 = \frac{\lambda'(0; P_n) X - \bar{X}_n X - \bar{X}_n' \lambda(0; P_n)}{\max \{\beta(P_n), 1\}}.$$ 

The choice of numerator in $\hat{\mu}_1$ is motivated by the formula for the higher order terms in the stochastic expansion of the estimator $\varphi(\mu_n; P_n)$. The denominator makes the bias $||\varphi(\mu_n; P_n)||^2 - ||\varphi(\kappa_n; P_n)||^2$, which may be non–zero for drifting DGP-2, bounded by 1. This choice gives good finite sample coverage properties for both CI$_{\alpha,n}$ and $\tilde{\text{CI}}_{\alpha,n}$. The theory behind this choice of tuning parameters is under investigation.

5.2 Two-dimensional designs

I consider four different MC designs. DGP 1-3 have 4 moment inequalities with coefficients that have the following expectation:

---

29As mentioned in Section 3, it is preferable to replace $\beta(P_n)$ by $||\varphi(-\mu_n; P_n)||^2$. One can use the estimator $\varphi(\mu_n; P_n)$ as an initial point for computation of $\varphi(-\mu_n; P_n)$ in the active set solver. For small $\mu_n$ the active sets for $\varphi(-\mu_n; P_n)$ and $\varphi(\mu_n; P_n)$ share elements which accelerates the local search. As a result, one do not have to perform the costly multistart algorithm.

---
Table 1: Parameter values for DGP 1-4

<table>
<thead>
<tr>
<th></th>
<th>(\omega_1)</th>
<th>(\omega_2)</th>
<th>(\sigma_0)</th>
<th>(p)</th>
<th>(\varphi(P))</th>
<th>(\bar{v}(P))</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGP-1</td>
<td>0.25</td>
<td>1.5</td>
<td>0.1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>DGP-2</td>
<td>(n^{-1/2})</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>DGP-3</td>
<td>0</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>DGP-4</td>
<td>(n^{-1/2})</td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>-1 + 2\tan\frac{\pi}{2}n^{-1/2}</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\mathbb{E}_P \mathbf{X} = \begin{pmatrix}
-\cos \omega_1 \pi/2 & -\sin \omega_1 \pi/2 \\
\cos \omega_1 \pi/2 & \sin \omega_1 \pi/2 \\
-\cos \omega_2 \pi/2 & -\sin \omega_2 \pi/2 \\
\cos \omega_2 \pi/2 & \sin \omega_2 \pi/2 
\end{pmatrix}, \quad \mathbb{E}_P \mathbf{w} = \begin{pmatrix}
-\cos \omega_1 \pi/2 & \zeta_1 - \sin \omega_1 \pi/2 & \zeta_2 \\
\cos \omega_1 \pi/2 & \psi_1 + \sin \omega_1 \pi/2 & \psi_2 \\
-\cos \omega_2 \pi/2 & \psi_1 - \sin \omega_2 \pi/2 & \psi_2 \\
\cos \omega_2 \pi/2 & \zeta_1 + \sin \omega_2 \pi/2 & \zeta_2 
\end{pmatrix}.
\]

The expectations are parametrized to guarantee \(\varphi(P) = \zeta_1, \bar{v}(P) = \psi_1\). DGP-4 has the first two inequality conditions replaced by a single moment equality condition

\[
\cos \omega_1 \pi/2 \theta_1 + \sin \omega_1 \pi/2 \theta_2 = \cos \omega_1 \pi/2 \zeta_1 + \sin \omega_1 \pi/2 \zeta_2.
\]

The components of \(\mathbf{W} = (\mathbf{X}, \mathbf{w})\) are independent Gaussian random variables with variance \(\sigma_0^2\). The parametrization used in the MC experiments is summarized in Table 1. The corresponding identified sets for DGP 1-3 are shown on Figure 2.

Table 2 summarizes the results of 2000 MC simulations for sample sizes \(10^2, 10^3, 10^4, 10^5\). \(\text{CI}_{\alpha,n}\) has tendency to undercover in designs with unique optimum (DGP-1,2,4) and exceed the nominal size in the presence of multiple optima (DGP-3). Both tendencies become less prominent as sample size grows.

\(\tilde{\text{CI}}_{\alpha,n}\) has only marginally higher coverage probability than \(\text{CI}_{\alpha,n}\) in DGP-1,2,4. The exact coverage of \(\tilde{\text{CI}}_{\alpha,n}\) for these DGPs is achieved by design \(\lim_{n \to \infty} \|\widehat{\beta}(-\mu_n; P_n)\|^2 = \lim_{n \to \infty} \beta(P_n)\)^{30}\n
The DGP-3, in contrast, does not meet this condition. As a result, \(\tilde{\text{CI}}_{\alpha,n}\) converges to \(\mathcal{S}(P)\) at rate \(\sqrt{n^{-1} \log \log n}\) which results in the coverage probability close to 100%.

---

30 One can achieve exact asymptotic coverage in designs with unique optimum using correction \(\|\widehat{\beta}(-\mu_n; P_n)\|^2\) instead of \(\beta(P_n)\). For DGP-1,2,4 \(\|\widehat{\beta}(-\mu_n; P_n)\|^2 = \beta(P_n)\). The theory for the alternative correction is in progress.
Table 2: MC results for DGP 1-4.

<table>
<thead>
<tr>
<th>DGP</th>
<th>1 − α</th>
<th>CL_{α,n}</th>
<th>CL_{α,n}</th>
<th>CL_{α,n}</th>
<th>CL_{α,n}</th>
<th>CL_{α,n}</th>
<th>CL_{α,n}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0.90</td>
<td>0.9620</td>
<td>0.9620</td>
<td>0.9640</td>
<td>0.9640</td>
<td>0.9640</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.95</td>
<td>0.9460</td>
<td>0.9460</td>
<td>0.9460</td>
<td>0.9460</td>
<td>0.9460</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.975</td>
<td>0.9700</td>
<td>0.9700</td>
<td>0.9700</td>
<td>0.9700</td>
<td>0.9700</td>
</tr>
<tr>
<td>Avg. Ex. Len.</td>
<td>0.0671</td>
<td>0.0671</td>
<td>0.0212</td>
<td>0.0212</td>
<td>0.0067</td>
<td>0.0067</td>
<td>0.0021</td>
</tr>
</tbody>
</table>

Note: Average excess length is measured as the difference between the average |CL_{α,n}| and |S(P)| for α = 0.1.
Figure 3: Average computation time and excess length for $\text{CI}_{\alpha,n}$ and $\widetilde{\text{CI}}_{\alpha,n}$, $n = 1000$

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

5.3 Length comparison with existing approaches

One can assess the conservativeness of $\widetilde{\text{CI}}_{\alpha,n}$ by comparing the excess average length of $\widetilde{\text{CI}}_{\alpha,n}$ with the projection of the AS confidence set as the dimension of the problem grows. DGP 5 is described in Example 5 and coincides with DGP-3 if $d = 2$. I change the standard deviation with dimension by formula $\sigma_0 = 0.1/\sqrt{1 + d}$ to keep the variance of the moment conditions at every vertex of $[-1, 1]^d$ constant. Figure 3 shows the excess average length for $\widetilde{\text{CI}}_{\alpha,n}$ and $\text{CI}_{\alpha,n}^{\text{AS}}$ for $n = 1000$. For all experiments (except for $d = 4$), the average length of $\widetilde{\text{CI}}_{\alpha,n}$ is shorter than $\text{CI}_{\alpha,n}^{\text{AS}}$. As in DGP-3, $\widetilde{\text{CI}}_{\alpha,n}$ converges to $S(P)$ at rate $\sqrt{n^{-1} \log \log n}$ while $\text{CI}_{\alpha,n}^{\text{AS}}$ has a faster rate $n^{-1/2}$. Nevertheless, Figure 3 suggests that it may be preferable to use $\widetilde{\text{CI}}_{\alpha,n}$ in samples of small and moderate size.

6 Conclusion

This paper shows that the regularization approach provides a fast way to construct both pointwise and uniform confidence sets for $\theta_1$ that can be shorter than those in the existing literature. Moreover, my CIs remain valid in some situations where the existing procedures cannot be used. Monte Carlo simulations show that the proposed CIs have good finite sample coverage properties. The computational benefits of the new approach are particularly prominent if the dimension of $\theta$
is large. My approach is attractive in applications like linear model with interval-valued outcome variable and a large number of regressors.

The analysis can be extended in a number of ways. First, the error bounds on the higher order terms can provide a theory for the optimal choice of the tuning parameters. Second, one can potentially allow for overidentifying moment conditions using regularization of the dual variables $\Lambda (\mu; P)$. Third, one can make the coverage probability of $\widehat{CI}_{\alpha,n}$ exact for any DGP with unique optimum using correction $\| \hat{\theta} (\mu_n; P_n) \|^2$ instead of $\beta (P_n)$.

References


MANSKI, C. F. AND E. TAMER (2002): “Inference on regressions with interval data on a regressor or outcome,” *Econometrica*, 70, 519–546. 2


Appendix

7 Proofs

7.1 Outline

7.1.1 Plan of the Lemmas

Lemma 1 proves that Assumption 4 is equivalent to Assumption 6. Linear Independence constraint Qualification is a necessary and sufficient condition for uniqueness of Lagrange multipliers in programs (3), (9), and (12).

Lemma 2 provides a representation of the minima to (3) and (9) as a solution to a KKT system. The KKT representation is used by the numerical software that finds the minima as a solution to the conditions. Lemma 2 also has a bound on the norm of the Lagrange multipliers in terms of singular values of the active constraints.

Lemma 3 shows that the closure of \( \mathcal{P} \) is compact and satisfies the conditions 1-4 of Definition 2. This Lemma is necessary to establish continuity of the solutions in both in \( \mu \) and \( P \in \mathcal{P} \) in Lemma 5. These continuity properties are crucial for Lemma 12 and 10.

Lemma 4 shows that the Lagrange multipliers have a uniformly bounded norm for \( P \in \overline{\mathcal{P}} \). The bound only depends on the maximal norm of \( \theta \) and \( \eta \) from Definition 2. This bound is used in Lemmas 5 and 12.

In Lemma 5 I show that the solution to the KKT system can be represented as a solution to a least squares problem. It allows me to use the Maximum theorem to establish continuity of Lagrange multipliers for any \( \mu \geq 0 \) and \( P \in \mathcal{P} \) and the hemicontinuity of the argmin.

Lemma 6 uses Lemma 5 to show that any point in \( \mathcal{P} \) satisfies the condition 5 of Definition 2. This result is important for Lemma 12 to avoid division by a zero variance.

Lemma 7 establishes that the change in the solution to the regularized program is linear in size of the perturbation. This result gives directional derivatives for any change in the coefficients of the constraints and \( \mu \). The perturbation in \( \mu \) corresponding to \( \alpha_\delta \neq 0 \) is used in Lemma 9. The perturbation with \( \delta (\theta) \neq 0 \) is used in the Delta–method to derive the asymptotic Gaussian
distribution of the estimator.

Lemma 8 provides explicit bounds on the error due to linear approximation. This result is important to prove asymptotic Gaussianity of the bound estimator and consistency of the variance estimator in [12].

Lemma 9 is crucial to show that the bias correction is exact for any fixed DGP.

Lemma 10 shows that the bias $e_j^{\prime} \hat{\theta} (\mu; P) - v (P)$ is zero for $\mu < \mu_1$. This result is important in Theorem 2, since limit of $\hat{\theta} (\mu_n; P_n)$ depends on the sequence $P_n$.

The prerequisite for the Delta method, Uniform LLN and FCLT, are proven in Lemma 11. They follow immediately from the uniform bound on the moments and linearity in $\theta$ on a compact set $\Theta$.

Lemma 12 puts uses differentiability of the value function and the uniform bound on the error to derive uniform convergence to a Gaussian distribution.

Proof of Theorem 1 uses Lemmas 9 and 12. Theorem 2 uses Lemmas 10 and 12.

7.2 Lemmas

7.2.1 Constraint qualification

Consider any distribution $P$ with support on $\mathbb{R}^{(k-2d)\times(d+1)}$ such that $\mathbb{E}_P W$ exist. Let $I^a (\theta; P) \subset \{1, ..., k\}$ be the set of indices of moment equality and inequality constraints active at $\theta$, i.e. all $j$ s.t.

$$e_j^{\prime} (\mathbb{E}_P X \theta + \mathbb{E}_P w) = 0.$$ 

$I^a (\theta; P)$ can be empty. Similarly,

$$I^0 (\mu; P) \triangleq \{ j = p + 1, ..., k | e_j^{\prime} \mathbb{E}_P W (1, \theta')' = 0, e_j^{\prime} \lambda (\mu; P) = 0 \},$$

$$I^+ (\mu; P) \triangleq \{1, ..., p\} \cup \{ j = p + 1, ..., k | e_j^{\prime} \lambda (\mu; P) > 0 \},$$

$$I^- (\mu; P) \triangleq \{ j = p + 1, ..., k | e_j^{\prime} \mathbb{E}_P W (1, \theta')' > 0 \},$$

$$I^a (\mu; P) \triangleq I^0 (\mu; P) \cup I^+ (\mu; P).$$
As before, I use symbols $I^a(\boldsymbol{\theta}; P)$, $I^a(\mu; P)$ etc to denote the projectors on the coordinates with the corresponding indices.

**Assumption 6** (Linear Independence Constraint Qualification (LICQ)). The matrix $I^a(\boldsymbol{\theta}; P) \mathbb{E}_P \mathbf{X}$ has full row rank for any $\boldsymbol{\theta} \in \Theta(P)$.

**Lemma 1** (Sufficient condition for LICQ). Assumption 4 implies Assumption 6. In addition, it implies that $p \leq d$. Assumption 6 implies Assumption 4.1 and that for any $\boldsymbol{\theta} \in \Theta(P)$

$$
rk [I^a(\boldsymbol{\theta}; P) \mathbb{E}_P (\mathbf{X}, -\mathbf{w})] = |I^a(\boldsymbol{\theta}; P)|.
$$

*Proof.* Assumption 4.1 implies that $I^a(\boldsymbol{\theta}; P)$ has at most $d$ elements at any $\boldsymbol{\theta} \in \Theta(P)$. The equality constraints are active at any point in $\Theta(P)$, so $p \leq d$.

Consider any point $\boldsymbol{\theta} \in \Theta(P)$. By the Rouché–Capelli theorem, the matrices $I^a(\boldsymbol{\theta}; P) \mathbb{E}_P \mathbf{X}$ and $\mathbf{A} = I^a(\boldsymbol{\theta}; P) \mathbb{E}_P (\mathbf{X}, -\mathbf{w})$ have the same rank. The set $I^{ineq}$ includes (1), so $k \geq 2d \geq d + 1$. It implies that the matrix $\mathbf{A} \in \mathbb{R}^{k \times (d+1)}$ is a submatrix to some full rank square matrix $(I^{eq'}, \mathcal{I})' \mathbb{E}_P (\mathbf{X}, -\mathbf{w})$ with $\mathcal{I} \subset I^{ineq}$. By Assumption 4.2, $rk (I^{eq'}, \mathcal{I})' \mathbb{E}_P (\mathbf{X}, -\mathbf{w}) = d$, so $\mathbf{A}$ has full rank which is also equal to $rk (I^a(\boldsymbol{\theta}; P) \mathbb{E}_P \mathbf{X})$. This result proves Assumption 6.

Suppose that Assumption 6 holds. It immediately implies Assumption 4.1. Consider any point $\boldsymbol{\theta} \in \Theta(P)$ such that $|I^a(\boldsymbol{\theta}; P)| = d$. By definition of $I^a(\boldsymbol{\theta}; P)$,

$$
I^a(\boldsymbol{\theta}; P) (\mathbb{E}_P \mathbf{X}\boldsymbol{\theta} + \mathbb{E}_P \mathbf{w}) = 0.
$$

By the Rouché–Capelli theorem and linear independence of $I^a(\boldsymbol{\theta}; P) \mathbb{E}_P \mathbf{X}$,

$$
rk (I^a(\boldsymbol{\theta}; P) \mathbb{E}_P (\mathbf{X}, -\mathbf{w})) = rk (I^a(\boldsymbol{\theta}; P) \mathbb{E}_P \mathbf{X}) = |I^a(\boldsymbol{\theta}; P)|.
$$

□
7.2.2 Topological properties

Lemma 2 (Equivalent characterizations of the optimal solution for \( \mu > 0 \)). Under Assumption 3 for any \( \mu \geq 0 \) any minimizer \( \theta = \vartheta (\mu; P) \in \Theta (P) \) for Program (9) is a solution to the corresponding Karush–Kuhn–Tucker (KKT) optimality conditions for some finite \( \lambda \in \mathbb{R}^k \),

\[
\begin{align*}
(e_1 + 2 \mu \theta)' &= -\lambda' \mathbb{E}_P X, \\
\lambda' \mathbb{E}_P W (\theta', 1)' &= -\lambda' \mathbb{E}_P X, \\
e_j' \mathbb{E}_P W (\theta', 1)' &= 0, \text{ for } j \in I^{eq}, \\
e_j' \mathbb{E}_P W (\theta', 1)' &\leq 0, \lambda_j \geq 0, \text{ for } j \in I^{ineq}, \\
\lambda' \mathbb{E}_P W (\theta', 1)' &= 0.
\end{align*}
\]  

(23)

If in addition, Assumption 4 is satisfied, then for any \( \mu \geq 0 \) the set of multipliers \( \lambda = \Lambda (\mu; P) \) in (23) is a singleton. The norm of \( \Lambda (\mu; P) \) has the following bound:

\[
\| \Lambda (\mu; P) \| \leq C_\Theta (\mu) \eta^{-1} (P),
\]

(24)

\[
\eta (P) \Delta \min_{\theta \in \partial \Theta (P)} \text{eig } I^a (\theta; P) \mathbb{E}_P \mathbb{E}_P X \mathbb{E}_P X I^a (\theta; P)' > 0,
\]

\[
C_\Theta (\mu) \Delta \max_{\theta \in \Theta} \| e_1 + 2 \mu \theta \|,
\]

where \( \partial \Theta (P) \) is the boundary of \( \Theta (P) \). Moreover, if \( \mu > 0 \), then (23) has a unique solution \((\vartheta (\mu; P), \Lambda (\mu; P))\).

Proof. Step 1. By Assumption 3 \( \Theta (P) \subset \Theta \) is non–empty and closed, so the global optima for Program (9) exist. Program (9) is convex for any \( \mu \geq 0 \), i.e. the objective function is convex, the constraints are affine. Assumption 3 implies Slater’s condition. Since the Program (9) is convex, any global optimum \( \vartheta (\mu; P) \) of Program (9) satisfies (23) for some finite vector of Lagrange multipliers \( \lambda \) (maybe non–unique) (see p.244 in Boyd and Vandenberghe (2004)).

Step 2. By Lemma 1 for any \( \theta \in \Theta (P) \) the matrix \( I^a (\theta; P) \mathbb{E}_P X \) has full rank, so \( \eta (P) > 0 \). For any \( \theta \in \text{int } (\Theta (P)) \cap \vartheta (\mu; P) \) the set of active constraints is empty, \( I^a (\theta; P) = \emptyset \). So \( \Lambda (\mu; P) = 0 \) and (24) trivially holds.

Consider any point \( \theta \in \partial (\Theta (P)) \triangleq \Theta (P) \setminus \text{int } (\Theta (P)) \). For such \( \theta \) there exists some element
in $I^a(\theta; P)$. As a result, for any $\theta \in \partial (\Theta (P))$ the spectral norm of the matrix

$$\mathbb{E}_P X' I^a(\theta; P)' \dagger = I^a(\theta; P) \mathbb{E}_P X' \mathbb{E}_P X' I^a(\theta; P)' (I^a(\theta; P) \mathbb{E}_P X).$$

is bounded by $\eta^{-1}(P)$. By Step 1 for any $\mu \geq 0$ and any $\theta \in \partial (\Theta (P)) \cap \vartheta(\mu; P)$ the vector of Lagrange multiplier $\lambda(\mu; P)$ satisfies

$$- \mathbb{E}_P X' I^a(\mu; P)' \dagger (e_1 + 2\mu \theta) = I^a(\mu; P) \lambda(\mu; P),$$

$$\lambda(\mu; P) = I^a(\mu; P)' I^a(\mu; P) \lambda(\mu; P).$$

Correspondingly, $\lambda(\mu; P)$ is unique for any $\mu \geq 0$ and (24) holds.

**Step 3.** The Hessian of $L(\theta, \lambda; \mu; P)$ with respect to $\theta$ at $(\vartheta(\mu; P), \lambda(\mu; P))$ 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 $\vartheta(\mu; P)$, i.e. $\exists \varepsilon > 0$ and $c > 0$ s.t. for $\forall \theta \in \Theta (P)$ s.t. $\|\theta - \vartheta(\mu; P)\| < \varepsilon$ the following inequality holds

$$\theta_1 + \mu \|\theta\|^2 \geq e_1' \vartheta(\mu; P) + \mu \|\vartheta(\mu; P)\|^2 + c \|\theta - \vartheta(\mu; P)\|^2.$$

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

**Definition 3.** Let $BL_1$ be the set of all real functions $f$ on $\mathbb{R}^{k \times (d+1)}$ with a norm

$$\|f\|_{1, \infty} \triangleq \sup_{x \neq y; x, y \in \mathbb{R}^{k \times (d+1)}} \frac{|f(x) - f(y)|}{\|x - y\|} \leq 1.$$

The bounded Lipschitz metric for $P_1, P_2 \in \mathcal{P}$ is

$$d_{BL}(P_1, P_2) = \sup_{f \in BL_1} \int f dP_1 - \int f dP_2.$$
Metric $d_{BL}(P_1, P_2)$ generates the topology of weak convergence on $\bar{\mathcal{P}}$ (Theorem 1.12.4 van der Vaart and Wellner (1996)).

**Lemma 3** (Topological properties of $\mathcal{P}$). The closure of $\mathcal{P}$, $\bar{\mathcal{P}}$, is compact in weak topology and satisfies conditions 1-4 of Definition 2. Moreover, for any convergent sequence $\{P_n\}_{n=1}^{\infty} \subset \mathcal{P}$ and any combination of indices $r, \ell, j, m$

$$E_{P_n} |W_{r,\ell} W_{j,m}| \to E_P |W_{r,\ell} W_{j,m}|, \quad (25)$$

$$E_{P_n} W \to E_P W, \quad (26)$$

where $P_n \Rightarrow P$.

**Proof.** Structure of the proof. First I show that $\mathcal{P}$ is uniformly tight so compactness of its closure, $\bar{\mathcal{P}}$, follows from the Prokhorov theorem. Then I prove conditions 1-4 in Definition 2 for any limiting point of $\mathcal{P}$ using the uniform integrability properties of $W$ and the corresponding continuity properties of the functions involved in Definition 2.

**Step 1.** By Definition 2 and Jensen’s inequality, for any $P \in \mathcal{P}$

$$E_P \|W\|^2 \leq k (d + 1) \bar{M}^{\frac{2}{d+2}}.$$

Consider a compact set for any positive $\xi$,

$$\mathcal{K}^\xi = \left\{ \|W\|^2 \leq \xi k (d + 1) \bar{M}^{\frac{2}{d+2}} \bigg| W \in \mathbb{R}^{k \times (d+1)} \right\}.$$

By Chebyshev’s inequality for any $P \in \mathcal{P}$,

$$P \ W \in \mathcal{K}^\xi > 1 - \frac{1}{\xi}. \quad (27)$$

The set $\mathcal{K}^\xi$ does not depend on $P \in \mathcal{P}$, so $\mathcal{P}$ is uniformly tight. By the Prohorov’s Theorem (see Theorem 1.3.9 in van der Vaart and Wellner (1996)) the uniformly tight set $\mathcal{P}$ is a precompact (or relatively compact) in weak topology. Since the $\mathcal{P}$ precompact, its closure $\bar{\mathcal{P}}$ is a sequential
compact in weak topology.

By Theorem 1.12.4 from van der Vaart and Wellner (1996) the weak convergence on $P$ is metrizable by $d_{BL}$. For a metric spaces sequential compactness is equivalent to compactness.

**Step 2.** Consider any weakly converging sequence $\{P_n\}_{n=1}^{\infty} \subset P$ such that $P_n \rightharpoonup P$. By Schwarz and Jensen’s inequalities, for any combination of indices $r, \ell, j, m$

$$
\mathbb{E}_{P_n} |W_{r, \ell} W_{j, m}|^{1+\varepsilon/2} \leq \mathbb{E}_{P_n} |W_{r, \ell}|^{2+2\varepsilon} \mathbb{E}_{P_n} |W_{j, m}|^{2+2\varepsilon} \frac{1}{2} \leq M.
$$

So for any $P_n \in P$ and any combination of indices $r, \ell, j, m$ the random variables $|W_{r, \ell}|$ and $|W_{r, \ell} W_{j, m}|$ have correspondingly finite $1 + \varepsilon/2$ and $2 + \varepsilon$ moments. The bound on the moments is independent of $P_n \in P$, so these random variables with measures $\{P_n\}$ are uniformly integrable (see p. 67 in Kallenberg (2006)). By Lemma 4.11 in Kallenberg (2006), the uniform integrability implies

$$
\mathbb{E}_{P_n} W_{r, \ell, j, m} \to \mathbb{E}_P W_{r, \ell, j, m},
$$

$$
\mathbb{E}_{P_n} W \to \mathbb{E}_P W,
$$

$$
\liminf_{n \to \infty} \mathbb{E}_{P_n} |W_{r, \ell}|^{2+\varepsilon} \geq \mathbb{E}_P |W_{r, \ell}|^{2+\varepsilon}.
$$

The last inequality implies that any $P \in \bar{P}$ satisfies Condition 1 in Definition 2.

**Step 3.** By Condition 2 of Definition 2 $\Theta(P_n)$ is not empty for every $n \in \mathbb{N}$. Pick a point $\vartheta_n \in \Theta(P_n)$ for every $n \in \mathbb{N}$. Since every $\vartheta_n$ belongs to the compact set $\Theta$, there exist a convergent subsequence $\{\vartheta_{n_\ell}\}_{\ell=1}^{\infty}$. Let $\vartheta^* \triangleq \lim_{\ell \to \infty} \vartheta_{n_\ell}$. By (26), $\vartheta^* \in \Theta(P)$. So Condition 2 of Definition 2 is satisfied for any $P \in \bar{P}$.

**Step 4.** By Step 1, $\bar{P}$ is a metric space. By (26), the set $\Theta(P)$ is a compact–valued continuous correspondence at any $P \in \bar{P}$. Analogously, the objective function of Program (16) is continuous in $P \in \bar{P}$. It implies that Program (16) satisfies the conditions of the Maximum Theorem (see the corresponding section in Ok (2007)). As a result, the value of Program (16) is a continuous function of $P \in \bar{P}$. By this continuity property, any $P \in \bar{P}$ satisfied Condition 3 of Definition 2.

By continuity of $\text{eig}(\cdot)$, any $P \in \bar{P}$ also satisfies Condition 4 of Definition 2.

\[\square\]
Lemma 4 (Uniform bound on Lagrange multipliers). For any $P \in \mathcal{P}$

$$
\eta(P) = \min_{\theta \in \partial \Theta(P)} \text{eig} \ I^a(\theta; P) E_P X E_P X' I^a(\theta; P)' \geq \eta.
$$

Correspondingly,

$$
\|\lambda(\mu; P)\| \leq \eta^{-1} C_\Theta(\mu),
$$

$$
C_\Theta(\mu) \triangleq \max_{\theta \in \Theta} \|e_1 + 2\mu \theta\|.
$$

Proof. Lemma 3 and Conditions 3-4 of Definition 2 imply that any $P \in \mathcal{P}$ satisfy Assumption (4). By Lemma (1), any $P \in \mathcal{P}$ satisfies Assumption (6). Set $I^a(\theta; P)$ is nonempty for any $\theta \in \partial \Theta(P)$. Assumption (6) implies that for any $\theta \in \partial \Theta(P)$

$$
eig I^a(\theta; P) E_P X E_P X' I^a(\theta; P)' > 0.
$$

The Rouché–Capelli theorem implies that ranks and, correspondingly, the numbers of the positive eigenvalues are equal for the following matrices

$$
E_P (X, -w)' I^a(\theta; P)' I^a(\theta; P) E_P (X, -w), 
$$

$$
E_P X I^a(\theta; P)' I^a(\theta; P) E_P X.
$$

Let this number be $\ell$. By the Courant–Fischer–Weyl min-max principle, the $\ell$-th largest eigenvalue of (28) can be represented as

$$
\tau_\ell = \max_{U: \dim U = \ell} \min_{\nu \in U \subset \mathbb{R}^{d+1}} \nu' E_P (X, -w)' I^a(\theta; P)' I^a(\theta; P) E_P (X, -w) \nu / \|\nu\|^2 ,
$$

where maximum is taken over all linear subspaces $U \subset \mathbb{R}^{d+1}$ with dimension $\ell$. Similarly, the $\ell$-th largest eigenvalue of (29) can be represented as

$$
\tilde{\tau}_\ell = \max_{U: \dim U = \ell} \min_{\nu \in U \subset \mathbb{R}^d} \nu' E_P X I^a(\theta; P)' I^a(\theta; P) E_P X \nu / \|\nu\|^2 .
$$
For any fixed $U$ the value of the minimization program in (31) is not smaller than the corresponding value in (30) since the corresponding domains are nested. It implies $\tilde{\tau} \geq \tau$.

Positive singular values do not change by matrix transposition operation. By this property of singular values the matrix $I^a(\theta; P)\mathbb{E}_P(X, -w)\mathbb{E}_P(X, -w)'I^a(\theta; P)'$ has the same set of positive eigenvalues as the matrix (28). By Condition 4 from $\tau \geq \eta^2$. Similarly, any eigenvalue of the matrix $I^a(\theta; P)\mathbb{E}_P X \mathbb{E}_P X' I^a(\theta; P)'$ is also an eigenvalue of (29). So

$$\text{eig} \quad I^a(\theta; P)\mathbb{E}_P X \mathbb{E}_P X' I^a(\theta; P)' \geq \eta,$$

and $\eta(P) \geq \eta$ for any $P \in \mathcal{P}$. The uniform bound on $\|\lambda(\mu; P)\|$ follows from Lemma 2. □

**Lemma 5** (Continuity properties of the solution). *Under Assumption 3 and 4, the solution to (23), $\vartheta(\mu; P)$, is upper hemicontinuous in $\mu$ for $\mu \geq 0$ and continuous for $\mu > 0$; the functions $\lambda(\mu; P)$ and $\upsilon(\mu; P)$ are continuous in $\mu \geq 0$. Moreover, the function $\vartheta(\mu; P)$ is upper hemicontinuous over the product space $\mathbb{R}_+ \times \bar{\mathcal{P}}$ and continuous on $(0, \infty) \times \bar{\mathcal{P}}$; $\lambda(\mu; P)$ and $\upsilon(\mu; P)$ are continuous functions of $(\mu, P) \in \mathbb{R}_+ \times \bar{\mathcal{P}}$.*

**Proof.** STRUCTURE OF THE PROOF. First, I show that solution to (23) can be represented as a solution to a continuous optimization problem (34). Then I apply the Maximum theorem to (34) to establish hemicontinuity of $\vartheta(\cdot; P)$ on $\mathbb{R}_+$ for a fixed $P$. Second, I repeat the argument for the product topology on $\mathbb{R}_+ \times \bar{\mathcal{P}}$. Finally, I use uniqueness of $\vartheta(\cdot; P)$ (and $\lambda(\mu; P)$) to show their continuity for any $\mu > 0$ ($\geq 0$).

**STEP 1.** The function $\phi(a, b) \triangleq \sqrt{a^2 + b^2} + a - b$ has the following property (see Fischer (1992) for details)

$$\phi(a, b) = 0 \text{ if and only if } a \leq 0, b \geq 0, ab = 0. \quad (32)$$

By Property (32), (23) is equivalent to

$$\begin{cases} 
L(\theta, \lambda; \mu; P) = 0, \\
e'_j\mathbb{E}_P W(\theta', 1)' = 0, \text{ for } j \in I^q, \\
\phi \ e'_j\mathbb{E}_P W(\theta', 1)', \lambda_j = 0, \text{ for } j \in I^{ineq},
\end{cases} \quad (33)$$

9
where $L(\theta, \lambda; \mu; P) \triangleq \theta_1 + \mu \|\theta\|^2 + \lambda^T \mathbb{E}_P W (\theta', 1)$. Solutions to (33) coincide with solutions to

$$\arg\min_{\theta \in \Theta, \lambda \in \mathbb{R}^k} \left\{ \|\partial_\theta L(\theta, \lambda; \mu; P)\|^2_2 + \sum_{j \in I^e} e_j^T \mathbb{E}_P W (\theta', 1)'^2 + \sum_{j \in I^{ineq}} \phi_j e_j^T \mathbb{E}_P W (\theta', 1)' , \lambda_j \right\}^2,$$

(34)

**Step 2.** By Assumption 3 the program (34) has a minimizer. By Lemma 2, Assumption 4 implies that $\lambda(\mu; P)$ is unique and satisfies (24) for any $\mu \geq 0$. So the solution to (34) coincides with

$$\arg\min_{\theta \in \Theta, \lambda \in \mathbb{R}^k, \|\lambda\| \leq C_\alpha (\eta^{-1}(P))} \left\{ \|\partial_\theta L(\theta, \lambda; \mu; P)\|^2_2 + \sum_{j \in I^e} e_j^T \mathbb{E}_P W (\theta', 1)'^2 + \sum_{j \in I^{ineq}} \phi_j e_j^T \mathbb{E}_P W (\theta', 1)' , \lambda_j \right\}^2.$$

The objective function of this program is a continuous function of $\mu$ and the domain is a compact valued continuous correspondence in $\mu$. By the Maximum Theorem (See Ok (2007)) $\xi (\mu; P) = (\vartheta (\mu; P), \lambda (\mu; P))$ is upper hemicontinuous function of $\mu \geq 0$.

Function $v(\mu; P) = e_1^T \vartheta (\mu; P) + \mu \|\vartheta (\mu; P)\|^2$ is a composition of upper hemicontinuous functions and hence, by Theorem VI.2.1’ from Berge (1963), is upper hemicontinuous in $\mu \in \mathbb{R}_+$. Since by definition $v(\mu; P)$ is a single–valued function, upper hemicontinuity implies continuity in $\mu \geq 0$ for any fixed $P$.

**Step 3.** The set $\check{P}$ is compact in weak topology by Lemma 3. By Theorem V.2.2 from Berge (1963) the set $[0, +\infty) \times \check{P}$ is a metric space with a 2-product metric

$$d_2 ((\mu, P), (\nu, Q)) = (\mu - \nu)^2 + d_{BL} (P, Q)^2.$$

By Lemma 4, $\eta(P)$ in Step 1 can be replaced by $\eta$. By 25 and 26 the objective function of Program (34) is continuous in $(\mu, P) \in [0, +\infty) \times \check{P}$. By Maximum Theorem $\xi (\mu; P)$ is upper hemicontinuous function of $(\mu, P) \in [0, +\infty) \times \check{P}$. Correspondingly, $v(\mu; P)$ is continuous for any $(\mu, P) \in \mathbb{R}_+ \times \check{P}$.

**Step 4.** By Lemma 2 Program 9 has a unique solution $\vartheta (\mu; P)$ for any $\mu \in (0, \infty)$ so $\vartheta (\mu; P)$ is a continuous function of $\mu > 0$ and a continuous function of $(\mu, P) \in (0, \infty) \times \check{P}$.

By Lemma 2 $\lambda (\mu; P)$ is single–valued. It implies that $\lambda (\mu; P)$ is a continuous function of
\( \mu \geq 0 \) for any fixed \( P \) and a continuous function of \( (\mu, P) \) over \( \mathbb{R}_+ \times \bar{P} \).

**Lemma 6.** The class \( \mathcal{P} \) is compact in weak topology.

**Proof.** By Lemma 5, the function \( \mathbf{X} (0; \cdot) \) is continuous in weak topology on \( \bar{P} \), i.e.

\[
\mathbf{X}'(P_n) \to \mathbf{X}'(P). \tag{35}
\]

Analogously, \( \mathbf{\theta} (\mu; P) \) is upper hemicontinuous in \( (\mu, P) \in [0, +\infty) \times \bar{P} \). So for any convergent sequence \( \{ (\mu_n, P_n) \}_{n=1}^{\infty} \subset [0, +\infty) \times \mathcal{P} \) with \( \mu_n \to 0 \) and \( P_n \to P \)

\[
\lim_{n \to \infty} \mathbf{\theta} (\mu_n; P_n) \in \mathbf{\theta} (0; P). \tag{36}
\]

By Condition 5 in Definition 2, for any \( \theta \in \mathbf{\theta} (0; P_n) \)

\[
\text{Var} \ \mathbf{X}'(P_n) \mathbf{W} (\theta, 1)' \geq \sigma_0^2.
\]

By (35), (36), (25) and (26) any \( P \in \bar{P} \) also satisfies condition 4 in Definition 2, so \( \bar{P} = \mathcal{P} \). The class \( \mathcal{P} \) is compact since it is closed and precompact. \( \square \)

### 7.2.3 Smoothness properties

**Lemma 7** (Local linear representation). Suppose that Assumptions (3) and (4) hold. Then for any \( \alpha_\delta \in \mathbb{R}, t \geq 0, \mu > 0, \) and affine vector function \( \delta (\theta) = \tilde{X}\theta + \tilde{w} \) with \( \tilde{W} = \tilde{X}, \tilde{w} \in \mathbb{R}^{k \times (d+1)} \) the program

\[
\begin{align*}
& \min_{\theta \in \Theta} \quad e_1' \theta + \mu \| \theta \|^2 + \alpha_\delta t \| \theta \|^2, \\
& \text{s.t.} \quad \begin{cases}
& e_j \ P \mathbf{W} (\theta', 1)' + t \delta (\theta) = 0, \text{ for } j \in \Gamma_{eq}, \\
& e_j \ P \mathbf{W} (\theta', 1)' + t \delta (\theta) \leq 0, \text{ for } j \in \Gamma_{ineq},
\end{cases}
\end{align*}
\tag{37}
\]

11
has a unique solution $\vartheta_\delta (\mu, t; P)$ and the corresponding Lagrange multipliers $\lambda_\delta (\mu, t; P)$, which can be represented for $0 \leq t < T$ as

$$
\vartheta_\delta (\mu, t; P) = \vartheta (\mu; P) + t \dot{\vartheta}_\delta (\mu; P),
\lambda_\delta (\mu, t; P) = \lambda (\mu; P) + t \dot{\lambda}_\delta (\mu; P),
$$

where

$$
\dot{\vartheta}_\delta (\mu; P) = \begin{pmatrix} (2\mu)^{-1} P_\delta (\mu; P), & A^\dagger_\delta (\mu; P) \end{pmatrix} \psi (\delta),
$$

$$
\dot{\lambda}_\delta (\mu; P) = \begin{pmatrix} A^\dagger_\delta (\mu; P), & -2\mu (A_\delta (\mu; P) A'_\delta (\mu; P))^{-1} \end{pmatrix} \psi (\delta),
$$

$$
\psi (\delta) \triangleq - \begin{pmatrix} \tilde{X}' \Lambda (\mu; P) \\ \mathbf{I}^\dagger_\delta (\mu; P) \delta (\vartheta (\mu; P)) \end{pmatrix} - 2\alpha_\delta (\vartheta' (\mu; P); 0)',
$$

$$
\mathbf{I}^\dagger_\delta (\mu; P) \triangleq \mathbf{I}^\dagger (\mu; P) \cup \{ j \in \mathbf{I}^0 (\mu; P) | e_j' \dot{\lambda}_\delta (\mu; P) > 0 \},
$$

$$
A_\delta (\mu; P) \triangleq \mathbf{I}^\dagger_\delta (\mu; P) \mathbf{E}_P X,
$$

$$
A^\dagger_\delta (\mu; P) \triangleq A'_\delta (\mu; P) (A_\delta (\mu; P) A'_\delta (\mu; P))^{-1},
$$

$$
P_\delta (\mu; P) \triangleq \mathbf{I}_d - A^\dagger_\delta (\mu; P) A_\delta (\mu; P).
$$

**Proof.** The proof is the application of generalized implicit function theorem to the KKT conditions for Program (37)

$$
\begin{cases}
\mathbf{e}_1 + 2 (\mu + \alpha_\delta t) \theta = -\lambda' \mathbf{E}_P X + t \tilde{X}, \\
\mathbf{e}_j' \mathbf{E}_P W (\theta', 1) = 0, \text{ for } j = 1, ..., p, \\
\mathbf{e}_j' \mathbf{E}_P W (\theta', 1) \leq 0, \mathbf{X}' e_j \geq 0, \text{ for } j = p + 1, ..., k, \\
\mathbf{X}' \mathbf{E}_P W (\theta', 1) = 0.
\end{cases}
$$

(38)

By Lemma 2 if $t = 0$ Program (37) has a unique solution $\vartheta (\mu; P)$ and a unique vector of Lagrange multipliers $\lambda (\mu; P)$ corresponding to that point. Proposition 5.38 from Bonnans and Shapiro (2000) states that a critical point $\xi' (\mu; P) = (\vartheta' (\mu; P), \lambda' (\mu; P))$ is strongly regular iff it satisfies LICQ (Assumption (6)) and SOSC, which is the case here. By Theorem 5.60 from
Bonanns and Shapiro (2000)\(^{31}\) there exist a unique solution \((h', q') = \hat{\vartheta}'(\mu; P), \hat{\lambda}'(\mu; P)\) to the following system of generalized equations

\[
\begin{cases}
2\mu h' I_d + q' \mathbb{E}_P X = -\mathbf{X}'(\mu; P) \tilde{X} - 2\alpha_\delta \vartheta'(\mu; P), \\
e_j' (\mathbb{E}_P X h + \delta (\vartheta(\mu; P))) = 0 \text{ for } j \in I^+(\mu; P), \\
e_j' (\mathbb{E}_P X h + \delta (\vartheta(\mu; P))) \leq 0, e_j' q \geq 0 \text{ for } j \in I^0(\mu; P), \\
e_j' q = 0 \text{ for } j \in I^-(\mu; P), \\
q' (\mathbb{E}_P X h + \delta (\vartheta(\mu; P))) = 0.
\end{cases}
\]

Since \(I_\delta^+(\mu; P)\) is well defined, this system can be represented in a matrix form:

\[
\begin{pmatrix}
2\mu I_d & A_\delta'(\mu; P) \\
A_\delta(\mu; P) & 0
\end{pmatrix}
\begin{pmatrix}
\hat{\vartheta}_\delta(\mu; P) \\
I_\delta^+(\mu; P) \hat{\lambda}_\delta(\mu; P)
\end{pmatrix} = \psi(\delta),
\]

\[(40)\]

\[
- \begin{pmatrix}
\tilde{X}' \lambda(\mu; P) \\
I_\delta^+(\mu; P) \delta (\vartheta(\mu; P))
\end{pmatrix}
- 2\alpha_\delta (\vartheta'(\mu; P); 0)' \triangleq \psi(\delta).
\]

\[(41)\]

In addition to that, \(\hat{\lambda}_\delta(\mu; P) = I_\delta^+(\mu; P)' I_\delta^+(\mu; P) \hat{\lambda}_\delta(\mu; P)\).

Since by the strong regularity property the solution \((h, \lambda)\) is unique, this matrix is invertible,

\[
\begin{pmatrix}
2\mu I_d & A_\delta'(\mu; P) \\
A_\delta(\mu; P) & 0
\end{pmatrix}^{-1} = \begin{pmatrix}
(2\mu)^{-1} P_\delta(\mu; P) & A_\delta'(\mu; P) \\
A_\delta'(\mu; P) & -2\mu (A_\delta(\mu; P) A_\delta'(\mu; P))^{-1}
\end{pmatrix}.
\]

So the solution can be represented as

\[
\hat{\vartheta}_\delta(\mu; P) = (2\mu)^{-1} P_\delta(\mu; P), \quad A_\delta'(\mu; P) \quad \psi(\delta)
\]

\[
I_\delta^+(\mu; P) \hat{\lambda}_\delta(\mu; P) = A_\delta'(\mu; P), \quad -2\mu (A_\delta(\mu; P) A_\delta'(\mu; P))^{-1} \quad \psi(\delta)
\]

By Theorem 5.60 from Bonnans and Shapiro (2000) \(\hat{\vartheta}_\delta(\mu, t; P)\) and \(\hat{\lambda}_\delta(\mu, t; P)\) are analytic in

\(^{31}\)The original proof of the result was discovered by Bonnans and Sulem (1995).
In some neighborhood \([0, T(\mu, \delta, P)]\), i.e. they can be represented as Taylor series. Since the higher order derivatives of every constraint function and the objective function with respect to \(t\) are zero, the higher order directional derivatives of \(\vartheta_\delta(\mu, t; P)\) and \(\lambda_\delta(\mu, t; P)\) are equal to zero at \(t = 0\). So the Taylor series representation is

\[
\vartheta_\delta(\mu, t; P) = \vartheta(\mu; P) + t \dot{\vartheta}_\delta(\mu; P), \\
\lambda_\delta(\mu, t; P) = \lambda(\mu; P) + t \dot{\lambda}_\delta(\mu; P).
\]

**Definition 4.** Let

\[
\begin{align*}
I^+_\delta(\mu; P) & \triangleq I^+(\mu; P) \cup \{ j \in I^0(\mu; P) | e_j' \dot{\lambda}_\delta(\mu; P) > 0 \}, \\
I^-_\delta(\mu; P) & \triangleq I^-(\mu; P) \cup \{ j \in I^0(\mu; P) | e_j' \dot{\lambda}_\delta(\mu; P) < 0 \}
\end{align*}
\]

**Lemma 8** (Uniform error bounds). Suppose that Assumptions 3 and 4 hold. Then the following inequalities hold for the solution of Program 37 with \(\alpha_\delta = 0\) and any \(\mu \in (0, 1]::

\[
\begin{align*}
&v_\delta(\mu, t; P) - v(\mu; P) - t \lambda'(\mu; P) \delta(\vartheta(\mu; P)) \leq t^2 C_v(\mu, \delta), \\
&\lambda_\delta(\mu, t; P) - \lambda(\mu; P) \leq \eta(-4) \left(1 + \beta \right) \bar{W}^2 + \mu^{-1} \lambda'(\mu; P) \bar{X}^2, \\
&\Phi(\delta_\delta(\mu, t; P)) \lambda_\delta(\mu, t; P) - \lambda'(\mu; P) \Phi(\vartheta(\mu; P)) \lambda(\mu; P) \leq t C_\Phi(\mu, \delta),
\end{align*}
\]

where

\[
\begin{align*}
\eta(P) & = \inf_{\theta \in \Theta(P)} \text{eig} \left( \Gamma^a(\theta; P) \mathbb{E}_P X \mathbb{E}_P X^\top(\theta; P) \right)^t > 0, \\
C_v(\mu, \delta) & = \mu \eta^{-4}(P) \left(1 + \beta \right) \bar{W}^2 + \mu^{-1} \lambda'(\mu; P) \bar{X}^2, \\
C_\Phi & = \eta^{-2}(P) \sqrt{1 + \beta \lambda} \sqrt{\mu^{-1} C_v(\mu, \delta)} + \eta^{-4}(P) \bar{W} + 2 \left(1 + \beta \right)^{3/2} \eta^{-3}(P) \bar{W}^2, \\
\beta & = \max \{1, \max_{\theta \in \Theta} \|\theta\|^2\}
\end{align*}
\]
\[ L_{\Phi} \triangleq \max_{\theta, \nu \in \Theta} \left\| \Phi(\nu) - \Phi(\theta) \right\|_2 \]

\[ M_{\Phi} \triangleq \max_{\theta \in \Theta} \left\| \Phi(\theta) \right\|_2 \]

**Proof.** Step 1. By Lemma \([2]\),

\[(e_1 + 2\mu \partial(\mu; P))' = -X'(\mu; P)E_XP\]

Since by definition \(I^+ (\mu; P) \subseteq I^+_\delta (\mu; P)\) for any affine \(\delta (\theta)\),

\[ X'(\mu; P) = X'(\mu; P)I^+_\delta (\mu; P)' I^+_\delta (\mu; P) \]

So \[(e_1 + 2\mu \partial (\mu; P))' P_\delta (\mu; P) = -X'(\mu; P)E_XP_\delta (\mu; P) \]

\[ = -X'(\mu; P)I^+_\delta (\mu; P)' A_\delta (\mu; P) P_\delta (\mu; P) = -X'(\mu; P) 0 = 0, \tag{42} \]

where \(P_\delta (\mu; P)\) is defined in Lemma \([7]\). In addition,

\[ (e_1 + 2\mu \partial (\mu; P))' A^+_\delta (\mu; P) = -X'(\mu; P)I^+_\delta (\mu; P)' A_\delta (\mu; P) \]

\[ = -X'(\mu; P)I^+_\delta (\mu; P)' I^+_\delta (\mu; P)' I^+_\delta (\mu; P)' = -X'(\mu; P). \tag{43} \]

By Lemma \([7]\) using (42) and (44), the value function \(v_\delta (\mu, t; P)\) can be represented a

\[ v_\delta (\mu, t; P) = e_1'A_\delta (\mu, t; P) + \mu \left\| \partial_\delta (\mu, t; P) \right\|^2 = \]

\[ = v(\mu; P) + tX'(\mu; P) \delta (\partial (\mu; P)) + \mu t^2 \partial_\delta (\mu; P)^2, \]

where \(\partial_\delta (\mu; P)^2\) is equal to

\[ (2\mu)^{-2} X'(\mu; P)X' P_\delta (\mu; P) X'(\mu; P)X + \]

\[ \bar{X} \partial (\mu; P) + \bar{w} I^+_\delta (\mu; P)' (A_\delta (\mu; P) A_\delta (\mu; P))^{-1} I^+_\delta (\mu; P) \bar{X} \partial (\mu; P) + \bar{w} \]
By Lemma \[\text{[2]},\] I can choose
\[
\eta(P) = \min_{\theta \in \Theta(p)} \text{eig} \ I^o(\theta; P) E_p X E_p X^o(\theta; P) > 0,
\]
so that the spectral norm of the matrix \((A_\delta(\mu; P) A_\delta(\mu; P))^{-1}\) is bounded,
\[
(A_\delta(\mu; P) A_\delta(\mu; P))^{-1} \leq \eta^{-2}(P).
\]

I can obtain the uniform bounds
\[
|v_\delta(\mu, t; P) - v(\mu; P) - t \tilde{X}'(\mu; P) \delta(\bar{\theta}(\mu; P))| \leq t^2 \mu \eta^{-2}(P) \tilde{X}^2 \beta + \|\bar{w}\|^2 + \mu^{-1} \tilde{X}'(\mu; P) \tilde{X}^2
\]
that is valid for small \(t > 0\) uniformly in \(\mu \in [0, 1]\).

Let \(\beta = \max\{1, \max_{\theta \in \Theta} \|\theta\|^2\}\). The norm of the difference is bounded by
\[
\|\bar{\theta}_\delta(\mu, t; P) - \bar{\theta}(\mu; P)\|^2 = t^2 \mu \eta^{-2}(P) \tilde{X}^2 \beta + \|\bar{w}\|^2 + \mu^{-1} \tilde{X}'(\mu; P) \tilde{X}^2 \tag{45}
\]

\text{STEP 2.} Consider the difference
\[
|X'_\delta(\mu, t; P) \Phi(\bar{\theta}_\delta(\mu, t; P)) \hat{\lambda}_\delta(\mu, t; P) - \hat{\lambda}(\mu; P) \Phi(\bar{\theta}(\mu; P)) \hat{\lambda}(\mu; P)| \leq \tag{46}
\]
\[
|X'_\delta(\mu, t; P) \Phi(\bar{\theta}_\delta(\mu, t; P)) \hat{\lambda}_\delta(\mu, t; P) - \hat{\lambda}_\delta(\mu, t; P) \Phi(\bar{\theta}(\mu; P)) \hat{\lambda}_\delta(\mu, t; P)|
+ |X'_\delta(\mu, t; P) \Phi(\bar{\theta}(\mu; P)) \hat{\lambda}_\delta(\mu, t; P) - X'_\delta(\mu, t; P) \Phi(\bar{\theta}(\mu; P)) \hat{\lambda}(\mu; P)|
+ |X'_\delta(\mu, t; P) \Phi(\bar{\theta}(\mu; P)) \hat{\lambda}(\mu; P) - \hat{\lambda}(\mu; P) \Phi(\bar{\theta}(\mu; P)) \hat{\lambda}(\mu; P)| \leq 
\eta^{-2}(P) \sqrt{1 + \beta L_\theta} \|\bar{\theta}_\delta(\mu, t; P) - \bar{\theta}(\mu; P)\| + 2 \tilde{X}'(\mu; P) \hat{\lambda}_\delta(\mu; P) M_\delta t.
\]

Using \[\text{[14]},\] I can obtain the bound for \(\mu \in [0, 1]\):
\[
X'(\mu; P) \hat{\lambda}_\delta(\mu; P) \leq X'(\mu; P) I^+_\delta(\mu; P)'^{\dagger} A^\gamma_\delta(\mu; P) I^+_\delta(\mu; P) \tilde{X}' \hat{\lambda}(\mu; P)
\]
\[ + 2\mu \nabla (\mu; P) I^+ (\mu; P)' (A_\delta (\mu; P) A'_{\delta} (\mu; P))^{-1} I^+ (\mu; P) \delta (\vartheta (\mu; P)) \]
\[ \leq \tilde{W} \eta^{-3} (P) + 2\mu (1 + \beta)^{3/2} \eta^{-3} (P) \tilde{W}^2 \]

So (46) is bounded by

\[ |\lambda'_\delta (\mu, t; P) \Phi (\vartheta_{\delta} (\mu, t; P)) - \lambda' (\mu; P) \Phi (\vartheta (\mu; P)) \lambda (\mu; P)| \leq \]
\[ \mu^{-1} \sqrt{2} \eta^{-4} (P) (1 + \beta) L_\Phi \tilde{W} t \]
\[ + \eta^{-4} (P) \left( 1 + 2 (1 + \beta)^{3/2} \eta (P) \tilde{W} \right) t . \]

Lemma 9. Suppose that Assumptions 3 and 4 hold. There exist some \( \bar{\mu} (P) > 0 \) such that for any \( \mu < \bar{\mu} (P) \) the solution to Program 9

\[ \vartheta (\mu; P) = \vartheta (0+; P) . \]

This \( \bar{\mu} (P) \) can be chosen equal to

\[ \bar{\mu} (P) \triangleq \frac{\eta (P)}{2\sqrt{\beta}} \min_{j=p+1,k} \text{ s.t. } e'_j \lambda (P) > 0 \] \( e'_j \lambda (P) \),

or 1 if \( e'_j \lambda (P) = 0 \) for all \( j = p + 1, \ldots, k \).

Proof. Step 1. Consider a the program 9. If \( e'_j \lambda (0; P) > 0 \) for some \( j = p + 1, \ldots, k \) then by continuity of \( \lambda (\mu; P) \) (Lemma 5) the component \( e'_j \lambda (\mu; P) > 0 \) in some neighborhood \( 0, e'_j \lambda \). So, for any \( \mu \in (0, \bar{\mu} (P)) \) with \( \bar{\mu} (P) = \min_j e'_j \lambda \) we get the inclusion \( I^+ (0; P) \subseteq I^+ (\mu; P) \). Consider any \( \mu \in (0, \bar{\mu} (P)) \). Using the same argument by continuity of \( \lambda (\mu; P) \), there exist some \( 0 < \epsilon < \bar{\mu} (P) - \mu \) such that

\[ I^+ (\mu; P) \subseteq I^+ (\mu'; P) \]

for any \( \mu' \in [\mu, \mu + \epsilon] \). Consider the program 9 with \( \delta (\vartheta) = 0 \) and \( \alpha_\delta = 1 \) for this \( \mu \). By definition
of $I_\delta^+ (\mu; P)$,

$$I^+ (\mu; P) \subseteq I_\delta^+ (\mu; P) \subseteq I^+ (\mu'; P).$$

Consequently, the following inclusions holds:

$$I^+ (0; P) \subseteq I^+ (\mu; P) \subseteq I_\delta^+ (\mu; P).$$

(47)

By definition of the index sets and (47),

$$\lambda (0; P) = I_\delta^+ (\mu; P) \delta I_\delta^+ (\mu; P) \lambda(0; P),$$

$$\lambda (\mu; P) = I_\delta^+ (\mu; P) \delta I_\delta^+ (\mu; P) \lambda(\mu; P).$$

Step 2. The first order condition for Program (9) takes form

$$e' + 2\mu \vartheta' (\mu; P) = -A (\mu; P) \Psi X.$$

So $\vartheta' (\mu; P) = -(2\mu)^{-1} (A (\mu; P) - A (0; P)) \Psi X$,

$$= -(2\mu)^{-1} (A (\mu; P) - A (0; P)) I_\delta^+ (\mu; P) \lambda(\mu; P).$$

(48)

By Lemma 7 for $\delta (\theta) = 0$ and $\alpha_\delta = 1$ we get

$$\frac{\partial \vartheta (\mu+; P)}{\partial \mu} = -\frac{1}{\mu} P_\delta (\mu; P) \vartheta (\mu; P) =$$

$$= P_\delta (\mu; P) A_\delta (\mu; P) \Psi.$$

with $\Psi = (2\mu^2)^{-1} I_\delta^+ (\mu; P) (A (\mu; P) - A (0; P)).$ By definition, $P_\delta (\mu; P) A_\delta (\mu; P) = 0$, so $\vartheta (\mu; P)$ is continuous function with r.h.s. derivative equal to zero, i.e. a constant for any $\mu \in (0, \mu (P)).$

The equation (48) becomes

$$\lambda (\mu; P) = \lambda (P) - 2\mu \vartheta' (0+; P) A_\delta^+ (\mu; P) I_\delta^+ (\mu; P).$$
The second term is bounded,

\[ 2 \varrho'(0+; P) A_\delta^+(\mu; P) \leq \frac{2\gamma_\Theta}{\eta(P)}. \]

So the bound \( \bar{\mu}(P) \) can be chosen as

\[
\bar{\mu}(P) = \min \left\{ 1, \frac{\eta(P)}{2\gamma_\Theta} \min_{j=p+1,\ldots,k, \text{ s.t. } \epsilon_j^\Delta(P) > 0} \epsilon_j^\Delta(P) \right\},
\]

to guarantee that any Lagrange multiplier \( \epsilon_j^\Delta(P) \) on inequality constraint, that is positive at \( \mu = 0 \), is positive on \( \mu \in (0, \bar{\mu}(P)) \).

\[ \square \]

**Lemma 10.** There exist \( \mu_1(P) > 0 \) such that for any \( 0 \leq \mu \leq \mu_1(P) \) and \( P \in \mathcal{P} \)

\[ \epsilon_1^\varrho(\mu; P) = \nu(P). \]

**Proof.** Consider

\[
\mu_1(P) = \max_{\epsilon_1^\varrho(\mu; P) = \nu(P), 0 \leq \mu \leq 1} \mu.
\]

By Lemma \([49]\) for any \( P \in \mathcal{P} \) and \( 0 \leq \mu \leq \bar{\mu}(P) \)

\[ \epsilon_1^\varrho(\mu; P) = \nu(P). \]

So \( 0 < \min \{ \bar{\mu}(P), 1 \} \leq \mu_1(P) \). By Lemma \([5]\) the function \( \epsilon_1^\varrho(\mu; P) \) is continuous. The set

\[ \mathcal{D}(P) = \{ \mu | \epsilon_1^\varrho(\mu; P) = \nu(P), 0 \leq \mu \leq 1 \} \]

is a compact–valued continuous correspondence for \( P \in \mathcal{P} \). By the Maximum Theorem (See [Ok 2007]) \( \mu_1(P) \) is a continuous function of \( P \in \mathcal{P} \). By Lemma \([6]\) \( \mathcal{P} \) is compact so

\[ \mu_1 = \min_{P \in \mathcal{P}} \mu_1(P) > 0. \]

\[ \square \]
7.2.4 Estimators

Lemma 11 (Convergence of the empirical processes). The following results hold uniformly $P \in \mathcal{P}$ in for any combination of indices $r, \ell, j, m$

\[
\lim_{n \to \infty} \sup_{P \in \mathcal{P}} P \left( (W_{r,\ell} W_{j,m})_n - \mathbb{E}_P |W_{r,\ell} W_{j,m}| \geq \xi \right) = 0, \quad (50)
\]

\[
\lim_{n \to \infty} \sup_{P \in \mathcal{P}} P \left( \overline{W}_{n,\ell} - \mathbb{E}_P W_{r,\ell} \geq \xi \right) = 0, \quad (51)
\]

\[
G_n(P) = \sqrt{n} \ \text{vec} \ \overline{W}_n - \text{vec} \left( \mathbb{E}_P W \right) \rightsquigarrow N(0, \Omega_P), \quad (52)
\]

where $\Omega_P = \text{Cov}_P (\text{vec} (W))$. These results also hold for $P$ that satisfy Assumptions 2&1.

Proof. Consider any combination of indices for any combination of indices $r, \ell, j, m$. By Step 2 in the proof of Lemma 3 a class of random variables $W_{r,\ell} W_{j,m}$ with measures in $P \in \mathcal{P}$ is uniformly integrable. The limit (50) follows immediately from Proposition A.5.1 in van der Vaart and Wellner (1996). The result (51) is analogous. The result (52) follows from Proposition A.5.2. \hfill \Box

Lemma 12. Consider any sequence $\mu_n$ such that $\mu_n \to 0$ and $\mu_n \sqrt{n} \to \infty$. Then

\[
\sqrt{n} \frac{v(\mu_n; \mathbb{P}_n) - v(\mu_n; P)}{\sigma(\mu_n; \mathbb{P}_n)} \rightsquigarrow N(0, 1),
\]

\[
\sqrt{n} \frac{-\bar{v}(\mu_n; \mathbb{P}_n) + \bar{v}(\mu_n; P)}{\bar{\sigma}(\mu_n; \mathbb{P}_n)} \rightsquigarrow N(0, 1),
\]

uniformly in $P \in \mathcal{P}$. The result is also true for a fixed $P$ that satisfies Assumptions 2&3.

Proof. Step 1. Let

\[
\Sigma(\theta; P) \triangleq \mathbb{E}_P \left( W(\theta', 1)'(\theta', 1) W' - \mathbb{E}_P [W](\theta', 1)'(\theta', 1) \mathbb{E}_P [W]' \right),
\]

\[
\Sigma(\theta; \mathbb{P}_n) \triangleq \mathbb{W}(\theta', 1)'(\theta', 1) \mathbb{W}'_n - \mathbb{W}_n(\theta', 1)'(\theta', 1) \mathbb{W}'_n.
\]

By triangular inequality,
\[
\left| \mathcal{X}'(\mu; P_n) \Sigma (\vartheta (\mu; P_n); P_n) \lambda (\mu; P_n) - \mathcal{X}'(\mu; P) \Sigma (\vartheta (\mu; P); P) \lambda (\mu; P) \right| \leq \\
\left| \mathcal{X}'(\mu; P_n) \Sigma (\vartheta (\mu; P_n); P_n) \lambda (\mu; P_n) - \mathcal{X}'(\mu; P) \Sigma (\vartheta (\mu; P); P_n) \lambda (\mu; P) \right| + \\
\left| \mathcal{X}'(\mu; P) \Sigma (\vartheta (\mu; P); P_n) \lambda (\mu; P) - \mathcal{X}'(\mu; P) \Sigma (\vartheta (\mu; P); P) \lambda (\mu; P) \right|
\]

Consider \( t = 1/\sqrt{n} \) and

\[
\delta (\theta) = \sqrt{n} \bar{X}_n - \mathbb{E}_P X \theta + \sqrt{n} (\bar{w}_n - \mathbb{E}_P w). \tag{53}
\]

By Lemma (11), \( \sqrt{n} \bar{w}_n - \mathbb{E}_P w = O_P (1) \) uniformly in \( P \in \mathcal{P} \). I use \( O_P (1) \) to denote a random variable with a second moment uniformly bounded over \( P \in \mathcal{P} \), i.e.

\[
\mathbb{E}_P \sqrt{n} \bar{w}_n - \mathbb{E}_P w^2 \leq k (d + 1) \bar{M}^{1/(1+\epsilon/2)}.
\]

By Lemma (11) the Lipschitz constant and the Maximum norm for \( \Sigma (\theta; P_n) \) on \( \Theta \) converge in probability to the corresponding constants of \( \Sigma (\theta; P) \).

By Lemma 3 the function \( \eta (P) \) is bounded from below by \( \eta > 0 \). I can replace \( \eta (P) \) in Lemma 8 by \( \eta \). By Lemma 8 the following inequality holds for any \( \mu \) if \( n \) large enough,

\[
\left| \mathcal{X}'(\mu; P_n) \Sigma (\vartheta (\mu; P_n); P_n) \lambda (\mu; P_n) - \mathcal{X}'(\mu; P) \Sigma (\vartheta (\mu; P); P_n) \lambda (\mu; P) \right| = \frac{1}{\mu \sqrt{n}} O_P (1).
\]

By the triangular inequality and Lemma (11),

\[
\| \lambda (\mu; P) \|_2^2 (\mathbb{E}_P - \mathbb{E}_P) \mathbf{W} \vartheta (\mu; P) , 1 \prime (\vartheta (\mu; P) , 1) \mathbf{W}' + \\
\| \lambda (\mu; P) \|_2^2 (\mathbb{E}_P - \mathbb{E}_P) \mathbf{W} \vartheta (\mu; P) , 1 \prime (\vartheta (\mu; P) , 1) \mathbb{E}_P \mathbf{W}' +
\]
\[ \|\mathbf{X}(\mu; P)\|_2^2 \cdot \mathbb{E}_{\mathbb{P}_n} \mathbf{W} \mathbf{\vartheta}(\mu; P)' , 1' (\mathbf{\vartheta}(\mu; P), 1) \left( \mathbb{E}_{\mathbb{P}_n} - \mathbb{E}_P \right) \mathbf{W}' = \frac{1}{\sqrt{n}} O_P (1) , \quad (57) \]

where \( \mathbb{E}_{\mathbb{P}_n} [\cdot] = \overline{[\cdot]}_n \). Finally,

\[ \mathcal{N}(\mu; \mathbb{P}_n) \Sigma(\mathbf{\vartheta}(\mu; \mathbb{P}_n); \mathbb{P}_n) \mathbf{\Lambda}(\mu; \mathbb{P}_n) = \mathcal{N}(\mu; P) \Sigma(\mathbf{\vartheta}(\mu; P); P) \mathbf{\Lambda}(\mu; P) + \frac{1}{\mu \sqrt{n}} O_p (1) . \quad (58) \]

**Step 2.** The compact correspondences \( \sigma^2(\mu; P) \) and \( \bar{\sigma}^2(\mu; P) \) can be multivalued for \( \mu = 0 \). The matrix function \( \Sigma(\cdot; P) \) is continuous. By Lemma \[ \mathbf{\vartheta}(\mu; P), \mathbf{\vartheta}(\mu; P), \mathbf{\Lambda}(\mu; P), \) and \( \mathbf{\Lambda}(\mu; P) \) are upper hemicontinuous for \( \mu \geq 0 \) and continuous for \( \mu > 0 \). So \( \sigma^2(\mu; P), \bar{\sigma}^2(\mu; P) \) are continuous for \( \mu > 0 \). So for any converging sequence \( \{(\mu_n, P_n)\}_{n=1}^{\infty} \subset [0, +\infty) \times \mathcal{P} \) with \( \mu_n \to 0 \) the limits \( \lim_{n \to \infty} \sigma^2(\mu_n; P_n) \) and \( \lim_{n \to \infty} \bar{\sigma}^2(\mu_n; P_n) \) exist and belong correspondingly to the sets \( \sigma^2(0; \lim_{n \to \infty} P_n) \) and \( \bar{\sigma}^2(0; \lim_{n \to \infty} P_n) \). The limits \( \lim_{n \to \infty} \sigma^2(\mu_n; P_n), \lim_{n \to \infty} \bar{\sigma}^2(\mu_n; P_n) \) are positive by Condition 5 of Definition 5. Correspondingly by Step 1 for any \( \xi > 0 \),

\[ \lim_{n \to \infty} \sup_{P \in \mathcal{P}} P \sigma^2(\mu_n; P_n) \leq \sigma_0^2 - \xi = 0 , \quad (59) \]

\[ \lim_{n \to \infty} \sup_{P \in \mathcal{P}} P \bar{\sigma}^2(\mu_n; P_n) \leq \sigma_0^2 - \xi = 0 . \quad (60) \]

**Step 3.** Consider any converging sequence \( \{(\mu_n, P_n)\}_{n=1}^{\infty} \subset [0, +\infty) \times \mathcal{P} \) with \( \mu_n \to 0 \). By continuity of \( \mathbf{\Lambda}(\mu; P) \) and u.h.c of \( \mathbf{\vartheta}(\mu; P) \) established in Lemma \[ \mathbf{\vartheta}(\mu; P), \mathbf{\vartheta}(\mu; P), \mathbf{\Lambda}(\mu; P), \) and \( \mathbf{\Lambda}(\mu; P) \) are continuous.

\[ \frac{\mathcal{N}(\mu_n; P_n) \sqrt{n} \overline{\mathbf{X}_n} - \mathbb{E}_{P_n} \mathbf{X} \mathbf{\vartheta}(\mu_n; P_n) + \sqrt{n} \left( \overline{\mathbf{w}_n} - \mathbb{E}_{P_n} \mathbf{w} \right)}{\lim_{n \to \infty} \sigma^2(\mu_n; P_n)} \sim N(0, 1) . \]

By Slutsky’s theorem,

\[ G_n(\mu_n; P) = \frac{\mathcal{N}(\mu_n; P) \sqrt{n} \overline{\mathbf{X}_n} - \mathbb{E}_{P} \mathbf{X} \mathbf{\vartheta}(\mu_n; P) + \sqrt{n} \left( \overline{\mathbf{w}_n} - \mathbb{E}_{P} \mathbf{w} \right)}{\bar{\sigma}(\mu_n; \mathbb{P}_n)} \sim N(0, 1) \]

uniformly in \( P \in \mathcal{P} \).
By Lemma 8 the following inequality

\[ |\xi_t (\mu, t; P) - \xi (\mu; P) + t \Delta (\mu; P) \delta (\theta (\mu; P))| \leq t^2 C_v (\mu, \delta), \]

holds for \( t \) small enough. Take \( t = 1/\sqrt{n} \) and \( \delta (\theta) \) as in (53). For any realization of \( (53) \) the following result holds

\[
\sqrt{n} \left( \frac{v (\mu_n; \mathbb{P}_n) - v (\mu; P)}{\sigma (\mu_n; \mathbb{P}_n)} - \mathcal{G}_n (\mu_n; P_n) \right) =
\sqrt{n} \frac{\bar{X}_n - \mathbb{E}_{P_n} X}{\mu_n \sqrt{n} \sigma (\mu_n; \mathbb{P}_n)} 2 O (1) + \frac{\bar{W}_n - \mathbb{E}_{P_n} W}{\sigma (\mu_n; \mathbb{P}_n)} 2 O (1)
\]

(61)

for a sufficiently large \( n \) with probability approaching 1. By Lemma 11 and 59 the r.h.s. of (61) converges to 0 in probability uniformly in \( P \in \mathcal{P} \) if \( \mu_n \sqrt{n} \to \infty \). So

\[
\sqrt{n} \frac{v (\mu_n; \mathbb{P}_n) - v (\mu; P)}{\sigma (\mu_n; \mathbb{P}_n)} \sim N (0, 1)
\]

(63)

uniformly in \( P \in \mathcal{P} \). The proof for the upper bound is analogous.

The proof for a fixed \( P \) is analogous. The only difference is (59) is need to be replaced by

\[
\sigma^2 (\mu_n; \mathbb{P}_n) \xrightarrow{P} \bar{\sigma}^2 (0+; P) > 0, \quad (64)
\]

\[
\bar{\sigma}^2 (\mu_n; \mathbb{P}_n) \xrightarrow{P} \bar{\sigma}^2 (0+; P) > 0. \quad (65)
\]

7.3 Proof of Theorem 1

Proof. Step 1. Using (45) and the argument from Lemma 12,

\[
\mu_n \sqrt{n} \| \hat{\phi} (\kappa_n; \mathbb{P}_n) \|^2 - \| \hat{\phi} (\kappa; P) \|^2 = \mu_n \left( \sqrt{n} \frac{\bar{X}_n - \mathbb{E}_{P_n} X}{\kappa_n \sqrt{n}} \right)^2 O (1) + \bar{W}_n - \mathbb{E}_{P_n} W^2 O (1).
\]

The r.h.s. of this equation converges to 0 in probability.
By Lemma 9, for \( n \) such that \( \kappa_n \leq \bar{\mu} (P) \)

\[
\| \vartheta (\kappa_n; P) \|^2 = \| \vartheta (\mu_n; P) \|^2 = \| \vartheta (0+; P) \|^2, \\
\epsilon_1' \vartheta (\mu_n; P) = \vartheta (P).
\]

Take

\[
\xi_{1,n} \overset{\triangle}{=} \sqrt{n - \frac{\vartheta (\mu_n; P) - \mu_n \| \vartheta (\kappa_n; P) \|^2 - \vartheta (P)}{\sigma (\mu_n; P)}}, \\
\xi_{2,n} \overset{\triangle}{=} \sqrt{n - \frac{-\bar{\vartheta} (\mu_n; P) - \mu_n \bar{\vartheta} (\kappa_n; P) - \bar{\vartheta} (P)}{\bar{\sigma} (\mu_n; P)}}.
\]

By Lemma (12) and the Slutsky’s theorem, \( \xi_{1,n} \) and \( \xi_{2,n} \) converge in distribution to \( N (0, 1) \).

**STEP 2.** Consider the one–sided confidence interval \( \text{CI}_{\alpha,n}^L \).

\[
\lim_{n \to \infty} P \ S (P) \subset \text{CI}_{\alpha,n}^L \\
= \lim_{n \to \infty} P \ v (P) \geq \vartheta (\mu_n; P) - \mu_n \| \vartheta (\kappa_n; P) \|^2 - \vartheta (P) z_{1-\alpha} n^{-1/2} \\
= \lim_{n \to \infty} P \{ \xi_{1,n} \leq z_{1-\alpha} \} \\
= \Phi (z_{1-\alpha}) = 1 - \alpha.
\]

The proof for \( \text{CI}_{\alpha,n}^R \) is analogous. Proof for \( \text{CI}_{\alpha/2,n}^* \) follows immediately from the Bonferroni inequality.

Suppose that \( p = 0 \). The following argument follows the proof of [Imbens and Manski (2004)].

By Lemma 1, \( \vartheta (0; P) < -\bar{\vartheta} (0; P) \). So

\[
\lim_{n \to \infty} \min \left\{ \frac{\vartheta (\mu_n; P) - \mu_n \| \vartheta (\kappa_n; P) \|^2 - \vartheta (P) \bar{\sigma} (\mu_n; P) z_{1-\alpha} n^{-1/2}}{\sigma (\mu_n; P)}, 1, \ldots \right\} = \bar{v} (P) \\
\lim_{n \to \infty} \frac{-\bar{\vartheta} (\mu_n; P) - \mu_n \bar{\vartheta} (\kappa_n; P) - \bar{\vartheta} (P) + z_{1-\alpha} \bar{\sigma} (\mu_n; P) n^{-1/2}}{\bar{\sigma} (\mu_n; P)} = \bar{v} (P) \\
\min \left\{ \lim_{n \to \infty} P \ S (P) \subset \text{CI}_{\alpha,n}^L, 1, \lim_{n \to \infty} P \ S (P) \subset \text{CI}_{\alpha,n}^R \right\} = (66) \\
\min \left\{ 1 - \alpha, 1, 1 - \alpha \right\} = 1 - \alpha.
\]
To understand the second equation, consider the following argument. Suppose that \( \theta \in \Theta (\mathbb{P}) \) is such that \( \underline{v}(\mathbb{P}) < e'_1 \theta < \bar{v}(\mathbb{P}) \). Then \( e'_1 \theta \) will be covered with probability 1 since the \( \text{CI}_{\alpha,n}^L \) and \( \text{CI}_{\alpha,n}^R \) and upper bounds of \( \text{CI}_{\alpha,n}^B \) cover correspondingly \( \underline{v}(\mathbb{P}) \) and \( \bar{v}(\mathbb{P}) \).

\[ \square \]

### 7.4 Proof of Proposition [1]

**Proof.** Structure of the proof. First I provide the asymptotic expansion of the length for \( \text{CI}_{\alpha,n}^{FH} \) and \( \text{CI}_{\alpha,n} \). Then I compare the critical values in \( \text{CI}_{\alpha,n}^{FH} \) and \( \text{CI}_{\alpha,n} \) with quantiles of the asymptotic distribution of the support functions without regularization. The order of quantiles implies the asymptotic length comparison.

Step 1. The \( \text{CI}_{\alpha,n}^{FH} \) takes form

\[ \underline{v}(0; \mathbb{P}_n) + n^{-1/2} \underline{\zeta}_{\alpha,n}^{FH} + \bar{v}(0; \mathbb{P}_n) + n^{-1/2} \bar{c}_{1-\alpha,n}^{FH} \, . \]  

(68)

By Section 4.1 of Freyberger and Horowitz (2015) \( \hat{c}_{1-\alpha,n}^{FH} \xrightarrow{P} c_{1-\alpha}^{FH} \). So

\[ \text{CI}_{\alpha,n}^{FH} = \bar{v}(\mathbb{P}) - \underline{v}(\mathbb{P}) + n^{-1/2} \underline{\zeta}_{\alpha}^{FH} + \bar{c}_{1-\alpha}^{FH} \, n^{-1/2} + o_p \, n^{-1/2} \, . \]  

(69)

By Theorem [1]

\[ |\text{CI}_{\alpha,n}| = \bar{v}(\mathbb{P}) - \underline{v}(\mathbb{P}) + z_{1-\alpha} n^{-1/2} (\underline{\sigma}(0+; \mathbb{P}) + \bar{\sigma}(0+; \mathbb{P})) + o_p \, n^{-1/2} \, . \]  

(70)

Step 2. By Lemma [1]

\[ \sqrt{n} \, W_n = \mathbb{E}_P W \sim Z. \]

By Theorem 3.5 of Shapiro (1991)

\[ \sqrt{n} (\underline{v}(0; \mathbb{P}_n) - \underline{v}(\mathbb{P})) \sim \min_{\theta \in \Theta (0; \mathbb{P})} X' (\mathbb{P}) Z (\theta', 1)' , \]  

(71)

\[ \sqrt{n} (\bar{v}(0; \mathbb{P}_n) - \bar{v}(\mathbb{P})) \sim \max_{\theta \in \Theta (0; \mathbb{P})} \bar{\lambda} (\mathbb{P}) Z (\theta', 1)' . \]  

(72)
By Lemmas 9 and 12,

$$\sqrt{n} \ v (\mu_n; \mathbb{P}_n) - \mu_n \| \theta (\kappa_n; \mathbb{P}_n) \|^2 - v (P) \sim \chi^2 (P) Z (\tilde{\theta}^\prime (0+; P), 1)^\prime,$$

$$\sqrt{n} \ - \tilde{v} (\mu_n; \mathbb{P}_n) + \mu_n \ \tilde{\theta} (\kappa_n; \mathbb{P}_n)^2 + \tilde{v} (P) \sim \chi (P) Z \ \tilde{\theta} (0+; P), 1^\prime.$$

Let $\zeta_\alpha$ and $\bar{c}_{1-\alpha}$ be the $\alpha$ and $1 - \alpha$ quantiles of (71) and (72) correspondingly. By Section 3.3 of Freyberger and Horowitz (2015) $\bar{c}_{1-\alpha}^F \geq \bar{c}_\alpha^F \geq -\zeta_\alpha$. Since by Lemma 3 $\theta (0+; P) \in \theta (0; P)$ and $\tilde{\theta} (0+; P) \in \tilde{\theta} (0; P)$,

$$z_{1-\alpha} \sigma (0+; P) \leq -\zeta_\alpha \ \text{and} \ z_{1-\alpha} \bar{\sigma} (0+; P) < \bar{c}_{1-\alpha}.$$

So

$$-\zeta_\alpha^F + \bar{c}_1^F \geq z_{1-\alpha} (\sigma (0+; P) + \bar{\sigma} (0+; P)).$$

The stochastic expansion for the difference between (69) and (70),

$$\text{CI}_{a,n}^F - |\text{CI}_{a,n}| = -\zeta_\alpha^F + \bar{c}_1^F - z_{1-\alpha} (\sigma (0+; P) + \bar{\sigma} (0+; P)) \ n^{-1/2} + o_p \ n^{-1/2},$$

implies the claim of the proposition.

### 7.5 Proof of Theorem 2

**Proof.** Consider arbitrary convergent sequence $\{P_n\}_{n=1}^\infty \in \mathcal{P}$. The proof is analogous for all CI. Consider, for example, $\tilde{\text{CI}}_{a,n}^L$.

Using a similar argument to Lemmas 12

$$\beta (\mathbb{P}_n) = \beta (P_n) + O_p \ \frac{1}{\sqrt{n}}.$$

By Lemma 10 for $\mu_n \leq \mu_1 (\mathbb{P})$ we get $v (P_n) = e^\prime_1 \theta (\mu_n; P_n)$.

By Lemma 12
\[ \sqrt{n} \frac{v(\mu_n; \mathbb{P}_n) - v(P_n) - \mu_n \beta(\mathbb{P}_n)}{\sigma(\mu_n; \mathbb{P}_n)} = \xi + \mu_n \sqrt{n} \| \vartheta(\mu_n; P_n) \|^2 - \beta(P_n) + \mu_n O_p(1). \]

By definition,

\[ \| \vartheta(\mu_n; P_n) \|^2 - \beta(P_n) \leq 0. \]

So

\[ P_n \frac{v(\mu_n; \mathbb{P}_n) - v(P_n) - \sigma(\mu_n; \mathbb{P}_n) z_{1 - \alpha} n^{-1/2}}{\sigma(\mu_n; \mathbb{P}_n)} \leq v(P_n) \leq P_n \frac{\sqrt{n} v(\mu_n; \mathbb{P}_n) - v(\mu_n; P_n)}{\sigma(\mu_n; \mathbb{P}_n)} + \mu_n \sqrt{n} (\beta(P_n) - \beta(P_n)) \leq z_{1 - \alpha} \to 1 - \alpha. \]

Since \( P_n \) is an arbitrary convergent sequence,

\[ \liminf_{n \to \infty} \inf_{P \in \mathcal{P}} \min_{\theta \in \Theta(P)} P_n \theta_1 \in \tilde{\mathcal{C}}_{1,n} \geq 1 - \alpha. \] (73)