The Numerical Delta Method and Bootstrap *

Han Hong           Jessie Li

October 28, 2016

Abstract

This paper studies inference on nondifferentiable functions using methods based on numerical differentiation. First we show that for an appropriately chosen sequence of step sizes, numerical derivative based delta methods provide consistent inference for functions of parameters that are directionally differentiable. Examples of directionally differentiable functions arise in a variety of economic applications such as moment inequalities models and threshold regression models. Second, we propose a numerical bootstrap method that provides asymptotically valid inference even for parameters that are not known to be directionally differentiable. The numerical bootstrap can consistently estimate the limiting distribution in many cases where the conventional bootstrap is known to fail and where subsampling has been the most commonly used inference approach. Applications include constrained and unconstrained M-estimators converging at both regular and nonstandard rates, misspecified simulated GMM models with nondifferentiable moments, LASSO, and 1-norm Support Vector Machine regression.

Keywords: Bootstrap, Numerical Differentiation, Directional Differentiability

JEL Classification: C12; C13; C50

*Department of Economics, Stanford University. We thank Andres Santos, Joe Romano, Xiaohong Chen, Zheng Fang, Bruce Hansen, David Kaplan, Adam McCloskey, Frank Wolak and participants at the Montreal Econometric Society World Congress and various conferences and seminars for helpful comments. We also acknowledge support by the National Science Foundation (SES 1164589), and both the IRiSS and the B.F.Haley and E.S.Shaw SIEPR dissertation fellowships.
1 Introduction

Inference on possibly nonsmooth functions of parameters has received much attention in the econometrics literature, as in Woutersen and Ham (2013) and Hirano and Porter (2012). In particular, a recent insightful paper by Fang and Santos (2014) studies inference for functions of the parameters that are only Hadamard directionally differentiable and not necessarily differentiable. Fang and Santos (2014) show that while the asymptotic distribution obtained using the bootstrap is invalid unless the target function of the parameter is differentiable, asymptotic inference using a consistent estimate of the first order directional derivative is valid as long as the target function is Hadamard directionally differentiable. In each of their examples studied, Fang and Santos (2014) constructed consistent analytical estimates of the directional derivative that are tailored to each particular case.

As an alternative to using analytical estimates, we show that numerical differentiation provides a comprehensive approach to estimating the directional derivative. The main advantage of using the numerical directional derivative is its computational simplicity and ease of implementation. In order to compute an estimate of the directional derivative, the user only needs to specify one tuning parameter (the stepsize), and she does not need to perform any additional calculations beyond evaluating the target function twice for each random draw from an approximation of the limiting distribution of the parameter estimates.

In some applications, the first order directional derivative may vanish on a set of parameters, which motivates the use of the second order numerical directional delta method. For example, the test statistics for moment inequality models often use the negative square test function, which has the property that the first order directional derivative is exactly zero over the null set. We demonstrate the pointwise consistency of the second order numerical directional derivative and demonstrate that it can be used to approximate the limiting distribution in the second order delta method.

The directional delta method assumes that we are conducting inference on a function that is directionally differentiable. However, in many cases of interest, we would like to obtain the limiting distribution of a parameter estimate that is the argmin of a nonsmooth objective function. To our knowledge, there are no results in the literature showing that the argmin functional of a nonsmooth objective function is directionally differentiable. Instances of such estimators include the maximum score
estimator of Manski (1975) and its constrained version, the LASSO estimator of Tibshirani (1996), and the 1-norm Support Vector Machine (SVM) regression estimator of Zhu et al. (2004). The limiting distribution of these estimators cannot be estimated consistently using the regular bootstrap, and currently, the most commonly used inference method is subsampling.

In order to conduct inference in these examples, we propose a new type of bootstrap called the numerical bootstrap which offers an alternative to subsampling in many cases where the regular bootstrap fails. We also show how to conduct inference for sample size dependent statistics such as the Laplace estimators of Chernozhukov and Hong (2003) and Jun et al. (2015) and generalize our results to the constrained versions. Additionally, we relate the numerical bootstrap to the partially identified moment inequality models of Andrews and Shi (2013), Bugni et al. (2015), and Bugni et al. (2014). An additional result we obtain, which appears to be new, is that misspecified nonsmooth simulated GMM estimators converge at the cube-root rate to limiting distributions which can be estimated by the numerical bootstrap.

We begin in section 2 with a description of the inference problem and provide motivating examples of nondifferentiable functions. Section 3 first shows pointwise consistency of the numerical delta method and demonstrates the (uniform) asymptotic validity under a convexity condition and a Lipschitz condition. These conditions are satisfied in all the examples provided in Fang and Santos (2014) as well as for the test statistics used in certain moment inequality models. Section 4 describes the second order numerical directional delta method and its applications to Andrews and Soares (2010), Bugni et al. (2014), and Bugni et al. (2015).

Section 5 provides an overview of the numerical bootstrap method. Section 5.1 contains some heuristic arguments comparing numerical bootstrap to subsampling. The later sections consider applications of the numerical bootstrap. Section 6 validates the consistency of the numerical bootstrap for a class of M-estimators that includes the maximum score estimator developed by Manski (1975) and whose asymptotics are derived in Kim and Pollard (1990) and Cavanagh (1987). We also allow the true parameter to lie on the boundary of a constrained set, as in the setup of Geyer (1994). Section 7 describes an application of the numerical bootstrap to misspecified simulated GMM models. Sections 8 and 9 contain applications of the numerical bootstrap to the LASSO estimator of Tibshirani (1996) and the 1-norm support vector machine regression estimator of Zhu et al. (2004). The role of recentering in hypoth-
esis testing is discussed in section 10. Section 11 discusses how to use the numerical bootstrap to estimate an unknown polynomial convergence rate. Section 12 compares the finite sample power of various tests for the entry game example in Bugni et al. (2014), section 13 provides an empirical application to maximum quantile treatment effects in the Tennessee STAR experiment, and section 14 concludes. The appendix contains a list of commonly used symbols and proofs of the main theorems.

2 Numerical directional delta method

Fang and Santos (2014) study inference on a nondifferentiable mapping $\phi(\theta)$ of the parameter $\theta \in \Theta$, where $\theta$ can be either finite or infinite dimensional, under the requirement that $\theta \in D_\phi$ and $\phi: D_\phi \subset D \rightarrow E$ for $D$ endowed with norm $|| \cdot ||_D$ and $E$ endowed with norm $|| \cdot ||_E$. The domain of $\phi$ is $D_\phi$.

The true parameter is denoted $\theta_0 \in D_\phi$, for which a consistent estimator $\hat{\theta}_n$ is available which converges in distribution at a suitable rate $r_n \rightarrow \infty$: $r_n \left( \hat{\theta}_n - \theta_0 \right) \Rightarrow G_0$ in $D$ in the sense of equation (2.8) of Kosorok (2007)

$1$ where the limit distribution $G_0$ is tight and is supported on $D_0 \subset D$. Examples of nondifferentiable $\phi(\cdot)$ functions arise in a variety of econometric applications such as moment inequalities models (Andrews and Shi (2013), Ponomareva (2010)) and threshold regression models (Hansen (2015)). Using the notation of Fang and Santos (2014), we describe each of these examples in more detail below.

**Fang and Santos (2014) Example 2.1 and Hansen (2015)** Define $\phi(\theta) = a\theta^+ + b\theta^-$, where $\theta^+ = \max\{\theta, 0\}$ and $\theta^- = -\min\{\theta, 0\}$. Let $X \in \mathbb{R}$, $\theta_0 = E[X]$, and $D = E = \mathbb{R}$.

**Fang and Santos (2014) Example 2.2** $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$, $\phi(\theta) = \max(\theta_1, \ldots, \theta_k)$. $D = \mathbb{R}^d \times \mathbb{R}^d \times \ldots \times \mathbb{R}^d$ and $E = \mathbb{R}$.

**Fang and Santos (2014) Example 2.3** Define $\phi(\theta_0) = \sup_{f \in \mathcal{F}} E(Yf(Z))$ as in Andrews and Shi (2013). Here, $Y \in \mathbb{R}$, $Z \in \mathbb{R}^d$, and $\theta_0 \in \ell^\infty(\mathcal{F})$. $\mathcal{F} \subset \ell^\infty(\mathcal{R}^d)$ is a set of functions satisfying $\theta_0(f) \equiv \theta(P)(f) = EYf(Z)$ for all $f \in \mathcal{F}$. $D = \ell^\infty(\mathcal{F})$ and $E = \mathbb{R}$.

$1$ $X_n \Rightarrow X_n$ in the metric space $(D, d)$ if and only if $\sup_{f \in BL_1} |E^*f(X_n) - Ef(X)| \rightarrow 0$ where $BL_1$ is the space of functions $f: D \rightarrow \mathbb{R}$ with Lipschitz norm bounded by 1.
**Ponomareva (2010) Example** \( \phi(\theta_0) = \max_{x \in X} E[m(Z_i; \beta) | X_i = x] \) where \( \beta \in B \) a compact subset of \( \mathbb{R}^d \). \( D = \ell^\infty(\mathbb{R}^d) \) and \( E = \mathbb{R} \).

The goal of our subsequent analysis is to approximate the distribution of \( r_n \left( \phi(\hat{\theta}_n) - \phi(\theta_0) \right) \) for statistical inference concerning \( \phi(\theta_0) \). The key requirement is that the function of interest must be Hadamard directionally differentiable.

**Definition 2.1** The map \( \phi \) is said to be Hadamard directionally differentiable at \( \theta \in D_\phi \) tangentially to a set \( D_0 \subset D \) if there is a continuous map \( \phi'_\theta : D_0 \rightarrow \mathbb{E} \) such that:

\[
\lim_{n \to \infty} \left\| \frac{\phi(\theta + t_nh_n) - \phi(\theta)}{t_n} - \phi'_\theta(h) \right\|_E = 0,
\]

for all \( \{h_n\} \subset D \) and \( \{t_n\} \subset \mathbb{R}_+ \) such that \( t_n \downarrow 0, h_n \to h \in D_0 \) as \( n \to \infty \) and \( \theta + t_nh_n \in D_\phi \).

When \( \phi(\cdot) \) is directionally differentiable in the sense defined above and when the support of the limiting distribution \( G_0 \) is contained in \( D_0 \), Fang and Santos (2014) showed that under suitable regularity conditions,

\[
r_n \left( \phi(\hat{\theta}_n) - \phi(\theta_0) \right) \rightsquigarrow \phi'_\theta(\mathcal{G}_0).
\]  

Based on this result, Fang and Santos (2014) suggested that this limiting distribution can be consistently estimated if we have a consistent estimate of the directional derivative and a consistent estimate \( Z^*_n \) of \( G_0 \), such as the following:

1. If we can consistently estimate the limiting distribution of \( r_n \left( \hat{\theta}_n - \theta_0 \right) \) using the bootstrap, then we can take \( Z^*_n = r_n \left( \theta \left( P^*_n \right) - \theta \left( P_n \right) \right) \), where \( P_n \) is the empirical measure and \( P^*_n \) is the bootstrap empirical measure. We can use the multinomial, wild, or other commonly used bootstrap implementations. The bootstrap sample size can also be different from the observed sample size. For example, we can take \( Z^*_n = r_{m_n} \left( \theta \left( P^*_n \right) - \theta \left( P_n \right) \right) \), where \( m_n \to \infty \) as \( n \to \infty \), and \( \theta \left( P^*_n \right) \) is computed from a multinomial bootstrap sample of size \( m_n \). Similar modifications apply to the next few methods.

2. When \( \theta \) is a finite dimensional parameter, typically \( r_n = \sqrt{n} \) and \( \mathcal{G}_0 = N(0, \Sigma) \) for some variance covariance matrix \( \Sigma \). Using a consistent estimate \( \hat{\Sigma} \) of \( \Sigma \), \( Z^*_n \) can be a random vector whose distribution given the data is \( N \left( 0, \hat{\Sigma} \right) \).
3. For correctly specified parametric models, one can use $Z^*_n = r_n \left( \hat{\theta}^*_n - \hat{\theta}_n \right)$, where $\hat{\theta}^*_n$ are MCMC draws from the (pseudo) posterior distribution based on the likelihood or other objective functions (Chernozhukov and Hong (2003)).

4. In section 5, we propose a technique called the numerical bootstrap, which produces estimates $\theta(\mathcal{Z}^*_n)$ based on the numerical bootstrap empirical measure $\mathcal{Z}^*_n \equiv P_n + \epsilon_n \sqrt{n} (P^*_n - P_n)$, where $\epsilon_n$ is a positive scalar step size parameter that satisfies $\epsilon_n \to 0$ and $n^{\gamma} \epsilon_n \to \infty$. We show that the finite sample distribution of $\mathcal{Z}^*_n = \epsilon_n^{-2\gamma} (\theta(\mathcal{Z}^*_n) - \theta(P_n))$ converges to the same limiting distribution as that of $n^\gamma \left( \hat{\theta}_n - \theta_0 \right)$ for a class of estimators that converge at rate $n^\gamma$ for some $\gamma \in \left[ \frac{1}{4}, 1 \right]$.

For $\epsilon_n \to 0$ slowly (in the sense that $r_n \epsilon_n \to \infty$, where $r_n$ is the convergence rate of $\hat{\theta}_n$ to $\theta_0$), define

$$\hat{\phi}'_n(h) \equiv \frac{\phi \left( \hat{\theta}_n + \epsilon_n h \right) - \phi \left( \hat{\theta}_n \right)}{\epsilon_n}$$

as the numerical directional derivative of $\phi$ at $\hat{\theta}$ in the direction of $h \in \mathbb{D}_0$. The rate requirement on the step size $\epsilon_n$ is needed to separate numerical differentiation error from the estimation error in $\hat{\theta}_n$, and serves the dual purposes of model selection and numerical differentiation.

Given the definition in (3), the numerical directional delta method estimates the limiting distribution of $r_n \left( \phi \left( \hat{\theta}_n \right) - \phi \left( \theta_0 \right) \right)$ using the distribution of the random variable:

$$\hat{\phi}'_n(\mathcal{Z}^*_n) \equiv \frac{\phi \left( \hat{\theta}_n + \epsilon_n \mathcal{Z}^*_n \right) - \phi \left( \hat{\theta}_n \right)}{\epsilon_n}$$

which can be approximated by the following:

1. Draw $\mathcal{Z}_s$ from the distribution of $\mathcal{Z}^*_n$ for $s = 1, \ldots, S$.

2. For the given $\epsilon_n$, evaluate for each $s$:

$$\hat{\phi}'_n(\mathcal{Z}_s) \equiv \frac{\phi \left( \hat{\theta}_n + \epsilon_n \mathcal{Z}_s \right) - \phi \left( \hat{\theta}_n \right)}{\epsilon_n}.$$

(5)
The empirical distribution of $\hat{\phi}_n'(Z_s), s = 1, \ldots, S$ can then be used for confidence interval construction, hypothesis testing, or variance estimation. For example, a $1 - \tau$ two-sided equal-tailed confidence interval for $\phi(\theta_0)$ can be formed by

$$[\phi(\hat{\theta}) - \frac{1}{r_n} c_{1-\tau/2}, \phi(\hat{\theta}) - \frac{1}{r_n} c_{\tau/2}]$$

where $c_{\tau/2}$ and $c_{1-\tau/2}$ are the $\tau/2$ and $1 - \tau/2$ empirical percentiles of $\hat{\phi}_n'(Z_s)$. Symmetric confidence intervals can be formed by

$$[\phi(\hat{\theta}) - \frac{1}{r_n} d_{1-\tau}, \phi(\hat{\theta}) + \frac{1}{r_n} d_{1-\tau}]$$

where $d_{1-\tau}$ is the $1 - \tau$ percentile of $|\hat{\phi}_n'(Z_s)|$.

3 Asymptotic validity

This section shows that the numerical directional delta method provides consistent inference under general conditions. We first verify pointwise consistency and then discuss uniform validity.

3.1 Pointwise asymptotic distribution

In this subsection we show pointwise consistency of the numerical delta method using the definition of Hadamard directional differentiability and (a bootstrap version of) the extended continuous mapping theorem. The first part of the following theorem is a directional Delta method due to D"umbgen (1993), Fang and Santos (2014), and references therein. The second part of the theorem shows consistency of the numerical Delta method. Let $BL_1$ be the space of Lipschitz functions $f : \mathbb{D} \mapsto \mathbb{R}$ with Lipschitz norm bounded by 1. For random variables $F_1$ and $F_2$, let $\rho_{BL_1}(F_1, F_2) = \sup_{f \in BL_1} |Ef(F_1) - Ef(F_2)|$ metricize weak convergence. As in Kosorok (2007) (pages 19-20), we use $\xrightarrow{p_{\mathbb{W}}}$ to denote weak convergence in probability conditional on the data. 2

\^X_n \xrightarrow{p_{\mathbb{W}}} X$ means that $\hat{X}_n$ is a random function of the data and $\sup_{f \in BL_1} \left| E\left[ f(\hat{X}_n) | X_n \right] - Ef(X) \right| \to 0$ (where $X_n$ denotes the data).
Theorem 3.1  Suppose $D$ and $E$ are Banach Spaces and $\phi : D_\phi \subseteq D \mapsto E$ is Hadamard directionally differentiable at $\theta_0$ tangentially to $D_0$, a convex set. Let $\hat{\theta}_n : \{X_i\}_{i=1}^n \mapsto D_\phi$ be such that for some $r_n \uparrow \infty$, $r_n\{\hat{\theta}_n - \theta_0\} \rightsquigarrow G_0$ in $D$ and assume the support of $G_0$ is included in $D_0$. Then $r_n\left(\phi\left(\hat{\theta}_n\right) - \phi(\theta_0)\right) \rightsquigarrow \phi'_{\theta_0}(G_0)$.

Let $Z^*_n \overset{P}{\rightsquigarrow} G_0$ satisfy certain measurability assumptions stated in the appendix. Then for $\epsilon_n \to 0$, $r_n\epsilon_n \to \infty$,

$$\hat{\phi}'_n(Z^*_n) \equiv \frac{\phi\left(\hat{\theta}_n + \epsilon_n Z^*_n\right) - \phi\left(\hat{\theta}_n\right)}{\epsilon_n} \overset{P}{\rightsquigarrow} \phi'_{\theta_0}(G_0).$$

3.2 Uniform Inference

Uniform asymptotic validity over a class of distributions can be a desirable feature to establish for an inference procedure (Romano and Shaikh 2008; 2012). The Lipschitz and convexity properties of $\phi(\cdot)$ are key to establishing uniform size control in hypothesis testing of the following form using the test statistic $r_n\phi\left(\hat{\theta}_n\right)$:

$$H_0 : \phi(\theta_0) \leq 0 \quad \text{against} \quad H_1 : \phi(\theta_0) > 0.$$

and suggested rejecting $H_0$ whenever $r_n\phi\left(\hat{\theta}_n\right) \geq \hat{c}_{1-\tau}$, where $\hat{c}_{1-\tau}$ is the $1-\tau$ quantile of $\hat{\phi}'_n(Z^*_n)$ or its simulated version in (5). This is related to the one-sided confidence interval in Part (i) of Theorem 2.1 in Romano and Shaikh (2012):

$$P\left(r_n\left(\phi\left(\hat{\theta}_n\right) - \phi(\theta_0)\right) \leq \hat{c}_{1-\tau}\right),$$

Whenever $\phi(\theta)$ is convex and Lipschitz in $\theta$, using the $1-\tau$ percentile of $\hat{\phi}'_n(Z^*_n)$ as $\hat{c}_{1-\tau}$ provides uniform size control for both (6) and (7) under the condition that $r_n\epsilon_n \to \infty$ without requiring $\epsilon_n \to 0$. Intuitively, convexity implies for $\epsilon_n > \frac{1}{r_n}$ and for a given distribution of $Z$,

$$r_n\left(\phi\left(\theta_0 + \frac{G_0}{r_n}\right) - \phi(\theta_0)\right) \leq \frac{1}{\epsilon_n}\left(\phi\left(\theta_0 + \epsilon_n G_0\right) - \phi(\theta_0)\right),$$

8
so that the left hand side is stochastically dominated by the right hand side in distribution. If we denote, using notations from Romano and Shaikh (2012), the distribution functions of the two sides of (8) by $J_n(x, G_0)$ and $J_{\epsilon_n}(x, G_0)$, then equation (8) immediately implies that

$$\sup_n \sup_{x \in \mathbb{R}} \{J_{\epsilon_n}(x, G_0) - J_n(x, G_0)\} \leq 0.$$ (9)

Next, $\phi(\theta)$ being Lipschitz ensures that the left hand side of (8) is close to $r_n \left( \phi\left(\hat{\theta}_n\right) - \phi\left(\theta_0\right)\right)$, whose distribution function is denoted $J_n(x, P)$, while the right hand side is close to $\hat{\phi}_n'(Z^*_n)$, whose conditional distribution function given the data is $J_{\epsilon_n}(x, P)$, so that $J_n(x, G_0)$ and $J_{\epsilon_n}(x, G_0)$ in (9) can be replaced by their feasible sample versions $J_n(x, P)$ and $J_{\epsilon_n}(x, P)$.

Uniformity statements in line with those in Romano and Shaikh (2012) are possible under the following assumptions. We focus on the finite dimensional case $D = \mathbb{R}^k$ and $E = \mathbb{R}$.

**Assumption 3.1** Let $\mathcal{P}$ be a class of distributions such that each $P \in \mathcal{P}$ governs $\theta_0 = \theta(P)$ and the laws of $r_n \left( \hat{\theta}_n - \theta(P) \right)$, $G_0$, and $Z^*_n$. Suppose

(i) $\lim_{n \to \infty} \sup_{P \in \mathcal{P}} \rho_{BL_1} \left( r_n \left( \hat{\theta}_n - \theta(P) \right), G_0 \right) = 0$, $\lim_{M \to \infty} \sup_{P \in \mathcal{P}} P \left( |G_0| \geq M \right) = 0$;

(ii) for each $\epsilon > 0$, $\lim_{n \to \infty} \sup_{P \in \mathcal{P}} P \left( \rho_{BL_1} \left( Z^*_n, G_0 \right) \geq \epsilon \right) = 0$.

Primitive conditions for assumption 3.1 can be found for example in the uniform central limit theorems of Romano and Shaikh (2008).

**Assumption 3.2** Define for each $x, a, d$, $C_{a,d,x} = \{ g : \phi \left( d + \frac{g}{a} \right) \leq x \}$, then

$$\sup_{P \in \mathcal{P}} P \left( G_0 \in \partial C_{a,d,x} \right) = 0 \quad \text{for all } x, a, d,$$

where $\partial C_{a,d,x}$ denotes the boundary of $C_{a,d,x}$.

---

3Equation (8) follows from rewriting it as, for $r_n \epsilon_n > 1$,

$$\phi \left( \theta_0 + \frac{Z_{r_n}}{r_n} \right) \leq \frac{1}{r_n \epsilon_n} \phi \left( \theta_0 + \epsilon_n Z \right) + \left( 1 - \frac{1}{r_n \epsilon_n} \right) \phi \left( \theta_0 \right).$$
Assumption 3.2 is mainly used to invoke versions of Theorem 2.11 of Bhattacharya and Rao (1986), as in Example 3.2 of Romano and Shaikh (2012). If $\phi(\cdot)$ is scale equivariant, then it is sufficient to check all $C_{d,x} \equiv \{ g : \phi(d + g) \leq x \}$. Convexity is crucial in the following.

**Theorem 3.2** If $r_n \epsilon_n \to \infty$ and assumptions 3.1, 3.2 both hold, when $\phi(\cdot)$ is Lipschitz and convex, then $\forall \epsilon > 0$,

$$
\limsup_{n \to \infty} P \left( \sup_{x \in A} J_{\epsilon_n}(x, P) - J_n(x, P) \leq \epsilon \right) \to 1;
$$

$$
\limsup_{n \to \infty} P \left( r_n \left( \phi\left( \hat{\theta}_n \right) - \phi\left( \theta(P) \right) \right) \geq \hat{c}_{1-\tau} \right) \leq \tau.
$$

where $A$ is any set for which $\lim_{\lambda \to 0} \sup_{P \in \mathcal{P}} \sup_{x \in A} P\left( J_{\epsilon_n}(x, \mathcal{G}_0) \in (x, x + \lambda) \right) = o(1)$ and contains a neighborhood of both $J_{\epsilon_n}^{-1}(1 - \tau, \mathcal{G}_0)$ and $J_n^{-1}(1 - \tau, P)$ for all large $n$.

It also turns out that proper convex functions are directionally differentiable at every interior point of a finite dimensional parameter space. This is shown in Example 7.27 and the remark before Example 10.28 in Rockafellar et al. (1998), who call Hadamard directional differentiability semidifferentiability. Therefore, $r_n \left( \phi\left( \hat{\theta}_n \right) - \phi\left( \theta(P) \right) \right)$ converges to a limiting distribution that is continuous at every quantile. Then by remark 2.1 in Romano and Shaikh (2012), the coverage statements in the previous theorem can be strengthened in the following way:

**Corollary 3.1** Let the conditions in Theorem 3.2 hold over $\theta_0$ in the interior of a finite dimensional parameter space. Then for $\epsilon_n \to 0$,

$$
\limsup_{n \to \infty} P \left( r_n \left( \phi\left( \hat{\theta}_n \right) - \phi\left( \theta(P) \right) \right) \geq \hat{c}_{1-\tau} \right) = \tau \tag{10}
$$

According to Theorem 3.2 and Corollary 3.1, whenever $\phi(\cdot)$ is convex, the lower one-sided confidence interval $\left( \phi\left( \hat{\theta}_n \right) - \frac{\hat{c}_{1-\tau}}{r_n}, \infty \right)$ will have uniformly asymptotically valid coverage. Similarly, if $\phi(\cdot)$ is a concave function instead, then the same arguments will establish that the upper one-sided confidence interval of the form $\left( -\infty, \phi\left( \hat{\theta}_n \right) - \frac{\hat{c}_\tau}{r_n} \right)$ has uniformly asymptotically valid coverage. Furthermore, if it is known that $\phi(\cdot) \geq 0$ (e.g. Andrews (2000)), we can use the first term in (4) $\epsilon_n^{-1} \phi\left( \hat{\theta}_n + \epsilon_n \mathbb{Z}_n^* \right)$ in place of
\( \hat{\phi}'_n (Z^*_n) \) at the cost of being more conservative. If we take \( t_n = r_n^{-1} \) and use the bootstrap distribution \( Z^*_n = r_n (\hat{\theta}^*_n - \hat{\theta}_n) \), a modified bootstrap uses \( r_n (\phi (\theta_0 + \hat{\theta}^*_n - \hat{\theta}_n) - \phi (\theta_0)) \) to approximate the null distribution of \( r_n (\phi (\hat{\theta}_n) - \phi (\theta_0)) \). However, it does not provide moment selection functions to improve the power of the test and does not offer uniform size control for \( r_n (\phi (\hat{\theta}_n) - \phi (\theta_0)) \) under drifting sequences of \( \theta_n \). In some cases, if only \( \phi (\theta) = \phi_0 \) but not \( \theta_0 \) is known under the null, \( \hat{\theta}_n \) can be either the constrained or unconstrained estimate. Note also that the only use of the convexity of \( \phi (\cdot) \) above is the stochastic dominance condition in (8) and (9). Therefore the convexity requirement of \( \phi (\cdot) \) can be replaced by the following stochastic dominance condition:

**Assumption 3.3** For all \( \theta_0 \) and \( t > 0 \), \( \frac{\phi(\theta_0 + tZ) - \phi(\theta_0)}{t} \) is nondecreasing in \( t \).

Even if \( \phi (\theta) \) is not convex and does not satisfy assumption 3.3, it is still possible to establish uniform size control over \( \theta_0 \) under sufficient conditions for the limiting distribution of the numerical directional derivative to stochastically dominate the analytic limiting distribution over all \( \theta_0 \) that lie in the null set.

**Assumption 3.4** For all \( \theta_0 \) and \( t > 0 \), \( \frac{\phi'(\theta_0 + tZ) - \phi'(\theta_0)}{t} \) is nondecreasing in \( t \).

Clearly A3.3 (which in turn is implied by \( \phi (\cdot) \) being convex) is a sufficient condition for A3.4. A3.4 is also satisfied if \( \phi'_0(h) \) is convex in \( h \) (which in turn follows from the convexity of \( \phi (\cdot) \)), since for \( t_2 > t_1 > 0 \),

\[
\frac{\phi'_0(\theta_0 + t_1 Z) - \phi'_0(\theta_0)}{t_1} \leq \frac{\phi'_0(\theta_0 + t_1 Z - \phi'_0(\theta_0))}{t_1} \leq \frac{\phi'_0(\theta_0 + t_2 Z) - \phi'_0(\theta_0)}{t_2}
\]

A3.4 plays a similar role to (8) and (9) and implies for \( \epsilon_n r_n > 1 \):

\[
r_n (\phi'_0 (\theta_0 + \frac{G_0}{r_n}) - \phi'_0 (\theta_0)) \leq \frac{\phi'_0 (\theta_0 + \epsilon_n G_0) - \phi'_0 (\theta_0)}{\epsilon_n} \leq \phi'_0 (\theta_0 + \epsilon_n G_0) - \phi'_0 (\theta_0) \leq \phi'_0 (\theta_0 + t_n h) - \phi'_0 (\theta_0) \leq \phi'_0 (\theta_0 + t_n h) - \phi'_0 (\theta_0) \leq \phi'_0 (\theta_0 + t_n h) - \phi'_0 (\theta_0)
\]

In order for both sides of (11) to provide good approximations to \( r_n (\phi (\hat{\theta}_n) - \phi (\theta_0)) \) and \( \hat{\phi}'_n (Z_n) \), we require the following additional assumptions.

**Assumption 3.5** For any \( t_n \downarrow 0 \):

\[
\lim_{t_n \downarrow 0} \left| \frac{1}{t_n} (\phi (\theta_0 + t_n h) - \phi (\theta_0)) - \left( \phi'_0 (\theta_0 + t_n h) - \phi'_0 (\theta_0) \right) \right| = 0.
\]
For $\theta_0 = O(t_n)$, this follows from the definition of Hadamard directional differentiability, which however does not handle $\theta_0/t_n \to \infty$.

We now state a uniformity result similar to Andrews and Soares (2010) without relying on convexity.

**Theorem 3.3** Let $\phi(\cdot)$ be Lipschitz, $r_n\epsilon_n \to \infty$, $\epsilon_n \to \infty$, and $r_n\left(\phi\left(\hat{\theta}_n\right) - \phi(\theta_0)\right)$ be asymptotically tight. Define $\mathcal{P}$ to be a class of DGPs such that assumptions 3.1 and 3.2 hold, and for which $\phi(\cdot)$ satisfies either assumption 3.3 or assumptions 3.4 and 3.5. Then for $\forall \epsilon, \delta > 0$ and $x = J_n^{-1}(1 - \tau - \epsilon, P)$,

$$
\sup_{P \in \mathcal{P}} \left( J_{\epsilon_n}(x, P) \leq J_n(x, P) + \epsilon \right) \geq 1 - \delta. \tag{12}
$$

Consequently, the one sided interval (7) has uniform coverage,

$$
\limsup_{n \to \infty} \sup_{P \in \mathcal{P}} P\left( r_n\left( \phi\left(\hat{\theta}_n\right) - \phi(\theta(P))\right) \geq \hat{c}_{1-\tau} \right) \leq \tau. \tag{13}
$$

It turns out that the following additional condition is also satisfied in most of the examples in Fang and Santos (2014) and in Andrews and Soares (2010): For all $v_n \to v$, $|v| = 1$, and all $|a_n| \to 0$, $\phi'_{\theta_0, v}(\cdot) = \lim_{n \to \infty} \phi'_{\theta_0 + |a_n|v_n}(\cdot)$, the limit of the directional derivative along direction $v$, is well defined. It is not required for results in this section, and its only additional implication is that the asymptotic size is exact along local parameter sequences drifting sufficiently slowly: for $\epsilon_n/|\theta_0| \to 0$:

$$
\lim_{n \to \infty} P\left( r_n\left( \phi\left(\hat{\theta}_n\right) - \phi(\theta_0)\right) \geq \hat{c}_{1-\tau} \right) = \tau.
$$

4 **Second Order Numerical Directional Delta Method**

In situations in which the first order delta method limiting distribution is degenerate, the second (or higher) order delta method may provide the necessary nondegenerate large sample approximation. For example, Andrews and Soares (2010) conduct inference using $\phi(\theta) = \sum_{k=1}^{K} \left(\theta_k\right)^2$, which has a first order directional derivative of $\phi'(h) = -\sum_{k=1}^{K} 2\theta_k h_k$. Under the null hypothesis of $\inf_{k=1,\ldots,K} \theta_k \geq 0$, $\phi'(h) = 0$, which leads to a degenerate first order delta method limiting distribution.

We will maintain the assumption that $\phi(\cdot)$ is first order Hadamard differentiable at $\theta_0$. The second order Hadamard directional derivative at $\theta_0$ in the direction $h$
tangential to \(D_0 \subseteq D\) is defined as

\[
\phi''_{\theta_0}(h) \equiv \lim_{t_n \downarrow 0, h_n \to h \in D_0} \frac{\phi(\theta_0 + t_n h_n) - \phi(\theta_0) - t_n \phi'_\theta(\theta_0) h_n}{\frac{1}{2} t_n^2}
\]

(14)

Note that \(\phi''_{\theta_0}(h)\) is continuous with respect to \(h \in D_0\), and it is also positively homogeneous of degree 2: \(\phi''_{\theta_0}(ch) = c^2 \phi''_{\theta_0}(h)\) for all \(c \geq 0\) and \(h \in D_0\). A simple illustrative example is \(\phi''_{\theta_0}(\theta) = (\theta - \theta_0)^2\). For this function, the first order directional derivative is \(\phi'_\theta(h) = -2\theta h\), which is identically zero for \(\theta \geq 0\). The second order directional derivative is \(\phi''_{\theta_0}(h) = 2(h - \theta_0)^2 1(\theta_0 = 0) + 2h^2 1(\theta_0 < 0)\).

The first part of the following theorem is due to Römsch (2005) and Shapiro (2000); in the second part we incorporate the numerical bootstrap principle.  

**Theorem 4.1** (Second Order Directional Delta Method): Suppose \(D\) and \(E\) are Banach spaces and \(\phi: D_0 \subseteq D \mapsto E\) is second order Hadamard directionally differentiable at \(\theta_0\) tangentially to \(D_0\). Let \(\hat{\theta}_n: \{X_i\}_{i=1}^n \mapsto D_0\) be such that for some \(r_n \uparrow \infty\), \(r_n(\hat{\theta}_n - \theta_0) \Rightarrow G_0\) in \(D\) and assume the support of \(G_0\) is included in \(D_0\), a convex set. Then,

\[
r_n^2 \left[ \phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_\theta(\hat{\theta}_n - \theta_0) \right] \Rightarrow \frac{1}{2} \phi''_{\theta_0}(G_0)
\]

(15)

Let \(\epsilon_n \to 0\), \(r_n \epsilon_n \to \infty\), and \(Z_n \overset{p}{\Rightarrow} G_0\). Then if \(\phi'_\theta(h) \equiv 0\), \(\forall h \in D_0\),

\[
\frac{\phi(\hat{\theta}_n + \epsilon_n Z_n) - \phi(\hat{\theta}_n)}{\epsilon_n^2} \overset{p}{\Rightarrow} \frac{1}{2} \phi''_{\theta_0}(G_0).
\]

(16)

The pointwise asymptotic validity of the numerical bootstrap is justified by (16). There are several alternatives for approximating \(\frac{1}{2} \phi''_{\theta_0}(G_0)\). First, the left hand side of (16) can be replaced by \(\hat{\phi}_n''(Z_n)\) where the second order numerical directional derivative can be estimated by

\[
\hat{\phi}_n''(h) \equiv \frac{\phi(\hat{\theta}_n + 2\epsilon_n h) - 2\phi(\hat{\theta}_n + \epsilon_n h) + \phi(\hat{\theta}_n)}{\epsilon_n^2}
\]

(17)

**Theorem 4.2** Under the same conditions as in Theorem 4.1, except without \(\phi'_\theta(h) \equiv 0\), for \(\hat{\phi}_n''(h)\) in (17), \(\hat{\phi}_n''(Z_n) \overset{p}{\Rightarrow} \phi''_{\theta_0}(G_0)\).

---

4Recent independent work by Chen and Fang (2015) also studies inference under first order degeneracy.
Theorem 4.1 applies when \( \phi'_{\theta_0}(h) \equiv 0 \), in which case \( r_n^2 \left( \phi(\hat{\theta}_n) - \phi(\theta_0) \right) \overset{\text{d}}{\to} J \).

By Theorems 4.1 and 4.2, \( \frac{\phi(\hat{\theta}_n + \epsilon_n Z_n) - \phi(\theta_n)}{\epsilon_n^2} \) in (16) and \( \frac{1}{2} \phi''_{\theta_0}(Z_n^*) \) in (17) converge to the same limiting distribution \( J = \frac{1}{2} \phi''_{\theta_0}(G_0) \) under fixed \( \theta_0 \) asymptotics. Under a local drifting sequence of parameters \( \theta_n \) where \( r_n (\theta_n - \theta_0) \to c \) for \( ||c|| < \infty \), they also converge to the same distribution:

\[
\frac{1}{r_n^2} \left( \phi \left( \frac{1}{r_n} (r_n (\theta_n - \theta_0) + Z_n) \right) - \phi \left( \frac{1}{r_n} (r_n (\theta_n - \theta_0)) \right) - \phi(\theta_0) \right) \overset{\text{d}}{\to} \frac{1}{2} \phi''_{\theta_0}(c + G_0) - \frac{1}{2} \phi''_{\theta_0}(c).
\]

The equalities follow from \( Z_n = r_n \left( \hat{\theta}_n - \theta_n \right), r_n (\theta_n - \theta_0) + Z_n \overset{\text{d}}{\to} c + G_0, r_n (\theta_n - \theta_0) \overset{\text{d}}{\to} c \), and the definition of the second order delta method.

The behaviors of \( \frac{1}{2} \phi''_{\theta_0}(Z_n^*) \) and \( \frac{\phi(\hat{\theta}_n + \epsilon_n Z_n^*) - \phi(\theta_n)}{\epsilon_n^2} \) differ under a more distant local drifting sequence of parameters \( \frac{\theta_n - \theta_0}{\epsilon_n} \to c \), when \( 0 < ||c|| < \infty \). While this has no impact on asymptotic test consistency and size control, since now \( |r_n (\theta_n - \theta_0)| \to \infty \), it does imply different finite sample behaviors. On the one hand,

\[
\frac{1}{\epsilon_n^2} \left( \phi(\hat{\theta}_n + \epsilon_n Z_n^*) - \phi(\hat{\theta}_n) \right) = \frac{1}{\epsilon_n^2} \left[ \phi \left( \epsilon_n \left( \theta_n - \theta_0 + \frac{Z_n^*}{r_n \epsilon_n} \right) \right) - \phi(\theta_0) \right] - \frac{1}{\epsilon_n^2} \left[ \phi \left( \epsilon_n \left( \theta_n - \theta_0 + \frac{Z_n^*}{r_n \epsilon_n} \right) \right) - \phi(\theta_0) \right] \overset{\text{d}}{\to} \frac{1}{2} \phi''_{\theta_0}(c + G_0) - \frac{1}{2} \phi''_{\theta_0}(c).
\]

On the other hand, for (17),

\[
\frac{1}{2} \phi''_{\theta_0}(Z_n^*) = \frac{1}{2} \frac{1}{\epsilon_n^2} \left[ \phi \left( \epsilon_n \left( \theta_n - \theta_0 + \frac{Z_n^*}{r_n \epsilon_n} + 2Z_n^* \right) \right) - \phi(\theta_0) \right] - \frac{1}{\epsilon_n^2} \left[ \phi \left( \epsilon_n \left( \theta_n - \theta_0 + \frac{Z_n^*}{r_n \epsilon_n} \right) \right) - \phi(\theta_0) \right] + \frac{1}{2} \frac{1}{\epsilon_n^2} \left[ \phi \left( \epsilon_n \left( \theta_n - \theta_0 + \frac{Z_n^*}{r_n \epsilon_n} \right) \right) - \phi(\theta_0) \right] \overset{\text{d}}{\to} \frac{1}{2} \phi''_{\theta_0}(c + 2G_0) - \frac{1}{2} \phi''_{\theta_0}(c + G_0) + \frac{1}{4} \phi''_{\theta_0}(c).
\]

4.1 Application to Partially Identified Models: General Case

The second order delta method can be used to perform hypothesis tests and conduct inference in moment inequalities models. Let \( \mathcal{B} \) be the parameter space for a partially
identified parameter $\beta_0$ defined by a set of $K$ moment inequalities $Pg(\cdot, \beta_0) \geq 0$. We are interested in testing

\[ H_0 : \inf_{\beta \in \mathbb{B}} \min_{k=1}^K Pg_k(\cdot, \beta) \geq 0 \]
\[ H_1 : \inf_{\beta \in \mathbb{B}} \min_{k=1}^K Pg_k(\cdot, \beta) < 0 \]

For example, in Bugni et al. (2014), $\mathbb{B} = \mathbb{B}(\gamma) = \{ \beta : f(\beta) = \gamma \}$. In Bugni et al. (2015), $\mathbb{B}$ corresponds to the entire parameter space $\mathbb{B}$. In Andrews and Soares (2010), $\mathbb{B} = \beta^*$ corresponds to a singleton parameter value for a pointwise testing procedure.

Redefining $\theta_0(\beta) = Pg(\cdot, \beta)$, the hypotheses can be converted into

\[ H_0 : \inf_{\beta \in \mathbb{B}} S(\theta_0(\beta)) = 0 \]
\[ H_1 : \inf_{\beta \in \mathbb{B}} S(\theta_0(\beta)) > 0 \]

where $S(\cdot)$ is a nonincreasing and continuous function that satisfies $S(\theta) \geq 0$ for all $\theta$, $S(\theta) = 0$ for all $\theta \geq 0$, and $S(c\theta) = c^\rho S(\theta)$ for either $\rho = 1$ or 2. Define

\[ \phi(\theta) \equiv \inf_{\beta \in \mathbb{B}} S(\theta(\beta)) = (f \circ S)(\theta) \]
\[ f(S) = \inf_{\beta \in \mathbb{B}} S(\beta) \]

We would like to obtain the limiting distribution of $\sqrt{n} \left( \phi(\hat{\theta}) - \phi(\theta_0) \right) = \inf_{\beta \in \mathbb{B}} S(\sqrt{n}/\hat{\theta}_n(\beta)) - \inf_{\beta \in \mathbb{B}} S(\sqrt{n}/\theta_0(\beta))$ under $H_0$. We can compute the first order directional derivative using the chain rule:

\[ \phi'_\theta(h) = (f \circ S)'_\theta(h) = (f'_{S(\theta)} \circ S)'_\theta(h) \]
\[ f'_S(h) = \inf_{\beta \in \mathbb{B}_0(S)} h(\beta) \]
\[ \mathbb{B}_0(S) = \left\{ \beta \in \mathbb{B} : S(\beta) = \inf_{\beta \in \mathbb{B}} S(\beta) \right\} \]
\[ \phi'_\theta(h) = \inf_{\beta \in \mathbb{B}_0(S(\theta))} S'_{\theta(\beta)}(h(\beta)) \]
If \( S(\theta(\beta)) = \sum_{j=1}^{J} \theta_j(\beta)^{-} \), then \( \rho = 1 \) and

\[
S'_\theta(h(\beta)) = \sum_{j=1}^{J} \left\{ -h_j(\beta) 1 (\theta_j(\beta)^- < 0) + h_j(\beta)^- 1 (\theta_j(\beta)^- = 0) \right\}
\]

It follows that

\[
\phi'_\theta(h) = \inf_{\beta \in B_0(S(\theta))} \sum_{j=1}^{J} \left\{ -h_j(\beta) 1 (\theta_j(\beta)^- < 0) + h_j(\beta)^- 1 (\theta_j(\beta)^- = 0) \right\}
\]

Therefore, for \( \sqrt{n}(\hat{\theta}_n(\beta) - \theta_0(\beta)) \equiv \sqrt{n}(P_n - P)g(\cdot, \beta) \rightsquigarrow G_0 \),

\[
\inf_{\beta \in B_0(S(\theta))} \sum_{j=1}^{J} \left\{ -G_{0j}(\beta) 1 (\theta_{0j}(\beta)^- < 0) + G_{0j}(\beta)^- 1 (\theta_{0j}(\beta)^- = 0) \right\}
\]

If \( S(\theta(\beta)) = \sum_{j=1}^{J}(\theta_j(\beta)^-)^2 \), then \( \rho = 2 \) and

\[
S''_\theta(h(\beta)) = -2 \sum_{j=1}^{J} \theta_j(\beta)^- h_j(\beta)
\]

\[
\phi''_\theta(h) = \inf_{\beta \in B_0(S(\theta))} \left\{ -2 \sum_{j=1}^{J} \theta_j(\beta)^- h_j(\beta) \right\}
\]

The chain rule for second order directional derivatives gives us

\[
\phi''_\theta(h) = (f \circ S)''' \theta(h) = f''_S(h) (S'_\theta(h), S''_\theta(h))
\]

\[
S''_\theta(h) = 2 \sum_{j=1}^{J} \left( (h_j^-)^2 1 (\theta_j(\beta) = 0) + h_j^2 1 (\theta_j(\beta) < 0) \right)
\]

Using equations (4.426), (4.429), and (4.430) of Bonnans and Shapiro (2013), we can obtain the second order directional derivative of \( f \):

\[
f''_S(\eta, w) = \inf_{\beta \in B} \{ w(\beta) - \tau_{S,\eta}(\beta) \}\]
\[
\tau_{S,\eta}(\beta) = \begin{cases} 
0, & \text{if } \beta \in \text{interior}(B_0(S)) \\
\limsup_{\beta' \to \beta} \frac{(\eta(\beta) - \eta(\beta'))^+}{2(S(\beta') - \inf_{b \in B} S(b))} & \text{if } \beta \in \text{boundary}(B_0(S)) \\
S(\beta') > \inf_{b \in B} S(b) & \\
-\infty & \text{otherwise}
\end{cases}
\]

where \(\eta(\beta) = \inf_{b \in B_0(S)} \eta(b)\).

Therefore, for \(\rho = 2\),

\[
\inf_{\beta \in B} S\left(\sqrt{n}\hat{\theta}_n(\beta)\right) - \inf_{\beta \in B} S\left(\sqrt{n}\theta_0(\beta)\right) \sim \frac{1}{2} \inf_{\beta \in B} \left\{ S_{\theta_0}(\beta) (G_0) - \tau_{S,\theta_0,\theta_0}(G_0)(\beta) \right\}
\]

Our test rejects when \(\sqrt{n}\phi(\hat{\theta}_n) > \hat{c}_{1-\alpha}\), where \(\hat{c}_{1-\alpha}\) is the \(1 - \alpha\) quantile of one of the following three ways of estimating the second order numerical derivative (the last way is only applicable for \(\rho = 2\)):

1. Andrews and Soares (2010): \(\frac{1}{\epsilon_n} \phi(\hat{\theta}_n + \epsilon_n Z_n^*)\)

2. Numerical Bootstrap: \(\frac{1}{\epsilon_n} \left( \phi(\hat{\theta}_n + \epsilon_n Z_n^*) - \phi(\hat{\theta}_n) \right)\)

3. Numerical Second Order Derivative: \(\frac{1}{2} \hat{\phi}''(Z_n^*) = \frac{1}{2} \frac{\phi(\hat{\theta}_n + 2\epsilon_n Z_n^*) - 2\phi(\hat{\theta}_n + \epsilon_n Z_n^*) + \phi(\hat{\theta}_n)}{\epsilon_n^2}\)

where \(Z_n^* \stackrel{P}{\sim} G_0\). For a finite sample power comparison, see section 12.

For the Bugni et al. (2014) example, we can construct a nominal \(1 - \alpha\) confidence set using test statistic inversion: \(C = \{ \gamma : \inf_{\beta \in B(\gamma)} S\left(\sqrt{n}P_n g(\cdot; \beta)\right) \leq \hat{c}_{1-\alpha} \}\).

In empirical work, researchers typically use a recentered form of the test statistic \(\inf_{\beta \in B} S\left(\sqrt{n}\hat{\theta}_n(\beta)\right) - \inf_{\beta \in B} S\left(\sqrt{n}\hat{\theta}_n(\beta)\right)\) because it results in a confidence set that is non-empty with probability one.

5 A generalized numerical bootstrap method

In this section we generalize the numerical delta method to what we will call the numerical bootstrap. It can be useful in some situations where the regular bootstrap fails and subsampling (Politis et al. (1999)) works. Subsampling is more generally applicable than the numerical bootstrap, but in situations where both are applicable
(but the standard bootstrap is not), the numerical bootstrap might offer a more
accurate approximation to the limiting distribution.

To motivate, most estimators and test statistics can be written as a function of
the empirical distribution \( \theta(P_n) \) with a population analog \( \theta(P) \). Typically, for an
increasing function \( a(n) \) of the sample size \( n \), and for a limiting distribution \( J \) (which
can depend on \( P \)): \( a(n)(\theta(P_n) - \theta(P)) \sim J \). This can be rewritten as

\[
a(n) \left( \theta \left( P + \frac{1}{\sqrt{n}} \sqrt{n} (P_n - P) \right) - \theta(P) \right) \sim J.
\]

Since it is often times the case that \( \hat{G}_n = \sqrt{n} (P_n - P) \sim \mathcal{G}_0 \) where \( \mathcal{G}_0 \) is a properly
defined Brownian bridge, we also expect that

\[
a(n) \left( \theta \left( P + \frac{1}{\sqrt{n}} \mathcal{G}_0 \right) - \theta(P) \right) \sim J.
\]

If we take \( \epsilon_n = \frac{1}{\sqrt{n}} \), so that \( a(n) = a \left( \sqrt{n}^2 \right) \) is replaced by \( a \left( \frac{1}{\epsilon_n^2} \right) \), then we also
anticipate that for other \( \epsilon_n \to 0 \),

\[
a \left( \frac{1}{\epsilon_n^2} \right) \left( \theta \left( P + \epsilon_n \mathcal{G}_0 \right) - \theta(P) \right) \sim J.
\]

The goal is to provide a consistent estimate of \( J \), which approximates the left hand
class above. To obtain such a consistent estimate we need to estimate the unknown \( P \)
and \( \mathcal{G}_0 \). Intuitively, \( P \) can be estimated by \( P_n \), and \( \mathcal{G}_0 \) can be consistently estimated
by the bootstrapped empirical process \( \hat{G}_n^* = \sqrt{n} (P_n^* - P_n) \). A popular choice for \( \hat{G}_n^* \)
is the multinomial bootstrap in which \( \hat{G}_n^* = \sqrt{n} (P_n^* - P_n) \) and \( P_n^* = \frac{1}{n} \sum_{i=1}^{n} M_{ni} \delta_i \),
where \( \delta_i \) is the point mass on observation \( i \), and \( M_{ni} \), \( i = 1, \ldots, n \) is a multinomial
distribution with parameters \( (n^{-1}, n^{-1}, \ldots, n^{-1}) \). Another common choice for \( \hat{G}_n^* \)
is the Wild bootstrap, where \( \hat{G}_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\xi_i - \bar{\xi}) \delta_i \) for \( \bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i \) and \( \xi_i \) are i.i.d
variables with variance 1 and finite 3rd moment. Other forms of \( \hat{G}_n^* \) that consistently
estimate \( \mathcal{G}_0 \) can also be used. For example, we can use \( \hat{G}_n^* = \sqrt{m_n} (P_{mn}^* - P_n) \) where
\( m_n \to \infty \) as \( n \to \infty \), and \( P_{mn}^* \) is a multinomial i.i.d sample from \( P_n \) of size \( m_n \).
While consistency only requires \( m_n \to \infty \), we follow the convention of the bootstrap
literature and focus on \( m_n = n \). Other exchangeable bootstrap schemes in \textit{van der Vaart and Wellner (1996)} (Chapter 3.6) can also be used.

We know that \( \hat{G}_n^* \) converges in distribution to \( \mathcal{G}_1 \) both conditionally on the sample
in probability, and unconditionally, where \( G_1 \) is an independent and identical copy of \( G_0 \). To offset the noise of estimating \( P \) with \( P_n \), the step size parameter \( \epsilon_n \) is chosen such that \( \sqrt{n \epsilon_n} \to \infty \). Therefore, we propose a numerical bootstrap method that estimates \( J \) with

\[
\hat{J}_n^* = a \left( \frac{1}{\epsilon_n^2} \right) \left( \theta \left( P_n + \epsilon_n \hat{G}_n^* \right) - \theta (P_n) \right)
\]

To see why the numerical bootstrap might work when bootstrap fails, note that

\[
\hat{J}_n^* = a \left( \frac{1}{\epsilon_n^2} \right) \left( \theta \left( P_n + \epsilon_n \left[ \hat{G}_n^* + \frac{P_n - P}{\epsilon_n} \right] \right) - \theta (P) \right) - a \left( \frac{1}{\epsilon_n^2} \right) (\theta (P_n) - \theta (P))
\]

In the above, we rewrite the second term as

\[
a \left( \frac{1}{\epsilon_n^2} \right) (\theta (P_n) - \theta (P)) = \frac{1}{a(n)} a \left( \frac{1}{\epsilon_n^2} \right) a(n) (\theta (P_n) - \theta (P)).
\]

Since \( a(n) (\theta (P_n) - \theta (P)) \rightsquigarrow J \) and typically \( \frac{1}{a(n)} a \left( \frac{1}{\epsilon_n^2} \right) \to 0 \) (e.g. when \( a(n) = n^\gamma \)) as \( n \epsilon_n^2 \to \infty \), the second term vanishes asymptotically:

\[
a \left( \frac{1}{\epsilon_n^2} \right) (\theta (P_n) - \theta (P)) = o_P(1).
\]

Using the conditional weak convergence notation defined in Kosorok (2007), \( \hat{G}_n^* \rightsquigarrow \mathcal{G}_1 \) in the first term of \( \hat{J}_n^* \). Additionally, since \( \sqrt{n \epsilon_n} \to \infty \), heuristically we expect that

\[
\frac{P_n - P}{\epsilon_n} = \sqrt{n} \frac{P_n - P}{\sqrt{n \epsilon_n}} \approx \frac{\mathcal{G}_0}{\sqrt{n \epsilon_n}} \rightsquigarrow 0.
\]

Therefore, since \( \mathcal{G}_1 \) has the same distribution as \( \mathcal{G}_0 \), we also expect that

\[
\hat{J}_n^* \approx a \left( \frac{1}{\epsilon_n^2} \right) \left( \theta (P + \epsilon_n \mathcal{G}_1) - \theta (P) \right) \overset{\mathbb{W}}{\rightsquigarrow} J.
\]

5.1 Comparison of Numerical Bootstrap with Subsampling

In situations where both subsampling and the numerical bootstrap method can be used, the numerical bootstrap can potentially offer a more accurate approximation
than subsampling to the limiting distribution. Recall that subsampling (Politis et al. (1999)) approximates the limit distribution $\mathcal{J}$ using the finite sample distribution of $a(b)(\theta(P_b) - \theta(P_n))$ which in large samples is close to $a(b)(\theta(P_b) - \theta(P))$ whenever $a(b)(\theta(P_n) - \theta(P)) = o_P(1)$. In turn, as $b \to \infty$,

$$a(b)(\theta(P_b) - \theta(P)) \rightsquigarrow \mathcal{J}$$

To compare subsampling to the numerical bootstrap, write the subsampling distribution as

$$a(b)(\theta(P_b) - \theta(P_n)) = a(b)\left(\theta\left(P_n + \frac{1}{\sqrt{b}}\sqrt{b}(P_b - P_n)\right) - \theta(P_n)\right)$$

In the numerical bootstrap setup, subsampling is essentially using $\epsilon_n = \frac{1}{\sqrt{b}}$ as the step size and using $\sqrt{b}(P_b - P_n)$ to estimate $\mathcal{G}_0$ based on subsamples of size $b$. The numerical bootstrap method is different and instead uses $\hat{G}_n^\ast$ to estimate $\mathcal{G}_0$ based on the entire sample of size $n$. In addition to using either the bootstrap $\hat{G}_n^\ast \equiv \sqrt{n}(P_n^\ast - P_n)$ or subsampling $\sqrt{b}(P_b - P_n)$, $\mathcal{G}_0$ can also be approximated by a joint normal distribution in finite dimensional situations.

While a formal analysis is beyond the current scope and left for future analysis, the following presents heuristic arguments based on simple finite dimensional examples. In some cases, $\phi(\cdot)$ may be directionally linear. For example, let $\phi(\mu) = \mu^-$ and $X_i \overset{i.i.d.}{\sim} (\mu, \sigma^2)$, where $\mu$ is a fixed parameter. Consider approximating the distribution of $\sqrt{n}(\bar{X}^- - \mu^-)$ using $Z_n^\ast = \left(\frac{\bar{X}^\ast + \epsilon_n(\bar{X}^\ast - \bar{X})}{\epsilon_n}\right)^- - \left(\frac{\bar{X}}{\epsilon_n}\right)^-$. First let $\mu = 0$ and $x > 0$. Then by standard Berry-Esseen arguments,

$$P\left(\sqrt{n}\bar{X}^- \leq x\right) = P\left(\sqrt{n}\bar{X} \geq -x\right) = \Phi\left(\frac{x}{\sigma}\right) + O\left(n^{-1/2}\right).$$
Also let \( X_n = (X_1, \ldots, X_n) \). Then, since \( \bar{X}/\epsilon_n = O_P\left(\frac{1}{\sqrt{n \epsilon}}\right) \), consider:

\[
P\left( \hat{\phi}'_n \left( \bar{Z}_n^* \right) \leq x \mid X_n \right) \\
= P\left( \frac{\bar{X} + \sqrt{n}(\bar{X}^* - \bar{X})}{\epsilon_n} \leq x \mid X_n \right) \\
= P\left( \sqrt{n}(\bar{X}^* - \bar{X}) \geq -\left(x + \frac{\bar{X}}{\epsilon_n}\right) \right) \\
= \Phi\left( \frac{1}{\sigma} \left( x + \frac{\bar{X}}{\epsilon_n} \right) \right) + O_P\left( \frac{1}{\sqrt{n}} \right) = \Phi\left( \frac{x}{\sigma} \right) + O_P\left( \frac{1}{\sqrt{n}} \right) + O_P\left( \frac{1}{\sqrt{n}} \right)
\]

Subsampling corresponds to \( \epsilon_n = 1/\sqrt{b} \) and replacing \( \sqrt{n}(\bar{X}^* - \bar{X}) \) with \( \sqrt{b} (\bar{X}_b - \bar{X}) \),

\[
P\left( \left( \sqrt{b} (\bar{X}_b) \right)^{-} \leq x \right) = \Phi\left( \frac{x}{\sqrt{b}} \right) + O\left( \sqrt{n} \right) + O\left( b^{-1/2} \right)
\]

There is also an error of \( O_P\left( \frac{1}{\sqrt{n}} \right) \) from approximating the distribution of \( \sqrt{b} (\bar{X}_b - \bar{X}) \) using \( \binom{n}{b} \) sub-block estimates \( \bar{X}_{bi} \). An optimal choice of \( b \) will then be \( b = O(n^{1/2}) \), resulting in an error of \( O_P\left( n^{-1/4} \right) \). On the other hand, the error in numerical bootstrap can be made close to \( O_P\left( n^{-1/2} \right) \) when \( \epsilon_n \to 0 \) slowly.

Next let \( \mu < 0 \). Then with probability converging to 1, the numerical bootstrap is identical to bootstrap, which has an error of \( O_P\left( n^{-1/2} \right) \) regardless of how \( \epsilon_n \to 0 \), while the subsampling error can still be \( O_P\left( n^{-1/4} \right) \).

Consider now the more general nonlinear case of approximating the limiting distribution of \( \sqrt{n} \left( \phi(\hat{\theta}_n) - \phi(\theta_0) \right) \) using either the distribution of the numerical bootstrap \( \hat{\phi}'_n \left( \bar{Z}_n^* \right) \) or subsampling \( \sqrt{b} \left( \phi(\hat{\theta}_b) - \phi(\hat{\theta}_n) \right) \), where \( Z_n^* = \sqrt{n} \left( \hat{\theta}_n^* - \hat{\theta}_n \right) \) is the bootstrapped statistic using samples of size \( n \) and \( \hat{\theta}_b \) is the estimate of \( \theta_0 \) using samples of size \( b \). If \( \phi(\theta) \) is twice Hadamard directionally differentiable at \( \theta_0 \) with directional derivatives \( \phi'_{\theta_0} \) and \( \phi''_{\theta_0} \) that can be continuously extended to \( \mathbb{D} \), then for some \( h_n = O_P(1) \),

\[
\phi(\theta_0 + \epsilon_n h_n) = \phi(\theta_0) + \epsilon_n \phi'_{\theta_0}(h_n) + \frac{1}{2} \epsilon_n^2 \phi''_{\theta_0}(h_n) + o_P(\epsilon_n^2)
\]  

(18)
On the one hand, when $h_n = Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n}$ in (18),

$$
\phi \left( \hat{\theta}_n + \epsilon_n Z_n^* \right) = \phi (\theta_0) + \epsilon_n \phi'_{\theta_0} \left( Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) + \frac{1}{2} \epsilon_n^2 \phi''_{\theta_0} \left( Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) + o_p (\epsilon_n^2)
$$

On the other hand, taking $h_n = \frac{\hat{\theta}_n - \theta_0}{\epsilon_n}$ in (18), we can write

$$
\phi \left( \hat{\theta}_n \right) = \phi (\theta_0) + \epsilon_n \phi'_{\theta_0} \left( \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) + \frac{1}{2} \epsilon_n^2 \phi''_{\theta_0} \left( \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) + o_p (\epsilon_n^2)
$$

Let $Z_n^* \xrightarrow{p} G_0$. As long as $\phi'_{\theta_0} (\cdot)$ is Lipschitz, we can bound

$$
\phi'_{\theta_0} \left( Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) - \phi'_{\theta_0} (G_0) - \phi'_{\theta_0} \left( \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) \leq C \left\| Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} - G_0 \right\| + C \left\| \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right\|
$$

$$
\leq O_P (Z_n^* - G_0) + O_P \left( \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right)
$$

If we assume that $O_P (Z_n^* - G_0) = O_P \left( \frac{1}{\sqrt{n}} \right)$, then

$$
\phi' \left( Z_n^* \right) - \phi'_{\theta_0} (G_0) = \frac{\phi \left( \hat{\theta}_n + \epsilon_n Z_n^* \right) - \phi (\theta_0)}{\epsilon_n} - \phi'_{\theta_0} (G_0)
$$

$$
= O_P (Z_n^* - G_0) + O_P \left( \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) + \frac{1}{2} \epsilon_n \phi''_{\theta_0} \left( Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) - \frac{1}{2} \epsilon_n \phi''_{\theta_0} \left( \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) + o_p (\epsilon_n^2)
$$

$$
= O_P \left( \frac{1}{\sqrt{n}} \right) + O_P \left( \frac{1}{\epsilon_n \sqrt{n}} \right) + O_p (\epsilon_n) \tag{19}
$$

When $\phi'_{\theta_0} (\cdot)$ is not a linear function and the second derivative is nonzero, $\phi' \left( Z_n^* \right) - \phi'_{\theta_0} (G_0) = O_p \left( \frac{1}{\epsilon_n \sqrt{n}} \right) + O_p (\epsilon_n)$. The optimal choice of $\epsilon_n$ equates the two terms, which leads to $\epsilon_n = O \left( n^{-1/4} \right)$. Note however that in the differentiable case when $\phi'_{\theta_0} (\cdot)$ is a linear function that is not degenerate at $\theta_0$ and the second derivative is nonzero, then $\phi' \left( Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) - \phi'_{\theta_0} (G_0) = \phi' \left( Z_n^* - G_0 \right) = O_P \left( \frac{1}{\sqrt{n}} \right)$ so $\phi' \left( Z_n^* \right) - \phi'_{\theta_0} (G_0) = O_p (\epsilon_n) + O_p \left( \frac{1}{\sqrt{n}} \right)$. An optimal choice of $\epsilon_n$ would then be $O \left( \frac{1}{\sqrt{n}} \right)$ and the error would be $O_p \left( \frac{1}{\sqrt{n}} \right)$. Intuitively, the linearity of the first derivative removes the first order effect of the $\epsilon_n$ and preserves only the second or
higher order effects. When the second order derivative is zero, \( \hat{\phi}'_n (Z^*_n) - \phi'_{\theta_0} (G_0) = O_p \left( \frac{1}{\sqrt{n}} \right) + O_p \left( \frac{1}{\epsilon_n \sqrt{n}} \right) \). Then the optimal choice of \( \epsilon_n \) is close to 1.

To summarize, in differentiable situations, the optimal magnitude of \( \epsilon_n \) is \( O \left( \frac{1}{\sqrt{n}} \right) \) as in the bootstrap. In directionally linear situations, the optimal \( \epsilon_n \) should converge to zero very slowly. The more general case of smooth directional differentiability is more difficult because the choice of \( \epsilon_n \) depends on whether \( \phi (\cdot) \) is differentiable or directionally differentiable at \( \theta_0 \) which is unknown. Note that these discussions are only based on pointwise asymptotics. The choice of \( \epsilon_n \) might also depend on considerations about local drifting sequences of \( \mu \) or \( \theta \).

Subsampling uses the distribution of \( \sqrt{b} \left( \phi(\hat{\theta}_b) - \phi(\hat{\theta}_n) \right) \) to estimate \( \phi'_{\theta_0} (G_0) \), which effectively replaces \( \epsilon_n \) with \( 1/\sqrt{b} \) and \( Z^*_n \) with \( Z^*_b = \sqrt{b} (\hat{\theta}_b - \hat{\theta}_n) \) in (19). Its approximation error thus takes the same form with these replacements:

\[
\sqrt{b} \left( \phi(\hat{\theta}_b) - \phi(\hat{\theta}_n) \right) - \phi'_{\theta_0} (G_0) = O_p \left( \frac{1}{\sqrt{b}} \right) + O_p \left( \sqrt{\frac{b}{n}} \right).
\]

The optimal choice of \( b \) is therefore \( b = O \left( n^{1/2} \right) \), which leads to an error on the order of \( n^{-1/4} \). If \( \phi'_{\theta_0} (\cdot) \) were a linear function that is not degenerate at \( \theta_0 \) or \( \phi(\theta) \) had zero second order or higher derivatives, the error would still be \( O_p \left( n^{-1/4} \right) \) because there is an additional error of \( O_p \left( \frac{b}{\sqrt{n}} \right) \) introduced when the distribution of \( \sqrt{b} (\hat{\theta}_b - \theta_0) \) needs to be estimated using the empirical distribution of \( \sqrt{b} (\hat{\theta}_{b,i} - \hat{\theta}_n) \) over \( i = 1, \ldots, \binom{n}{b} \) sub-blocks. In other words, \( O_p \left( Z^*_b - G_0 \right) = O_p \left( \frac{1}{\sqrt{b}} \right) + O_p \left( \sqrt{\frac{b}{n}} \right) \).

The presence of the error of \( O_p \left( \sqrt{\frac{b}{n}} \right) \) is also demonstrated in Theorem 1 of Bertail et al. (1997) and Theorem 3 of Babu and Singh (1985). Therefore, the numerical bootstrap should outperform subsampling when the first derivative is linear and the second order derivative is nonzero, or when the second order derivative is zero.

6 Consistency of Numerical Bootstrap for M-estimators

In this section, we demonstrate the asymptotic consistency of the numerical bootstrap for a class of M-estimators \( \hat{\theta}_n \) that converge at rate \( n^{\gamma} \) for some \( \gamma \in \left( \frac{1}{4}, 1 \right) \). Our proofs in this section assume that the econometrician knows \( \gamma \), but in practice, we
can estimate an unknown $\gamma$ using methods described in section 11.

$$\hat{\theta}_n \equiv \arg\max_{\theta \in \Theta} P_n \pi(\cdot, \theta) = \frac{1}{n} \sum_{i=1}^{n} \pi(z_i, \theta).$$

We approximate the limiting distribution of $n^\gamma(\hat{\theta}_n - \theta_0)$ using the finite sample distribution of $\epsilon_n^{-2\gamma} \left( \hat{\theta}_n^* - \hat{\theta}_n \right)$, where $\hat{\theta}_n^* \equiv \arg\max_{\theta \in \Theta} Z_n^* \pi(\cdot, \theta)$, and $Z_n^* = P_n + \epsilon_n \hat{G}_n^*$ is a linear combination between the empirical distribution and the bootstrapped empirical process. For example, when $\hat{G}_n^*$ is the multinomial bootstrap, for each bootstrap sample $z_i^*, i = 1, \ldots, n,$

$$\hat{\theta}_n^* = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \pi(z_i, \theta) + \epsilon_n \sqrt{n^{-1}} \sum_{i=1}^{n} \left( Z_i^* - \pi(z_i, \theta) \right).$$

On the other hand, when $\hat{G}_n^*$ is the Wild bootstrap,

$$\hat{\theta}_n^* = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \pi(z_i, \theta) + \epsilon_n \sqrt{n^{-1}} \sum_{i=1}^{n} \left( \xi_i - \bar{\xi} \right) \pi(z_i, \theta).$$

In the following theorem, we show that for a suitable choice of the step size $\epsilon_n$, $n^\gamma(\hat{\theta}_n - \theta_0)$ and $\epsilon_n^{-2\gamma} \left( \hat{\theta}_n^* - \hat{\theta}_n \right)$ converge to the same limiting distribution for a large class of estimators that includes the typical $\sqrt{n}$ consistent estimators like OLS and IV as well as $n^{1/3}$ consistent estimators like the maximum score estimator studied in Manski (1975), Kim and Pollard (1990) and Abrevaya and Huang (2005). Another valid bootstrap method for the maximum score estimator is developed in Seijo and Sen (2011). Let $\hat{g}_n^* = o_P(1)$ (hence also $o_P(1)$) if the law of $\hat{X}_n^*$ is governed by $P_n$ and if $P_n(|\hat{X}_n^*| > \epsilon) = o_P(1)$ for all $\epsilon > 0$. Also define $M_n^* = O_P(1)$ (hence also $O_P(1)$) if $\lim_{m \to \infty} \limsup_{n \to \infty} P(P_n(M_n^* > m > \epsilon) \to 0, \forall \epsilon > 0$.

**Theorem 6.1 (Consistency of Numerical Bootstrap for M-estimators):** Define $g(\cdot, \theta) \equiv \pi(\cdot, \theta) - \pi(\cdot, \theta_0)$. Suppose the following conditions are satisfied for some $\rho \in (0, 3/2)$ and for $\gamma \equiv \frac{1}{2(3\rho - 2)}$.

(i) $P_n g(\cdot, \hat{\theta}_n) = \sup_{\theta \in \Theta} P_n g(\cdot, \theta) - o_P(n^{-2\gamma})$ and $Z_n^* g(\cdot, \hat{\theta}_n^*) = \sup_{\theta \in \Theta} Z_n^* g(\cdot, \theta) - o_P(\epsilon_n^{4\gamma})$

(ii) $\hat{\theta}_n \overset{P}{\to} \theta_0$ and $\hat{\theta}_n^* - \hat{\theta}_n = o_P(1)$.

(iii) $\theta_0$ is an interior point of $\Theta \in \mathbb{R}^d$. 

24
(iv) The class of functions $\mathcal{G}_R = \{g(\cdot, \theta) : |\theta - \theta_0| \leq R\}$ is uniformly manageable with envelope function $G_R(\cdot) \equiv \sup_{g \in \mathcal{G}_R} |g(\cdot)|$.

(v) $P g(\cdot, \theta)$ is twice differentiable at $\theta_0$ with negative definite Hessian matrix $-H$.

(vi) $\Sigma_{\rho}(s, t) = \lim_{\alpha \to \infty} \alpha^{2\rho} P g(\cdot, \theta_0 + \frac{\alpha}{\rho}) g(\cdot, \theta_0 + \frac{\alpha}{\rho})$ exists for each $s, t$ in $\mathbb{R}^d$.

(vii) $\lim_{\alpha \to \infty} \alpha^{2\rho} P g(\cdot, \theta_0 + \frac{\alpha}{\rho})^2 \{ |g(\cdot, \theta_0 + \frac{\alpha}{\rho})| > \epsilon \alpha^{2(1-\rho)} \} = 0$ for each $\epsilon > 0$ and $t \in \mathbb{R}^d$.

(viii) There exists a $R_0 > 0$ such that $PG_R^2 = O(R^{2\rho})$ for all $R \leq R_0$.

(ix) $\sqrt{n} \epsilon_n \to \infty$.

(x) For some $\eta > 0$, there exists a $K$ such that $PG_R^2 \{ G_R > K \} < \eta R^{2\rho}$ for $R \to 0$.

(xi) $P \{ g(\cdot, \theta_1) - g(\cdot, \theta_2) \} = O(\{ \theta_1 - \theta_2 \}^{2\rho})$ for $|\theta_1 - \theta_2| \to 0$.

Then $\hat{\theta}_n - \theta_0 = O_p(n^{-\gamma})$ and $\hat{\theta}_n^* - \theta_0 = O_p^*(\epsilon_n^{-1})$. Furthermore, for $Z_0(h)$ a mean zero Gaussian process with covariance kernel $\Sigma_{\rho}$ and nondegenerate increments,

$$n^{-1/2} \{ \hat{\theta}_n - \theta_0 \} \sim J \equiv \arg\max_{h \in \mathbb{R}} Z_0(h) - \frac{1}{2} h'Hh$$

and

$$Z^*_n \equiv \epsilon_n^{-2\gamma} \{ \hat{\theta}_n - \hat{\theta}_n \} \xrightarrow{p} J \quad \text{and} \quad Z_n^* \sim J.$$

The assumptions above are modeled after Kim and Pollard (1990) but generalized so that results for both the $\sqrt{n}$ and $n^{1/3}$ cases can be stated concisely.

To explain the intuition for the above theorem, note that for $\hat{h}_n = n^{-1/2} (\hat{\theta}_n - \theta_0)$,

$$\hat{h}_n = \arg\max_{h \in n^{-1} (\theta_0 - \theta_0)} n^{2\gamma} P_n g \left( Z; \theta_0 + \frac{h}{n^{\gamma}} \right) = n^{2\gamma - \frac{1}{2}} \sqrt{n} (P_n - P) g \left( \cdot; \theta_0 + \frac{h}{n^{\gamma}} \right) + n^{2\gamma} P \left( \cdot; \theta_0 + \frac{h}{n^{\gamma}} \right). \quad (21)$$

Under the stated conditions, $n^{2\gamma} P \left( \cdot; \theta_0 + \frac{h}{n^{\gamma}} \right) \to -\frac{1}{2} h'Hh$, and

$$n^{2\gamma - \frac{1}{2}} \sqrt{n} (P_n - P) g \left( \cdot; \theta_0 + \frac{h}{n^{\gamma}} \right) = n^{\rho} \mathcal{G}_n g \left( \cdot; \theta_0 + \frac{h}{n^{\gamma}} \right) \sim Z_0(h).$$
The numerical bootstrap seeks to approximate the limiting distribution \( \mathcal{J} \) with the distribution of

\[
\epsilon_n^{-2\gamma} \left( \hat{\theta}_n^* - \hat{\theta}_n \right) = \epsilon_n^{-2\gamma} \left( \hat{\theta}_n^* - \theta_0 \right) - \epsilon_n^{-2\gamma} \left( \hat{\theta}_n - \theta_0 \right),
\]

which will be valid if (1) \( \epsilon^{-2\gamma} \left( \hat{\theta}_n - \theta_0 \right) = o_p \( 1 \) and (2) \( \epsilon^{-2\gamma} \left( \hat{\theta}_n^* - \theta_0 \right) \overset{p}{\to} \mathcal{J} \). Part (1) follows from \( \sqrt{n} \epsilon \to \infty \) since \( \epsilon^{-2\gamma} \left( \hat{\theta}_n - \theta_0 \right) = \frac{1}{(\sqrt{n} \epsilon)^{\gamma}} n^{\gamma} \left( \hat{\theta}_n - \theta_0 \right) = o_p \( 1 \). Regarding part (2), write \( \mathcal{Z}_n^* g (\cdot, \theta) = (\mathcal{Z}_n^* - P) g (\cdot, \theta) + P g (\cdot, \theta) \), so that

\[
\hat{\theta}_n^* = \arg \max_{\theta \in \Theta} \mathcal{Z}_n^* g (\cdot, \theta) = (\mathcal{Z}_n^* - P) g (\cdot, \theta) - \frac{1}{2} (\theta - \theta_0)' \left( H + o_p \( 1 \) \right) (\theta - \theta_0).
\]

For the first term, note that \( (\mathcal{Z}_n^* - P) = \frac{1}{\sqrt{n}} \sqrt{n} (P_n - P) + \epsilon_n \hat{\theta}_n^* \overset{p}{\to} \mathcal{G}_0 + \epsilon_n \mathcal{G}_1 \) where \( \mathcal{G}_0 \) and \( \mathcal{G}_1 \) are independent copies of the same Gaussian process. Since \( \epsilon_n \gg \frac{1}{\sqrt{n}} \), the second term should dominate, so that \( \left( \mathcal{Z}_n^* - P \right) \sim \epsilon_n \mathcal{G}_1 \). Consequently, we expect

\[
\hat{\theta}_n \sim \arg \max_{\theta \in \Theta} \epsilon_n \mathcal{G}_1 g (\cdot, \theta) - \frac{1}{2} (\theta - \theta_0)' H (\theta - \theta_0)
\]

\[
= \epsilon_n O_p \( |\theta - \theta_0|^\rho \) - \frac{1}{2} (\theta - \theta_0)' H (\theta - \theta_0).
\]

By definition of \( \hat{\theta}_n^* \), \( \epsilon_n O_p \left( |\hat{\theta}_n^* - \theta_0|^\rho \right) + \left( \hat{\theta}_n^* - \theta_0 \right)' H \left( \hat{\theta}_n^* - \theta_0 \right) \geq 0 \), implying that

\[
|\hat{\theta}_n^* - \theta_0|^{2-\rho} \leq O_p \( \epsilon_n \) \quad \Rightarrow \quad |\hat{\theta}_n^* - \theta_0| \leq O_p \left( \frac{1}{\epsilon_n^{\rho}} \right) = O_p \left( \epsilon_n^{2\gamma} \right).
\]

To be more formal, let \( \hat{h}_n^* = \epsilon^{-2\gamma} \left( \hat{\theta}_n^* - \theta_0 \right) \). Then

\[
\hat{h}_n^* = \max_{h \in \epsilon_n^{-2\gamma} (\Theta - \theta_0), \epsilon_n^{-4\gamma}} \left\{ (\mathcal{Z}_n^* - P) g (\cdot, \theta_0 + \epsilon_n^{2\gamma} h) + P g (\cdot, \theta_0 + \epsilon_n^{2\gamma} h) \right\}.
\]

The second term \( \epsilon_n^{-4\gamma} P g (\cdot, \theta_0 + \epsilon_n^{2\gamma} h) \to -\frac{1}{2} h' H h \). It is shown in the appendix that the first term satisfies

\[
\epsilon_n^{-4\gamma} (\mathcal{Z}_n^* - P) g (\cdot, \theta_0 + \epsilon_n^{2\gamma} h) \sim \epsilon_n^{-4\gamma} \left( \frac{1}{\sqrt{n}} \mathcal{G}_0 + \epsilon_n \mathcal{G}_1 \right) g (\cdot, \theta_0 + \epsilon_n^{2\gamma} h)
\]

26
and that for a suitable Gaussian process $Z_0$ (as in Kim and Pollard (1990)),

$$
\epsilon_{n}^{-4\gamma} \left( \frac{1}{\sqrt{n}} G_0 + \epsilon_n G_1 \right) g \left( \cdot, \theta_0 + \epsilon_n^2 h \right) \sim \epsilon_{n}^{-4\gamma} \left( G_1 g \left( \cdot, \theta_0 + \epsilon_n^2 h \right) \right) \xrightarrow{p} Z_0(h).
$$

Combining the first and second terms implies that $\hat{h}_n^* \xrightarrow{p} J = \arg \max_h Z_0(h) - \frac{1}{2} h' H h$. Altogether, parts (1) and (2) imply that $\frac{1}{\epsilon_n^2} \left( \hat{\theta}_n^* - \hat{\theta}_n \right) \xrightarrow{p} J$, which validates the asymptotic consistency of the numerical bootstrap method.

In a more conventional approach such as Jun et al. (2015), $J$ is approximated by $\hat{J}^* = \arg \max_h \hat{Z}_0(h) - \frac{1}{2} h' \hat{H} h$ where $\hat{H} \xrightarrow{p} H$ and $\hat{Z}_0(h)$ is a Gaussian process with estimated covariance kernel $\hat{\Sigma}_\rho(s, t)$ for $\hat{\Sigma}_\rho(s, t) \xrightarrow{p} \Sigma_\rho(s, t)$. Instead, the numerical bootstrap essentially replaces $\hat{Z}_0(h)$ with $\epsilon_{n}^{-4\gamma} Z_n^* \left( \hat{\theta}_n + \epsilon_n h \right)$ since $\hat{J}^* = \frac{1}{\epsilon_n^{-2\gamma}} \left( \hat{\theta}_n^* - \hat{\theta}_n \right) = \arg \max_{h \in \epsilon_n^{-2\gamma} (\Theta - \theta_0)} \epsilon_{n}^{-4\gamma} Z_n^* \left( \hat{\theta}_n + \epsilon_n h \right)$.

There are two leading cases for Theorem 6.1: the smooth case and the cubic root case. In the smooth case, $\rho = 1$ and $\gamma = \frac{1}{2}$, and the Gaussian process $G_0 g (\cdot; \theta)$ is linearly separable in $\theta$. Typically there exists a multivariate normal random vector $W_0 \sim N(0, \Omega)$ such that $G_0 g (\cdot; \theta) = W_0 (\theta - \theta_0)$, and for an independent copy $W_1$ of $W_0$, $G_1 g (\cdot; \theta) = W_1 (\theta - \theta_0)$. The regular bootstrap is valid in this case due to linear separability, and corresponds to $\epsilon_n = 1/\sqrt{n}$. In particular,

$$
\hat{\theta}_n^* = \arg \max_{\theta \in \Theta} Z_n^* g (\cdot; \theta) \equiv (Z_n^* - P_n) g (\cdot; \theta) + (P_n - P) g (\cdot; \theta) + P g (\cdot; \theta) \\
\approx \frac{W_0 + W_1}{\sqrt{n}} (\theta - \theta_0) - \frac{1}{2} (\theta - \theta_0)' H (\theta - \theta_0),
$$

since $(Z_n^* - P_n) g (\cdot; \theta) \approx W_1/\sqrt{n}$ and $(P_n - P) g (\cdot; \theta) \approx W_0/\sqrt{n}$. Likewise the sample estimate satisfies

$$
\hat{\theta}_n = \arg \max_{\theta \in \Theta} P_n g (\cdot; \theta) = (P_n - P) g (\cdot; \theta) + P g (\cdot; \theta) \\
\approx \frac{\hat{G}_n}{\sqrt{n}} g (\cdot; \theta) - \frac{1}{2} (\theta - \theta_0)' H (\theta - \theta_0) \approx \frac{W_0}{\sqrt{n}} (\theta - \theta_0) - \frac{1}{2} (\theta - \theta_0)' H (\theta - \theta_0).
$$

Hence if we let $\hat{h}_n^* = \sqrt{n} \left( \hat{\theta}_n^* - \theta_0 \right)$ and $\hat{h}_n = \sqrt{n} \left( \hat{\theta}_n - \theta_0 \right)$, then $\hat{h}_n^* \xrightarrow{p} H^{-1} (W_0 + W_1)$.
and \( \hat{h}_n \overset{p}{\to} H^{-1}W_0 \), so that
\[
\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n) = \hat{h}_n - \hat{h}_n \overset{p}{\to} H^{-1}W_1 = N(0, H^{-1}\Omega H^{-1}).
\]

## 6.1 Constrained M estimation

A related application is to constrained M-estimators when the parameter (in a correctly specified model) can possibly lie on the boundary of the constrained set. In the following we verify the consistency of the numerical bootstrap, under conditions given in Geyer (1994), Knight (1999), and in Theorem 6.1. Alternative approaches to similar problems are provided in Römsch (2005) and Bonnans and Shapiro (2013). While the latter approach provides a closer tie between the numerical bootstrap and the numerical delta method, the former approach seems more in line with the convention in the statistics literature. To simplify notation when we make use of results from Geyer (1994), we consider arg min instead of arg max.

Following the notation in section 6, replace the parameter space \( \Theta \) by a constrained subset \( C \) such that for \( \hat{\theta}_n \in C \),
\[
P_n \pi \left( \cdot, \hat{\theta}_n \right) \leq \inf_{\theta \in C} P_n \pi \left( \cdot, \theta \right) + o_P \left( n^{-2\gamma} \right),
\]
and for \( \hat{\theta}^*_n \in C \),
\[
\mathcal{Z}_n^* \pi \left( \cdot, \hat{\theta}^*_n \right) \leq \inf_{\theta \in C} \mathcal{Z}_n^* \pi \left( \cdot, \theta \right) + o^* \left( \epsilon_n^{4\gamma} \right).
\]

Let \( C \) be approximated by a cone \( T_C(\theta_0) \) at \( \theta_0 \) in the sense of Theorem 2.1 in Geyer (1994), which implies (pp 2002 Geyer (1994)) that for \( n \to \infty \),
\[
+\infty \left( \delta \notin n^{-\gamma} (C - \theta_0) \right) \overset{\mathcal{F}}{\to} +\infty \left( \delta \notin T_C(\theta_0) \right).
\]

Here, \( \mathcal{F} \) denotes epigraphical convergence as defined in Geyer (1994) pp1997. The difficulty of practical inference lies in the challenge of estimating the approximating cone \( T_C(\theta_0) \) (Shapiro (1989)), which is easily handled by the numerical bootstrap method.

The following theorem combines the results in Geyer (1994), Knight (1999) and Theorem 6.1. A restricted version of Theorem 6.2 corresponding to \( \rho = 1 \) and \( \gamma = 1/2 \) can also be stated using only Assumptions A-D, Lemma 4.1, and Theorem 4.4 in Geyer (1994). It also includes Theorem 6.1 as a special case when \( T_C(\theta_0) = R^d \).
Theorem 6.2 Assume $\theta_0 = \arg\min_{\theta \in \mathcal{C}} P\pi(\cdot, \theta)$. Let (23) and the conditions in Theorem 6.1 hold (except for (i) and (iii) and replace (v) with a positive definite $H$). Also assume that $\mathcal{J} \equiv \arg\min_{h \in \mathcal{C}(\theta_0)} \mathcal{Z}_0(h) + \frac{1}{2} h'Hh$ is almost surely unique, then $n^{-\frac{1}{2}}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{J}$, and $\mathcal{Z}_n^* \equiv \epsilon_n^{-2\gamma} (\hat{\theta}_n^* - \hat{\theta}_n) \xrightarrow{\mathcal{L}} \mathcal{J}$, and $\mathcal{Z}_n^* \xrightarrow{\mathcal{L}} \mathcal{J}$ unconditionally.

If $\theta_0$ is in the interior of $C$, then $T_C(\theta_0) = R^d$ and the proof of Theorem 6.1 can be applied. In other special cases, the proof of Theorem 6.1 can also be applied without change to Theorem 6.2, without having to appeal to the notion of epi-convergence. For example, it applies when $\theta_0$ is on the boundary of $C$ and $C - \theta_0$ already contains a cone at the origin, meaning for any compact set $K$, there exists $\alpha > 0$ such that $T_C(\theta_0) \cap K \subset \alpha (C - \theta_0)$ where $C - \theta_0$ is the tensor product between a cone at the origin and an open set.

Theorem 6.2 is based on the M-estimation framework, but generalization to (correctly specified) GMM models is immediate. To simplify notation, consider a fixed weighting matrix. In GMM models, $\hat{\theta}_n = \arg\min_{\theta \in \mathcal{C}} n\hat{Q}_n(\theta)$, where

$$\hat{Q}_n(\theta) = \hat{\pi}(\theta)' W \hat{\pi}(\theta) \quad \text{and} \quad \hat{\pi}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \pi(z_i, \theta).$$

Assumption 6.1 1. $\Theta$ is compact and $\pi(\theta) = E\pi(z_i, \theta)$

2. $\pi(\theta)$ is four times continuously differentiable

3. $\{\pi(\cdot, \theta) : \theta \in \Theta\}$ is a VC class of functions

4. $\pi(\theta) = 0$ if and only $\theta = \theta_0$ and $\theta_0 \in \mathcal{C}$

Define $G_0 = \frac{\partial}{\partial \theta_0} \pi(\theta_0)$, let $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \pi(z_i, \theta_0) \xrightarrow{\mathcal{L}} Z = N(0, \Omega)$. Also define $\Delta_n = G_0 W \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \pi(z_i, \theta_0)$, $\Delta_0 = G_0 WZ$, and $H = G_0 WG_0'$. It is known (e.g. Chernozhukov and Hong (2003)) that Assumption 6.1 implies the following identification condition and quadratic expansion of the objective function $\hat{Q}_n(\theta)$:

$$\forall \delta > 0, \sup_{\epsilon > 0} \limsup P \left( \inf_{|\theta - \theta_0| \geq \delta} \hat{Q}_n(\theta) - \hat{Q}_n(\theta_0) \geq \epsilon \right) = 1. \quad (24)$$

and for $R_n(\theta) = \hat{Q}_n(\theta) - \hat{Q}_n(\theta_0) - \Delta_n \sqrt{n}(\theta - \theta_0) - n(\theta - \theta_0)' H^{-1} (\theta - \theta_0)$,

$$\forall \delta_n \to 0, \sup_{|\theta - \theta_0| \leq \delta_n} \frac{|R_n(\theta)|}{1 + \sqrt{n}|\theta - \theta_0| + n|\theta - \theta_0|^2} = o_P(1). \quad (25)$$
Under (25), which also holds for most M-estimators, \( \dot{Q}_n(\theta) \) is locally approximated by a quadratic function:

\[
\dot{Q}_n(\theta) = \frac{1}{2} \left( \sqrt{n}(\theta - \theta_0) + H^{-1} \Delta_n \right)' H \left( \sqrt{n}(\theta - \theta_0) + H^{-1} \Delta_n \right) - \frac{1}{2} \Delta_n' H^{-1} H \Delta_n. 
\]

This leads to the asymptotic distribution

\[
\dot{J}_n = \sqrt{n} \left( \dot{\theta}_n - \theta_0 \right) \sim \mathcal{J} = \arg \min_{h \in T_C(\theta_0)} \left( h + H^{-1} \Delta_0 \right)' H \left( h + H^{-1} \Delta_0 \right). \tag{26}
\]

Each of the three unknown components can be consistently estimated. (1) If \( \pi(z_i, \theta) \) is twice differentiable, let \( \hat{H} = \frac{\partial^2}{\partial \theta \partial \theta} \frac{1}{n} \sum_{i=1}^{n} \pi \left( z_i; \hat{\theta}_n \right) \). Otherwise \( \hat{H} \) can be replaced by numerical differentiation (Hong et al. (2015)). (2) Let \( \hat{G} \) be either \( \frac{\partial}{\partial \theta} \frac{1}{n} \sum_{i=1}^{n} \pi \left( z_i; \hat{\theta}_n \right) \) or a numerical derivative analog, and let \( \hat{\Omega} = \frac{1}{n} \sum_{i=1}^{n} \pi \left( z_i; \hat{\theta}_n \right) \pi \left( z_i; \hat{\theta}_n \right)' \). Then let \( \hat{Z}_n = N \left( 0, \hat{\Omega} \right) \) be such that \( \hat{Z}_n \xrightarrow{p} Z, \hat{\Delta}_n = \hat{G} \hat{W} \hat{Z}_n \) so that \( \hat{\Delta}_n \xrightarrow{p} \Delta \). (3) Since \( T_C(\theta_0) \) is the limit of \( \sqrt{n}(C - \theta_0) \), we can also estimate \( T_C(\theta_0) \) by \( \epsilon^{-1}(C - \hat{\theta}_n) \).

Therefore we define, with \( \hat{\Gamma}_n = -\hat{H}^{-1} \hat{\Delta}_n \),

\[
\hat{J}_n^* = \arg \min_{h \in \epsilon^{-1}(C - \hat{\theta}_n)} \left( h + \hat{H}^{-1} \hat{\Delta}_n^* \right)' \hat{H} \left( h + \hat{H}^{-1} \hat{\Delta}_n^* \right) = \left( h - \hat{\Gamma}_n \right)' \hat{H} \left( h - \hat{\Gamma}_n \right). \tag{27}
\]

For example, if \( C = (\theta \geq 0) \), then \( \{ h \in \epsilon^{-1}(C - \hat{\theta}_n) \} = \{ h \geq -\epsilon^{-1}(C - \hat{\theta}_n) \} \).

In the regular M-estimator problem where \( \dot{Q}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \pi \left( z_i, \theta \right) \), we typically have \( \hat{H} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta \partial \theta} \pi \left( z_i; \hat{\theta}_n \right) \) or a numerical derivative analog, and \( \hat{\Delta}_n \sim N \left( 0, \hat{\Sigma} \right) \), where \( \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \pi \left( z_i; \hat{\theta}_n \right) \frac{\partial}{\partial \theta} \pi \left( z_i; \hat{\theta}_n \right)' \), or a numerical derivative analog.

**Theorem 6.3** Under Assumption 6.1, or more generally (24) and (25), (26) holds, and \( \hat{J}_n^* \xrightarrow{p} \mathcal{J} \).

Theorem 6.2 allows for more general nonstandard asymptotics with \( \gamma = 1/3 \). Theorem 6.3 is only confined to the regular case of \( \gamma = 1/2 \), but can be easier to implement since the objective function \( \left( h - \hat{\Gamma}_n \right)' \hat{H} \left( h - \hat{\Gamma}_n \right) \) is convex whenever \( \hat{H} \) is positive semi-definite. In particular, if \( C \) is a polyhedron, then the problem becomes a Quadratic Program, which can be quickly solved using off-the-shelf optimization routines.
7 Misspecified Nonsmooth GMM

In this section we investigate the properties of overidentified and misspecified GMM models that involve nonsmooth moments. Examples of GMM models with nonsmooth moments include the multinomial discrete choice models in (McFadden (1989) and Pakes and Pollard (1989)) and the quantile instrumental variable models in (Chernozhukov and Hansen (2005)). The asymptotic distribution of simulation based GMM under correct specification is developed by Pakes and Pollard (1989). The properties of smooth non-simulation based misspecified GMM models are developed by Hall and Inoue (2003) and Lee (2014). In this section, we first show that for nonsmooth globally misspecified GMM models, cube root asymptotics such as those in Kim and Pollard (1990) and Jun et al. (2015) apply. Then we demonstrate the validity of the numerical bootstrap for this model. Note that as pointed out in Hong et al. (2015) and Jun et al. (2015), the numerical implementation can potentially change the rate of convergence and the asymptotic distribution to resemble a nonparametric estimator. The estimation methods and the resulting asymptotic distributions in these two papers also obviously apply to this model.

To simplify exposition, consider a fixed weighting matrix $W$ and a misspecified model:

Assumption 7.1 Let $\hat{\theta}_n = \arg\min_{\theta \in \Theta} \hat{Q}_n (\theta)$, where $\hat{Q}_n (\theta) = \hat{\pi} (\theta)' W \hat{\pi} (\theta)$, and $\hat{\pi} (\theta) = \frac{1}{n} \sum_{i=1}^{n} \pi (z_i, \theta)$. Let Assumption 6.1 hold except we replace the last condition by (1) $\theta_0 = \arg\min_{\theta \in \Theta} \pi (\theta)' W \pi (\theta)$ is unique and an interior point of $\Theta$ and (2) $\pi (\theta_0) \neq 0$.

By Assumption 7.1, $G (\theta_0)' W \pi (\theta_0) = 0$ where $G (\theta) = \frac{\partial}{\partial \theta} \pi (\theta)$. It follows immediately from Assumption 7.1 that $\hat{\theta}_n \overset{p}{\rightarrow} \theta_0$.

Assumption 7.2 Let $g (\cdot, \theta) \equiv \pi (\cdot, \theta) - \pi (\cdot, \theta_0)$ satisfy conditions (iv), (v), (vi), (vii), (viii), (x) and (xi) of Theorem 6.1 with $\rho = \frac{1}{2}$, where $H$ is replaced by $G = G (\theta_0)$ in condition (v).

Theorem 7.1 Under assumptions 7.1 and 7.2, $n^{1/3} \left( \hat{\theta}_n - \theta_0 \right) \rightsquigarrow J$ where

$$J = \arg\min_{h} \tilde{Z}_0 (h) + \frac{1}{2} h' \bar{H} h, \quad \tilde{Z}_0 (h) = \pi (\theta_0)' W Z_0 (h) \quad \text{and}$$
where $H = \frac{\partial^2}{\partial \theta \partial \theta'} \pi (\theta_0)$ and $Z_0 (h)$ is a mean zero Gaussian process with covariance kernel $\Sigma_\rho (s,t)$ in condition (vi) of Theorem 6.1.

For the numerical bootstrap, $\hat{\theta}_n^* = \text{arg min}_{\theta \in \Theta} (Z_n^* \pi (\cdot, \theta))^\prime W (Z_n^* \pi (\cdot, \theta))$, where $Z_n^* = P_n + \epsilon_n \hat{G}_n^*$, so that with the multinomial bootstrap,

$$Z_n^* (\cdot, \theta) = \frac{1}{n} \sum_{i=1}^{n} \pi (z_i, \theta) + \epsilon_n \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\pi (z_i^*, \theta) - \pi (z_i, \theta)).$$

**Theorem 7.2** Under the same conditions as in Theorem 7.1, if $\sqrt{n} \epsilon_n \to \infty$,

$$\epsilon_n^{-2/3} \left( \hat{\theta}_n^* - \hat{\theta}_n \right) \overset{p}{\to} J.$$

To avoid the complexity in Theorems 7.1 and 7.2, in practice either insistence on correct specification of the model is needed, or model specification tests need to be conducted. Confidence intervals constructed using Theorem 7.2 will be conservatively correct but will overcover under correct specification. In linear models, the validity of misspecified GMM asymptotics is justified by Lee (2015). It remains an open question whether his results can be extended to nonlinear models. Another question is whether it is possible to conduct asymptotically accurate adaptive inference regardless of whether the model conditions are correctly specified or misspecified.

8 Application to Lasso, finite dimensional case

In this section, we apply the numerical bootstrap to estimate the limiting distribution of the LASSO estimator of Tibshirani (1996). It is well known that the LASSO estimator is not regular and that its asymptotic distribution cannot be consistently estimated by the conventional bootstrap. We consider the finite dimensional case as in Knight and Fu (2000):

$$\hat{\beta}_n = \text{arg min}_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i' \beta)^2 + \frac{\lambda_n}{\sqrt{n}} \sum_{k=1}^{p} |\beta_k|. \quad (28)$$
We define
\[
\hat{\beta}^*_n = \arg \min_{\beta} \mathcal{Z}_n^* (y - x' \beta)^2 + \lambda_n \epsilon_n \sum_{k=1}^p |\beta_k| \tag{29}
\]
where $\mathcal{Z}_n^* = P_n + \epsilon_n \hat{\mathcal{Z}}_n^*$. For example, with multinomial bootstrap,
\[
\mathcal{Z}_n^* (y - x' \beta)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - x'_i \beta)^2 + \frac{\epsilon_n}{\sqrt{n}} \left( \sum_{i=1}^n (y_i^* - x'_i \hat{\beta})^2 - \sum_{i=1}^n (y_i - x'_i \beta)^2 \right) \tag{30}
\]
Knight and Fu (2000) show that $\hat{\beta}_n \overset{P}{\to} \beta_0$ if $\lambda_n = o(\sqrt{n})$. Under a stronger condition, the distribution of $\sqrt{n} \left( \hat{\beta}_n - \beta_0 \right)$, even if it diverges, can be approximated by the numerical bootstrap $\frac{1}{\epsilon_n} \left( \hat{\beta}_n^* - \hat{\beta}_n \right)$. In the following we first present a general theorem, and then an illustrative example. Assume that $\lambda_n \to \lambda_0 \in [0, \infty]$.

**Theorem 8.1 (Consistency of Numerical Bootstrap for Lasso)** Define $\hat{\mathcal{G}}_n$ to be the law of $\sqrt{n} \left( \hat{\beta}_n - \beta_0 \right)$ and $\hat{\mathcal{G}}_n^*$ to be the law of $\epsilon_n^{-1} \left( \hat{\beta}_n^* - \hat{\beta}_n \right)$. If $\frac{\lambda_n}{\sqrt{n}} \to 0$ and $\lambda_n \epsilon_n \to 0$, then $\rho_{BL_1} \left( \hat{\mathcal{G}}_n, \mathcal{G}_n^0 \right) = o(1)$, and $\rho_{BL_1} \left( \hat{\mathcal{G}}_n^*, \mathcal{G}_n^0 \right) = o_P(1)$, where $\mathcal{G}_n^0$ is the law of
\[
\arg \min_u -2u^TW + u^TCu + \lambda_n \sum_{j=1}^p [u_j \text{sgn} (\beta_{0j}) 1 (\beta_{0j} \neq 0) + |u_j| 1 (\beta_{0j} = 0)]
\]
for $C = \text{plim} \left( \frac{1}{n} \sum_{i=1}^n x_i x'_i \right)$ and $\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \epsilon_i \rightsquigarrow W$.

**Example** Consider the orthonormal design case in Tibshirani (1996) where $X'X = I$ and let $\hat{\beta} = \hat{y}$ be the OLS estimate. Then it is easy to show that for each $j = 1, \ldots, p$, in (28)
\[
\hat{\beta}_{nj} = \text{sgn} \left( \bar{\beta}_j \right) \left( |\bar{\beta}_j| - \lambda_n / \sqrt{n} \right)^+ = (\bar{\beta}_j - \lambda_n / \sqrt{n})^+ - (\bar{\beta}_j + \lambda_n / \sqrt{n})^-
\]
The OLS estimate satisfies $Z_n \equiv \sqrt{n} \left( \hat{\beta} - \beta_0 \right) \rightsquigarrow Z = N (0, \Sigma)$. Let $\hat{\mathcal{G}}_n$ be the law of $\sqrt{n} \left( \hat{\beta}_n - \beta_0 \right)$ and $\mathcal{G}_n^0$ the law of
\[
\left\{ 1 (\beta_{0j} \neq 0) (Z_j - \text{sgn} (\beta_{0j}) \lambda_n) + 1 (\beta_{0j} \equiv 0) ((Z_j - \lambda_n)^+ - (Z_j + \lambda_n)^-) ; j = 1, \ldots, p \right\}.
\]
It is straightforward to show that $\rho_{BL_3} \left( \hat{\mathbb{G}}_n, \mathbb{G}_n^0 \right) = o(1)$, by considering sign combinations of $\beta_{0j}$ and noting that

$$\sqrt{n} \left( \hat{\beta}_n - \beta_0 \right) = \left\{ (\sqrt{n} \beta_{0j} + Z_{nj} - \lambda_n)^+ - (\sqrt{n} \beta_{0j} + Z_{nj} + \lambda_n)^- - \sqrt{n} \beta_{0j}, j = 1, \ldots, p \right\}$$

While Theorem 2 in Knight and Fu (2000) requires $\lambda_0 < \infty$, $\hat{\mathbb{G}}_n$ is consistent for $\mathbb{G}_n^0$ even when $\lambda_0 = \infty$. In both cases, $\mathbb{G}_n^0$ can be consistently estimated by the numerical bootstrap which takes the form

$$\hat{\beta}_n^* = \left\{ (\hat{\beta}_j + \epsilon_n \sqrt{n} (\hat{\beta}_j - \beta_j) - \epsilon_n \lambda_n)^+ - (\hat{\beta}_j + \epsilon_n \sqrt{n} (\hat{\beta}_j - \beta_j) + \epsilon_n \lambda_n)^-, \forall j \right\}.$$

More generally, $\hat{\beta}_n^* = \left\{ (\hat{y}_j + \epsilon_n \hat{\mathbb{G}}_n^* y_j - \epsilon_n \lambda_n)^+ - (\hat{y}_j + \epsilon_n \hat{\mathbb{G}}_n^* y_j + \epsilon_n \lambda_n)^-, \forall j \right\}$. It should be clear that $\hat{\mathbb{G}}_n^* y$ can be replaced by $\hat{Z}_n^* \overset{P}{\rightarrow} Z$, or any consistent estimate of the OLS limiting distribution. For the Wild bootstrap,

$$\hat{\beta}_{nj}^* = \left( \bar{\beta}_j + \frac{\epsilon_n}{\sqrt{n}} \sum_{i=1}^n (\xi_i - \xi_j) y_{ij} - \epsilon_n \lambda_n \right)^+ - \left( \bar{\beta}_j + \frac{\epsilon_n}{\sqrt{n}} \sum_{i=1}^n (\xi_i - \xi_j) y_{ij} + \epsilon_n \lambda_n \right)^-$$

The numerical bootstrap distribution $\hat{\mathbb{G}}_n^*$ is defined as the law of

$$\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\epsilon_n} = \left\{ \left( \frac{\hat{y}_j}{\epsilon_n} + \hat{\mathbb{G}}_n^* y_j - \lambda_n \right)^+ - \left( \frac{\hat{y}_j}{\epsilon_n} + \hat{\mathbb{G}}_n^* y_j + \lambda_n \right)^- - \frac{\hat{\beta}_{nj}}{\epsilon_n}, \forall j \right\}.$$

Under the condition that $\min(1, \lambda_{n}^{-1}) \sqrt{n} \epsilon_n \rightarrow \infty$, both $\hat{y}_j/\epsilon_n$ and $\hat{\beta}_{nj}/\epsilon_n$ can be replaced by $\beta_{0j}/\epsilon_n$ up to an $o_P(1)$ term. If we replace $\hat{\beta}_n$ by $\beta$ in $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\epsilon_n}$, then only $\sqrt{n} \epsilon_n \rightarrow \infty$ is needed to approximate $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\epsilon_n}$ by

$$\left\{ \left( \frac{\beta_{0j}}{\epsilon_n} + \hat{\mathbb{G}}_n^* y_j - \lambda_n \right)^+ - \left( \frac{\beta_{0j}}{\epsilon_n} + \hat{\mathbb{G}}_n^* y_j + \lambda_n \right)^- - \frac{\beta_{0j}}{\epsilon_n} + o_P(1), \forall j \right\}.$$

If additionally, $\lambda_n \epsilon_n \rightarrow 0$, then $\beta_{0j}/\epsilon_n$ will determine the signs w.p.1, and then using $\hat{\mathbb{G}}_n^* y \overset{P}{\rightarrow} Z$, we conclude that $\rho_{BL_3} \left( \hat{\mathbb{G}}_n, \mathbb{G}_n^0 \right) = o_P(1)$. As long as not all $\beta_{0j} = 0$, $\mathbb{G}_n^0$ will not be degenerate at zero and $\hat{\mathbb{G}}_n^*$ can be used for asymptotic inference. Note that $\lambda_n \epsilon_n \rightarrow 0$ and $\lambda_{n}^{-1} \sqrt{n} \epsilon_n \rightarrow \infty$ would require that $\lambda_n = o \left( n^{1/4} \right)$. However,
\[ \lambda_n^{-1} \sqrt{n} \epsilon_n \to \infty \] is not required if we replace \( \hat{\beta}_n \) by \( \bar{\beta} \) in \( \hat{G}_n^* \), in which case we only need \( \lambda_n = o(\sqrt{n}) \) to allow for a feasible \( \epsilon_n \) sequence. When \( \lambda_n \to \infty \) and \( \beta_{0j} \neq 0 \), \( \frac{\hat{\beta}_n - \beta_{0j}}{\epsilon_n} \) does not need to be stochastically bounded.

9 Application to 1-norm SVM, finite dimensional case

The analysis for the LASSO estimator suggests that similar arguments also apply to M-estimators whose objective functions admit a local quadratic expansion in the parameters. In this section, we investigate the asymptotics of the finite dimensional 1-norm support vector machine (SVM) model of Zhu et al. (2004), which is similar to the LASSO quantile regression model of Belloni and Chernozhukov (2011). We first describe its asymptotic distribution, and then apply the numerical bootstrap to consistently estimate that distribution. For \( \kappa > 0, \lambda_n > 0 \), the 1-norm SVM estimator is

\[ \hat{\beta}_n = \arg \min_{\beta} \hat{Q}_n (\beta) = \frac{1}{n} \sum_{i=1}^{n} (\rho_{\tau} (y_i - x_i' \beta) - \kappa)^+ + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{k} |\beta_j|. \]

where \( \rho_{\tau}(u) \) is the Koenker and Bassett (1978) check function. When \( \tau = \frac{1}{2} \), \( \rho_{\tau}(y - x' \beta) = \frac{1}{2} |y - x' \beta| \), and if additionally \( \kappa = 0 \), then \( \hat{\beta} \) is the LASSO quantile regression estimator of Belloni and Chernozhukov (2011). By a standard uniform law of large numbers, when \( \frac{\lambda_n}{\sqrt{n}} \to 0 \), \( \hat{\beta}_n \overset{p}{\to} \beta_0 \), where \( \beta_0 = \arg \min_{\beta \in B} E (\rho_{\tau} (y_i - x_i' \beta) - \kappa)^+ \) solves the asymptotic first order condition

\[ Ex_i \left[ \tau_1 \left( y_i \geq x_i' \beta_0 + \frac{\kappa}{\tau} \right) - (1 - \tau) \left( y_i \leq x_i' \beta_0 - \frac{\kappa}{1 - \tau} \right) \right] = 0. \]

When \( \tau = 1/2 \) and \( x_i \equiv 1 \), this reduces to \( F_Y (\beta_0 - \frac{\kappa}{2}) = \bar{F}_Y (\beta_0 + \frac{\kappa}{2}) \). When \( Y \) is symmetrically distributed, \( \beta_0 \) becomes the median. To state the asymptotic distribution for \( \sqrt{n} (\hat{\beta}_n - \beta_0) \), define

\[ \Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \left[ \tau \left( y_i \geq x_i' \beta_0 + \frac{\kappa}{\tau} \right) - (1 - \tau) \left( y_i \leq x_i' \beta_0 - \frac{\kappa}{1 - \tau} \right) \right], \]
and \( H = E x_i x_i' (\tau f_y (x_i' \beta_0 + \frac{\kappa}{\tau} x_i) + (1 - \tau) f_y (x_i' \beta_0 - \frac{\kappa}{1 - \tau} x_i)) \). Note that \( \Delta_n \sim \Delta \) where

\[
\Delta \equiv N \left( 0, V = E x_i x_i' \left( \tau^2 P \left( y_i > x_i' \beta_0 + \frac{\kappa}{\tau} \right) + (1 - \tau)^2 P \left( y_i < x_i' \beta_0 - \frac{\kappa}{1 - \tau} \right) \right) \right)
\]

Bootstrap does not work when \( \beta_0 = 0 \), but the numerical bootstrap will provide a consistent estimate of the limiting distribution. In the following theorem, define \( \hat{\beta}_n^* = \arg \min_{\beta} Z_n^* (\rho (y - x' \beta) + \kappa) + \lambda_n \epsilon_n \sum_{j=1}^{k} |\beta_j| \) for \( Z_n^* = P_n + \epsilon_n \hat{G}_n^* \) and denote \( L(X) \) as the law of the random variable \( X \).

**Theorem 9.1 (Consistency of Numerical Bootstrap for SVM)** Let \( \beta_0 \) be uniquely defined, and \( H \) nonsingular. Then \( \rho_{BL_1} \left( L \left( \sqrt{n} \left( \hat{\beta}_n - \beta_0 \right) \right) , L \left( \tilde{J}_n \right) \right) = o(1) \), where \( \tilde{J}_n = \arg \min_h \tilde{L}_n (h) \) for

\[
\tilde{L}_n (h) = \frac{1}{2} \left( h + H^{-1} \Delta \right)' H \left( h + H^{-1} \Delta \right) + \lambda_n \sum_{j=1}^{p} (\text{sgn} (\beta_{0j}) h_j + 1 (\beta_{0j} = 0) |h_j|).
\]

Furthermore, if \( \sqrt{n} \epsilon_n \to \infty \) and \( \lambda_n \epsilon_n \to 0 \), then

\[
\rho_{BL_1} \left( L \left( \epsilon_n^{-1} \left( \hat{\beta}_n^* - \hat{\beta}_n \right) \right) , L \left( \tilde{J}_n \right) \right) = o_P(1) .
\]

To illustrate, consider the scalar case when \( k = 1 \) and \( x_i \equiv 1 \), then \( \tilde{J}_n = -H^{-1} (\Delta + \text{sgn} (\beta_0) \lambda_n) \) if \( \beta_0 \neq 0 \). And if \( \beta_0 = 0 \), then analogous to the LASSO estimator, \( \tilde{J}_n = -H^{-1} \left( (-\Delta - \lambda_n)^+ - (-\Delta + \lambda_n)^- \right) \). When \( \lambda_n \to \infty \) and \( \beta_0 = 0 \), then \( \tilde{J}_n = 0 \) w.p.c.1.

10 **Recentering**

In hypothesis testing or in confidence set construction based on test statistic inversion, subsampling does not require recentering (unlike Bootstrap) to achieve consistency and local power (Politis et al. (1999)). However, recentering does improve finite sample power (Chernozhukov and Fernández-Val (2005)). The same insight applies to the numerical bootstrap method, which this section illustrates.

Consider, for example, testing \( H_0 : \theta (P) = \theta_0 \) vs \( H_1 : \theta (P) > \theta_0 \). The difference between the centered and noncentered versions of the numerical bootstrap method is analogous to those in subsampling tests. For \( \tilde{\theta}_n^* \equiv \theta (P_n + \epsilon_n \sqrt{n} (P_n^* - P_n)) \), 36
Kosorok (2007) and the fact that the CDF of $J$ and let following we assume that (1) power analysis come from studying these three distributions. To illustrate, in the following recentered version:

\[ a(n) \left( \hat{\theta}_n - \theta_* \right) = a(n) \left( \hat{\theta}_n - \theta_0 \right) + a(n) \mu \sim J + a(n) \mu \]

where $X_n \sim Y_n \iff \rho_{BL_1}(X_n, Y_n) = o(1)$. The noncentered numerical bootstrap distribution diverges to $\infty$ at the rate $a \left( \frac{1}{e^2 n} \right)$, which is slower than $a(n)$, since $a \left( \frac{1}{e^2 n} \right) / a(n) \to 0$, and

\[ a \left( \frac{1}{e^2 n} \right) \left( \hat{\theta}_n^* - \theta_* \right) = a \left( \frac{1}{e^2 n} \right) \left( \hat{\theta}_n^* - \theta_0 \right) + a \left( \frac{1}{e^2 n} \right) \mu \sim \rho_{BL_1}(X_n, Y_n) = o(1). \]

Therefore the noncentered test is consistent. More formally, let $\hat{c}_{1-a} = \inf \{ x : P_n \left( a \left( \frac{1}{e^2 n} \right) \left( \hat{\theta}_n^* - \theta_* \right) \leq x \right) \geq 1 - a \}$ and let $J_{1-a} = \inf \{ x : P(J \leq x) \geq 1 - a \}$. By arguments in Lemma 10.11 in Kosorok (2007) and the fact that the CDF of $J$ is strictly increasing on its support, $\hat{c}_{1-a} - J_{1-a} - a \left( \frac{1}{e^2 n} \right) \mu = o_P(1)$. Then by Slutsky,

\[ P \left( a(n) \left( \hat{\theta}_n - \theta_* \right) > \hat{c}_{1-a} \right) = P \left( a(n) \left( \hat{\theta}_n - \theta_* \right) + o_P(1) > J_{1-a} + a \left( \frac{1}{e^2 n} \right) \mu \right) = P \left( J > J_{1-a} + a \left( \frac{1}{e^2 n} \right) - a(n) \mu \right) + o(1) \to 1. \]

The noncentered test is consistent but can be less powerful in finite sample than the following recentered version:

\[ a \left( \frac{1}{e^2 n} \right) \left( \hat{\theta}_n^* - \theta_n \right) = a \left( \frac{1}{e^2 n} \right) \left( \hat{\theta}_n^* - \theta_0 \right) - \frac{a(1/e^2 n)}{a(n)} a(n) \left( \hat{\theta}_n - \theta_0 \right) \sim J. \]
If we let \( \hat{c}_{1-\alpha} = \inf \{ x : P_n \left( a \left( \frac{1}{\epsilon_n^2} \right) \left( \hat{\theta}_n - \theta_n \right) \leq x \right) \geq 1 - \alpha \} \), then we also have

\[
P \left( a (n) \left( \hat{\theta}_n - \theta_* \right) > \hat{c}_{1-\alpha} \right) = P \left( a (n) \left( \hat{\theta}_n - \theta_* \right) + o_P (1) > \mathcal{J}_{1-\alpha} \right)
\]

\[
= P (\mathcal{J} > \mathcal{J}_{1-\alpha} - a (n) \mu + o(1) \to 1.
\]

Observing that \( \hat{c}_{1-\alpha} - \hat{c}_{1-\alpha} = a \left( \frac{1}{\epsilon_n^2} \right) \mu + \frac{a \left( \frac{1}{\epsilon_n^2} \right)}{a(n)} a (n) \left( \hat{\theta}_n - \theta_0 \right) \), the power difference derives from

\[
P \left( \hat{c}_{1-\alpha} - \hat{c}_{1-\alpha} > a \left( \frac{1}{\epsilon_n^2} \right) \mu - c \right) \to 1 \quad \text{for all} \quad c > 0.
\]

Under the local alternative that \( \theta_0 = \theta_* + \frac{c}{a(n)} \) for \( c > 0 \), the sample distribution satisfies

\[
a (n) \left( \hat{\theta}_n - \theta_* \right) = a (n) \left( \hat{\theta}_n - \theta_0 \right) + c \sim \mathcal{J} + c.
\]

As \( a \left( \frac{1}{\epsilon_n^2} \right) \to 0 \), the noncentered numerical bootstrap distribution converges to the null limit:

\[
a \left( \frac{1}{\epsilon_n^2} \right) \left( \hat{\theta}_n^* - \theta_* \right) = a \left( \frac{1}{\epsilon_n^2} \right) \left( \hat{\theta}_n^* - \theta_0 \right) + \frac{a \left( \frac{1}{\epsilon_n^2} \right)}{a(n)} c \xrightarrow{p} \mathcal{J},
\]

and has the correct asymptotic local power. The centered numerical bootstrap has a limit that also does not depend on the local drift \( c \):

\[
a \left( \frac{1}{\epsilon_n^2} \right) \left( \hat{\theta}_n - \hat{\theta}_n \right) = a \left( \frac{1}{\epsilon_n^2} \right) \left( \hat{\theta}_n^* - \theta_0 \right) - \frac{a \left( \frac{1}{\epsilon_n^2} \right)}{a(n)} a (n) \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{p} \mathcal{J}.
\]

The relation between the centered and non-centered numerical bootstrap distributions depends on \( \mathcal{J}_n = a (n) \left( \hat{\theta}_n - \theta_0 \right) \sim \mathcal{J} \). When \( \mathcal{J} \) is a univariate centered normal distribution, the noncentered critical value \( \hat{c}_{1-\alpha} \) is more likely than not larger than the centered critical value \( \hat{c}_{1-\alpha} \), leading to less rejection. For \( b (n) = a \left( \frac{1}{\epsilon_n^2} \right) \) and \( \hat{c}_{1-\alpha} - \hat{c}_{1-\alpha} = b (n) \left( c + \mathcal{J}_n \right),

\[
P \left( \hat{c}_{1-\alpha} > \hat{c}_{1-\alpha} \right) = P \left( \mathcal{J}_n + c > 0 \right) > \frac{1}{2} + \delta \quad \text{for some} \quad \delta > 0 \quad \text{and all large} \ n.
11 Unknown polynomial convergence rate

Similar to subsampling, the numerical bootstrap can be used to estimate the unknown rate of convergence when the convergence rate is a polynomial function $a(n) = n^\beta$ of the sample size and when the numerical bootstrap consistently estimates the limiting distribution. This is done by comparing the empirical distributions estimated by two (or more) sequences of step sizes $\epsilon_n$. Let $L_{n,\epsilon_n}(x)$ denote the distribution of $\theta(P_n + \epsilon_n \hat{G}_n^*) - \theta(P_n)$, which is estimated by bootstrap simulations. Then

$$\epsilon_n^{-2\beta} L_{n,\epsilon_n}^{-1}(t) = \hat{J}_{\epsilon_n}^{-1}(t) = J^{-1}(t, P) + o_P(1).$$

For $t_1 \in (0, 0.5)$, $t_2 \in (0.5, 1)$, $\epsilon_n^{-2\beta} \left( L_{n,\epsilon_n}^{-1}(t_2) - L_{n,\epsilon_n}^{-1}(t_1) \right) = J^{-1}(t_2, P) - J^{-1}(t_1, P) + o_P(1)$, or

$$-2\beta \log \epsilon_n + \log \left( L_{n,\epsilon_n}^{-1}(t_2) - L_{n,\epsilon_n}^{-1}(t_1) \right) = J^{-1}(t_2, P) - J^{-1}(t_1, P) + o_P(1).$$

Using two step size sequences, $\epsilon_{n,1}$ and $\epsilon_{n,2}$, it is then natural to estimate $\beta$ by

$$\hat{\beta}_n = \frac{\log \left( L_{n,\epsilon_{n,2}}^{-1}(t_2) - L_{n,\epsilon_{n,2}}^{-1}(t_1) \right) - \log \left( L_{n,\epsilon_{n,1}}^{-1}(t_2) - L_{n,\epsilon_{n,1}}^{-1}(t_1) \right)}{2 \left( \log \epsilon_{n,2} - \log \epsilon_{n,1} \right)} = \beta + o_P \left( (\log n)^{-1} \right).$$

For example, take $\epsilon_{n,1} = n^{-\gamma_1/2}$ and $\epsilon_{n,2} = n^{-\gamma_2/2}$ where $0 < \gamma_2 < \gamma_1 < 1$. Then

$$\hat{\beta}_n = \beta + o_P \left( (\log n)^{-1} \right).$$

12 Moment Inequalities Simulation

We investigate the hypothesis test in Bugni et al. (2014):

$$H_0 : f(\beta) = \beta_k = \gamma_0 \text{ for } k = 1, 2 \quad H_1 : f(\beta) = \beta_k \neq \gamma_0 \text{ for } k = 1, 2$$

where $\beta$ is the parameter defined by a set of moment inequalities $Pg(\cdot; \beta) \geq 0$. We consider the entry game example of Bugni et al. (2014) with moment inequalities
The test statistic is

\[
P_g(z; \beta) = \begin{bmatrix}
    z_{1i}z_{2i} - (1 - \beta_1)(1 - \beta_2) \\
    -(z_{1i}z_{2i} - (1 - \beta_1)(1 - \beta_2)) \\
    z_{1i}(1 - z_{2i}) - \beta_2(1 - \beta_1) \\
    \beta_2 - z_{1i}(1 - z_{2i})
\end{bmatrix}
\]

The test statistic is \( \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S(\sqrt{n}P_n g(\cdot; \beta)) \) where \( S(\sqrt{n}P_n g(\cdot; \beta)) = n \sum_{q=1}^{Q} ((P_n g_q(\cdot; \beta))^{-2} \)

and \( \mathbb{B}_k(\gamma_0) \equiv \{ \beta \in \mathbb{B} : \beta_k = \gamma_0 \} \).

The non-recentered level \( \alpha \) one-sided test rejects when \( \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S(\sqrt{n}P_n g(\cdot; \beta)) > \hat{c}_{1-\alpha} \), where \( \hat{c}_{1-\alpha} \) is the \( (1 - \alpha) \) percentile of one of the following distributions:

1. **Numerical Bootstrap:** \( \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\frac{1}{\epsilon_n} Z_n^* g(\cdot; \beta)\right) - \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\frac{1}{\epsilon_n} P_n g(\cdot; \beta)\right) \).

2. **Minimum Resampling Test:** \( \min \left\{ \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\frac{1}{\epsilon_n} Z_n^* g(\cdot; \beta)\right), \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\frac{1}{\epsilon_n} Z_n^* g(\cdot; \beta)\right) \right\} \)

3. **Subsampling Test:** \( \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\sqrt{n}P_n g(\cdot; \beta)\right) - \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\sqrt{n}P_n g(\cdot; \beta)\right) \).

4. **Andrews and Soares (2010):** \( \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\frac{1}{\epsilon_n} Z_n^* g(\cdot; \beta)\right) \)

5. **Numerical Second Order Derivative:**

\[
\frac{1}{2} \left( \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\frac{1}{\epsilon_n} Z_{2n}^* g(\cdot; \beta)\right) - 2 \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\frac{1}{\epsilon_n} Z_n^* g(\cdot; \beta)\right) + \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\frac{1}{\epsilon_n} P_n g(\cdot; \beta)\right) \right)
\]

Here, \( Z_{2n}^* = P_n + 2 \epsilon_n \hat{g}_n^* \) and \( \mathbb{B}_k(\gamma_0) = \{ \beta : S(\sqrt{n}P_n g(\cdot; \beta)) \leq \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S(\sqrt{n}P_n g(\cdot; \beta)) + \sqrt{\log(n)}^{1/3} \} \) is an estimate of the identified set. Note that the second distribution corresponds to the minimum resampling test described in Bugni et al. (2014) when the Andrews and Soares (2010) GMS function is \( \phi_4(x) = x \). Also note that Bugni et al. (2014) is effectively setting \( \epsilon_n = \frac{\sqrt{\log(n)}}{\sqrt{n}} \), which we also use here, but without the recentering term \( \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S\left(\frac{1}{\epsilon_n} P_n g(\cdot; \beta)\right) \). The third distribution is the subsampling distribution discussed in remark 3.2 of Romano and Shaikh (2008) using \( b = n^{2/\beta} \).

We next consider the recentered test that rejects when \( \inf_{\beta \in \mathbb{B}_k(\gamma_0)} S(\sqrt{n}P_n g(\cdot; \beta)) - \inf_{\beta \in \mathbb{B}} S(\sqrt{n}P_n g(\cdot; \beta)) > \hat{c}_{1-\alpha}^* \), where \( \hat{c}_{1-\alpha}^* \) is the \( (1 - \alpha) \) percentile of the following two
recentered distributions

1. \( \inf_{\beta \in B_k(\gamma_0)} S \left( \frac{Z_n g(\cdot; \beta)}{\epsilon_n} \right) - \inf_{\beta \in B_k(\gamma_0)} S \left( \frac{P_n g(\cdot; \beta)}{\epsilon_n} \right) - \left( \inf_{\beta \in B} S \left( \frac{Z_n g(\cdot; \beta)}{\epsilon_n} \right) - \inf_{\beta \in B} \left( \frac{1}{\epsilon_n} P_n g(\cdot; \beta) \right) \right) \)

2. \( \inf_{\beta \in B_k(\gamma_0)} S \left( \sqrt{b} P_{bg}(\cdot; \beta) \right) - \inf_{\beta \in B_k(\gamma_0)} S \left( \sqrt{b} P_{bg}(\cdot; \beta) \right) - \left( \inf_{\beta \in B} S \left( \sqrt{b} P_{bg}(\cdot; \beta) \right) - \inf_{\beta \in B} \left( \sqrt{b} P_{bg}(\cdot; \beta) \right) \right) \)

Additionally, we also consider the test that rejects when

\[ \inf_{\beta \in B_k(\gamma_0)} S \left( \sqrt{n} P_n g(\cdot; \beta) \right) - \inf_{\beta \in B} S \left( \sqrt{n} P_n g(\cdot; \beta) \right) > \hat{c}_{1-\alpha} \]

which involves recentering only the test statistic, but not the distribution.

For the simulation exercise, we search over a grid of \( \beta \in [0.1, 0.5] \times [0.3, 0.7] \), setting \( \gamma_0 \) to be either \( \beta_1 \) or \( \beta_2 \). The true values are \( \beta_1 = 0.3 \) and \( \beta_2 = 0.5 \). The rejection frequencies when testing \( H_0 : \beta_1 = \gamma_0 \) against \( H_1 : \beta_1 \neq \gamma_0 \) for \( \gamma_0 \in [0.1, 0.5] \) are shown in figures 1 and 2.
Figure 1: Rejection frequency as a function of $\beta_1$

Rejection frequencies as a function of $\beta_1$ N=1000 B=1000

- Numerical Bootstrap (NB)
- Numerical Bootstrap Recentered
- NB Test Stat Recentered
- Second Order
- Andrews and Soares
- BCS
- Subsampling
- Subsampling Recentered
- Subsampling Test Stat Recentered
Notice that the non-recentered and recentered numerical bootstrap tests (red solid line, cross, and squares) have better finite sample power than the GMS test in Andrews and Soares (2010) (black dashed line), the minimum resampling test in Bugni et al. (2014) (green solid line with diamonds), and the non-recentered and recentered subsampling tests Politis et al. (1999) (blue dotted line, stars, and circles). For the numerical bootstrap and subsampling, recentering only the test statistic without recentering the distribution produces practically identical results as recentering both the test statistic and the distribution. Interestingly, using the second order numerical derivative (magenta dotted line) results in better finite sample power than subsampling and BCS, but not as much power as Andrews and Soares.

The rejection frequencies when testing $H_0 : \beta_2 = \gamma_0$ against $H_1 : \beta_2 \neq \gamma_0$ for $\gamma_0 \in [0.3, 0.7]$ are shown in figures 3 and 4.
Figure 3: Rejection frequency as a function of $\beta_2$
The rankings are almost the same as for $\beta_1$, except that the second order numerical derivative now performs worse than subsampling, while still performing better than BCS.

13 Empirical Application for Numerical Delta Method

Sometimes policy makers would like to determine for which individuals in a population a treatment is most effective or least effective. For instance, do smaller class sizes lead to the greatest/smallest improvement on the outcomes of high performing students or low performing students? One way to answer this question is to estimate quantile treatment effects at several different quantiles and then take their maximum or minimum. Because the pointwise maximum and minimum functions are directionally differentiable, we can use the numerical delta method to conduct inference.
We use the publicly available Tennessee STAR data inside the “AER” package of R which contains math and reading test scores, class size indicators, and covariates such as the teacher’s experience, whether the teacher has a higher degree, the teacher’s position on the career ladder, whether the student has free lunch, the student’s gender, race, and age, and whether the school is urban or rural. The observations are at the student-grade level over kindergarten through second grade, with different students entering the program at different grades. In order to run our regressions at the student level, we determine which grade the student first enters the program and use the variables corresponding to that grade in our regressions, which is in line with the approach taken in Chetty et al. (2011). We also use the same transformation discussed in Krueger (1999) to scale the test scores into percentile ranks and then take the average of the math and reading percentile ranks as our dependent variable $Y_i$. Our treatment variable is an indicator for whether a student is in a small classroom in the year in which she enters the program. The covariates are the teacher’s experience, the student’s gender, the student’s race, whether the student has free lunch, the teacher’s position on the career ladder, whether the teacher has a higher degree, the student’s age, and whether the school is urban or rural. Quantile regression was performed using the quantreg package in R.

More specifically, let $\hat{\theta}_n$ be the vector of quantile treatment effects and let $\hat{\theta}_n^*$ be the bootstrapped values. For $\phi(\theta) \equiv \max(\theta_1, \ldots, \theta_p)$ or $\min(\theta_1, \ldots, \theta_p)$ and $Z_n^* \equiv \sqrt{n} \left( \hat{\theta}_n^* - \hat{\theta}_n \right)$, we use the critical values of the distribution of the numerical derivative $\hat{\phi}_n^* (Z_n^*) \equiv \frac{\phi(\theta_n^*+\epsilon_n Z_n^*)-\phi(\theta_n)}{\epsilon_n}$ to form confidence intervals. The equal-tailed, upper, and lower nominal 95% confidence intervals using $B = 10000$ bootstrap iterations are shown in Table 1. The largest effect of 6.35% is at the 70th percentile while the smallest effect of 2.81% is at the 10th percentile.

<table>
<thead>
<tr>
<th></th>
<th>Point Estimate</th>
<th>Equal-Tailed</th>
<th>Upper</th>
<th>Lower</th>
<th>Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>6.35%</td>
<td>(4.21%, 6.91%)</td>
<td>($-\infty$,6.73%)</td>
<td>(4.45%, $\infty$)</td>
<td>70%</td>
</tr>
<tr>
<td>Min</td>
<td>2.81%</td>
<td>(2.29%, 5.07%)</td>
<td>($-\infty$,4.82%)</td>
<td>(2.47%, $\infty$)</td>
<td>10%</td>
</tr>
</tbody>
</table>
Conclusion

We showed how to conduct pointwise valid inference on Hadamard directionally differentiable functions $\phi(\cdot)$ using the numerical directional delta method. If we also assume that $\phi(\cdot)$ is convex and Lipschitz, then uniformly valid inference is possible. Additionally, we demonstrated how to consistently estimate the second order directional derivative and use the second order delta method to conduct pointwise valid inference in the partially identified models of Andrews and Soares (2010) and Bugni et al. (2014).

We proposed a numerical bootstrap principle which can be used to conduct consistent inference in situations where the regular bootstrap fails, such as the maximum score estimator, the LASSO estimator, and the 1-norm Support Vector Machine estimator. We demonstrated consistency of the numerical bootstrap for a large class of M-estimators and their constrained versions, whose asymptotic distribution involves a projection onto a cone that is difficult to estimate. We also discussed applications to misspecified GMM models and models with an unknown polynomial convergence rate. We demonstrated through monte carlo simulations that the numerical bootstrap offers a viable competing approach to subsampling and the methods proposed in Andrews and Soares (2010) and Bugni et al. (2014) both in terms of size control and finite sample power.

While the order of magnitude of the error of approximation in the numerical bootstrap can be heuristically understood, we realize that the choice of the step size parameter is a difficult task much like the choice of the block size in subsampling. We intend to study the adaptive choice of the step size parameter in future investigations.

References


Bugni, F., I. Canay, and X. Shi (2014): “Inference for functions of partially identified parameters in moment inequality models,” Duke University, Northwestern University and University of Wisconsin, cemmap working paper, Centre for Microdata Methods and Practice. 3, 4, 15, 17, 39, 40, 43, 47


Woutersen, T. and J. C. Ham (2013): “Calculating confidence intervals for continuous and discontinuous functions of parameters,” cemmap working paper, Centre for Microdata Methods and Practice. 2


50
A Appendix

A.1 List of Commonly Used Symbols

\( P_n \) empirical measure
\( P_n^* \) bootstrap empirical measure
\( Z_n^* \) numerical bootstrap empirical measure
\( \rightsquigarrow \) weak convergence
\( \overset{F}{\rightsquigarrow} \) weak convergence conditional on the data
\( Z_n^* \) consistent estimate of \( G_0 \)
\( \theta^- \) \(-\min(\theta, 0)\)
\( \theta^+ \) \(\max(\theta, 0)\)
\( \rho_{BL_1}(F_1, F_2) \) \(\sup_{f \in BL_1} |Ef(F_1) - Ef(F_2)|\)

\( BL_1 \) the space of Lipschitz functions \( f : D \mapsto \mathbb{R} \) with Lipschitz norm bounded by 1

A.2 Verification of Lipschitz property of \( \phi(\cdot) \) in Fang and Santos (2014) examples

Fang and Santos (2014) Example 2.1 \( \phi(\theta) = |\theta|, D = \mathbb{R}, E = \mathbb{R} \).

\[
\|\phi(\theta + h) - \phi(\theta)\|_E = |\phi(\theta + h) - \phi(\theta)| = ||\theta + h| - |\theta|| \leq |h| \equiv \|h\|_D
\]

Fang and Santos (2014) Example 2.2 \( \phi(\theta) = \max\{\theta^{(1)}, \theta^{(2)}\}, D = \mathbb{R}^2, E = \mathbb{R} \).

\[
\|\phi(\theta + h) - \phi(\theta)\|_E = |\phi(\theta + h) - \phi(\theta)| = |\max\{\theta^{(1)} + h^{(1)}, \theta^{(2)} + h^{(2)}\} - \max\{\theta^{(1)}, \theta^{(2)}\}|
\]

\[
= \begin{cases} 
|\theta^{(1)}|, & \theta^{(1)} + h^{(1)} \geq \theta^{(2)} + h^{(2)} \text{ and } \theta^{(1)} \geq \theta^{(2)} \\
|\theta^{(1)} - \theta^{(2)} + h^{(1)}|, & \theta^{(1)} + h^{(1)} \geq \theta^{(2)} + h^{(2)} \text{ and } \theta^{(1)} < \theta^{(2)} \\
|\theta^{(2)} - \theta^{(1)} + h^{(2)}|, & \theta^{(1)} + h^{(1)} < \theta^{(2)} + h^{(2)} \text{ and } \theta^{(1)} \geq \theta^{(2)} \\
|h^{(2)}|, & \theta^{(1)} + h^{(1)} < \theta^{(2)} + h^{(2)} \text{ and } \theta^{(1)} < \theta^{(2)} 
\end{cases}
\]

\[
\leq 2(|h^{(1)}| + |h^{(2)}|) \equiv 2 \|h\|_{\mathbb{R}^2}
\]
Fang and Santos (2014) Example 2.3 \( \phi(\theta) = \sup_{f \in \mathcal{F}} \theta(f) \), \( \mathbb{D} = \ell^\infty(\mathcal{F}) \),

\[
\|\phi(\theta + h) - \phi(\theta)\|_\mathbb{E} \leq \sup_{f \in \mathcal{F}} |\phi(\theta + h) - \phi(\theta)| = \left| \sup_{f \in \mathcal{F}} (\theta(f) + h(f)) - \sup_{f \in \mathcal{F}} \theta(f) \right| \\
\leq \sup_{f \in \mathcal{F}} |h(f)| \equiv \|h\|_{\ell^\infty(\mathcal{F})}
\]

Fang and Santos (2014) Example 2.4 For any \( \lambda \) in a convex, compact set \( \Lambda \subseteq \mathbb{R}^d \), \( \phi(\theta) = \sup_{p \in \mathbb{S}^d} \{ \langle p, \lambda \rangle - \theta(p) \} \), \( \mathbb{D} = \mathcal{C}(\mathbb{S}^d) \), \( \mathbb{E} = \mathbb{R} \)

\[
\|\phi(\theta + h) - \phi(\theta)\|_\mathbb{E} = |\phi(\theta + h) - \phi(\theta)| \\
= \left| \sup_{p \in \mathbb{S}^d} \{ \langle p, \lambda \rangle - \theta(p) - h(p) \} - \sup_{p \in \mathbb{S}^d} \{ \langle p, \lambda \rangle - \theta(p) \} \right| \leq \sup_{p \in \mathbb{S}^d} |h(p)| \equiv \|h\|_{\mathcal{C}(\mathbb{S}^d)}
\]

Fang and Santos (2014) Example 2.5 \( \phi((\theta^{(1)}, \theta^{(2)})) = \int_\mathbb{R} \max \{ \theta^{(1)}(u) - \theta^{(2)}(u), 0 \} w(u)du \), where \( w : \mathbb{R} \rightarrow \mathbb{R}_+ \) is a positive, integrable weighting function. \( \mathbb{D} = \ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R}) \), \( \mathbb{E} = \mathbb{R} \).

\[
\|\phi(\theta + h) - \phi(\theta)\|_\mathbb{E} = |\phi((\theta^{(1)} + h^{(1)}, \theta^{(2)} + h^{(2)})\) - \phi((\theta^{(1)}, \theta^{(2)}))| \\
= \left| \int_\mathbb{R} \max \{ \theta^{(1)}(u) - \theta^{(2)}(u) + h^{(1)}(u) - h^{(2)}(u), 0 \} w(u)du - \int_\mathbb{R} \max \{ \theta^{(1)}(u) - \theta^{(2)}(u), 0 \} w(u)du \right| \\
\leq C \sup_{u \in \mathbb{R}} |h^{(1)}(u) - h^{(2)}(u)| \text{, where } C = \int_\mathbb{R} w(u)du \\
\leq C (\sup_{u \in \mathbb{R}} |h^{(1)}(u)| + \sup_{u \in \mathbb{R}} |h^{(2)}(u)|) \equiv C \| (h^{(1)}, h^{(2)}) \|_{\ell^\infty(\mathbb{R}) \times \ell^\infty(\mathbb{R})}
\]
**Fang and Santos (2014) Convex Projection**  
Let $\mathbb{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and norm $\| \cdot \|_{\mathbb{H}}$. Let $\Lambda \subseteq \mathbb{H}$ be a known closed and convex set. Consider the following convex set projection function.

$$
\phi(\theta) \equiv \|\theta - \Pi_{\Lambda} \theta\|_{\mathbb{H}} = \inf_{\lambda \in \Lambda} \|\theta - \lambda\|_{\mathbb{H}} 
$$

(31)

It is straightforward to show that the projection function (regardless of whether $\Lambda$ is convex or not) is Lipschitz in the argument:

$$
|\phi(\theta_1) - \phi(\theta_2)| \leq \|\theta_1 - \theta_2\|_{\mathbb{H}}.
$$

(32)

To see this, consider two cases. Suppose first that $\phi(\theta_1) \geq \phi(\theta_2)$, then by the definition of $\Pi_{\Lambda}(\theta_1)$,

$$
|\phi(\theta_1) - \phi(\theta_2)| \leq \|\theta_1 - \Pi_{\Lambda} \theta_2\|_{\mathbb{H}} - \|\theta_2 - \Pi_{\Lambda} \theta_2\|_{\mathbb{H}} \leq \|\theta_1 - \theta_2\|_{\mathbb{H}}
$$

(33)

where the second inequality is the triangular inequality. Similarly, in the second case when $\phi(\theta_1) \leq \phi(\theta_2)$, then by the definition of $\Pi_{\Lambda}(\theta_2)$,

$$
|\phi(\theta_1) - \phi(\theta_2)| \leq \|\theta_2 - \Pi_{\Lambda} \theta_1\|_{\mathbb{H}} - \|\theta_2 - \Pi_{\Lambda} \theta_1\|_{\mathbb{H}} \leq \|\theta_1 - \theta_2\|_{\mathbb{H}}
$$

(34)

### A.3 Convexity of $\phi(\cdot)$ in Fang and Santos (2014) Examples

**Fang and Santos (2014) Example 2.1**  
For any $\lambda \in [0, 1]$ and $\theta_1, \theta_2 \in \mathbb{R}$,

$$
\phi(\lambda \theta_1 + (1 - \lambda) \theta_2) = |\lambda \theta_1 + (1 - \lambda) \theta_2| \leq \lambda |\theta_1| + (1 - \lambda) |\theta_2| = \lambda \phi(\theta_1) + (1 - \lambda) \phi(\theta_2)
$$

(35)

**Fang and Santos (2014) Example 2.2**  
For any $\lambda \in [0, 1]$ and $\theta_1, \theta_2 \in \mathbb{D}$,

$$
\phi(\lambda \theta_1 + (1 - \lambda) \theta_2) = \max\{\lambda \theta_1^{(1)} + (1 - \lambda) \theta_2^{(1)}, \lambda \theta_1^{(2)} + (1 - \lambda) \theta_2^{(2)}\} 
\leq \lambda \max\{\theta_1^{(1)}, \theta_1^{(2)}\} + (1 - \lambda) \max\{\theta_2^{(1)}, \theta_2^{(2)}\} = \lambda \phi(\theta_1) + (1 - \lambda) \phi(\theta_2)
$$

**Fang and Santos (2014) Example 2.3**  
For any $\lambda \in [0, 1]$ and $\theta_1, \theta_2 \in \mathbb{D}$,

$$
\phi(\lambda \theta_1 + (1 - \lambda) \theta_2) = \sup_{f \in \mathbb{F}} \{\lambda \theta_1(f) + (1 - \lambda) \theta_2(f)\}
$$
\[
\leq \lambda \sup_{f \in F} \theta_1(f) + (1 - \lambda) \sup_{f \in F} \theta_2(f) = \lambda \phi(\theta_1) + (1 - \lambda) \phi(\theta_2)
\]

**Fang and Santos (2014) Example 2.5** Note that \(\phi((\theta^{(1)}, \theta^{(2)})) = \int_{\mathbb{R}} \max\{\theta^{(1)}(u) - \theta^{(2)}(u), 0\} w(u) du\), where \(w : \mathbb{R} \to \mathbb{R}_{+}\), can be written as \(h(\theta)\), where \(g(\theta) = g(\theta^{(1)}, \theta^{(2)}) = \max\{\theta^{(1)} - \theta^{(2)}, 0\}\) and \(h(\gamma) = \int_{\mathbb{R}} \gamma(u) w(u) du\). We can show that \(g(\theta)\) is convex and \(h(\gamma)\) is linear and nondecreasing, which implies that their composition \(h(g(\theta^{(1)}, \theta^{(2)}))\) is convex. \(g(\theta)\) is convex because for any \(\lambda \in [0, 1]\) and \(\theta_1, \theta_2 \in \mathbb{D}\),

\[
g(\lambda \theta_1 + (1 - \lambda) \theta_2) = \max\{\lambda \theta_1^{(1)} + (1 - \lambda) \theta_2^{(1)} - (\lambda \theta_1^{(2)} + (1 - \lambda) \theta_2^{(2)}), 0\}
\leq \lambda \max\{\theta_1^{(1)} - \theta_2^{(2)}, 0\} + (1 - \lambda) \max\{\theta_2^{(1)} - \theta_2^{(2)}, 0\}
= \lambda g(\theta_1) + (1 - \lambda) g(\theta_2).
\]

Also \(h(\gamma) = \int_{\mathbb{R}} \gamma(u) w(u) du\) is nondecreasing because \(w(u)\) is positive, and it’s linear because integration is a linear operator.

**Convex Set Projection** Let \(\mathbb{H}\) be a Hilbert space with inner product \(\langle \cdot, \cdot \rangle_{\mathbb{H}}\) and norm \(\|\cdot\|_{\mathbb{H}}\). Let \(\Lambda \subseteq \mathbb{H}\) be a known closed and convex set. \(\phi(\theta) \equiv \|\theta - \Pi_{\Lambda} \theta\|_{\mathbb{H}} = \inf_{v \in \Lambda} \|\theta - v\|_{\mathbb{H}}\) can be interpreted as the shortest distance between \(\theta\) and a point in \(\Lambda\). Because \(\Lambda\) is a closed convex subset of a Hilbert space, by the Nearest Point Theorem (Theorem 6.53 Aliprantis and Border (1999)), for each \(\theta \in \mathbb{H}\), there exists a unique \(\Pi_{\Lambda} \theta \in \Lambda\) satisfying \(\|\theta - \Pi_{\Lambda} \theta\|_{\mathbb{H}} \leq \|\theta - v\|_{\mathbb{H}}\) for all \(v \in \Lambda\).

Therefore, for any \(\lambda \in [0, 1]\) and \(\theta_1, \theta_2 \in \mathbb{H}\),

\[
\phi(\lambda \theta_1 + (1 - \lambda) \theta_2) = \|\lambda \theta_1 + (1 - \lambda) \theta_2 - \Pi_{\Lambda}(\lambda \theta_1 + (1 - \lambda) \theta_2)\|_{\mathbb{H}}
\leq \|\lambda \theta_1 + (1 - \lambda) \theta_2 - (\lambda \Pi_{\Lambda}(\theta_1) + (1 - \lambda) \Pi_{\Lambda}(\theta_2))\|_{\mathbb{H}}
\leq \lambda \|\theta_1 - \Pi_{\Lambda} \theta_1\|_{\mathbb{H}} + (1 - \lambda) \|\theta_2 - \Pi_{\Lambda} \theta_2\|_{\mathbb{H}}
= \lambda \phi(\theta_1) + (1 - \lambda) \phi(\theta_2)
\]

The first inequality follows from the fact that \(\lambda \Pi_{\Lambda}(\theta_1) + (1 - \lambda) \Pi_{\Lambda}(\theta_2) \in \Lambda\) since \(\Lambda\) is convex.
A.4 Proofs for the Theorems

Proof of Theorem 3.1 Part 1 is exactly in Theorem 2.1 of Fang and Santos (2014). We make use of Lemma A.2 to show part 2. Let the following measurability assumptions hold:

- $f(Z_n^*)$ is measurable for every bounded, continuous map $f : \mathbb{D} \mapsto \mathbb{R}$.
- $g_0$ is Borel measurable and separable.

Define $g_n(h) = \frac{1}{\epsilon_n} (\phi(\theta_0 + \epsilon_n h) - \phi(\theta_0))$. By Hadamard directional differentiability, for $h_n \to h$, $h \in \mathbb{D}_0$, $\theta + \epsilon_n h_n \in \mathbb{D}$, $g_n(h_n) \to g(h) = \phi'_{\theta_0}(h)$. Then we write

$$ \frac{\phi(\hat{\theta}_n + \epsilon_n Z_n^*) - \phi(\hat{\theta}_n)}{\epsilon_n} = \frac{\phi(\hat{\theta}_n + \epsilon_n (Z_n^* + \hat{\theta}_n - \theta_0)) - \phi(\hat{\theta}_n + \epsilon_n (\hat{\theta}_n - \theta_0))}{\epsilon_n} \to g_n(\hat{\theta}_n - \theta_0) $$

$$ = g_n(Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n}) - g_n(\epsilon_n^{-1}(\hat{\theta}_n - \theta_0)) $$

Since $Z_n^* \overset{p}{\rightsquigarrow} G_0$, $\epsilon_n^{-1}(\hat{\theta}_n - \theta_0) = o_P(1)$, $Z_n^* + \epsilon_n^{-1}(\hat{\theta}_n - \theta_0) \overset{p}{\rightsquigarrow} G_0$ (also $\rightsquigarrow G_0$).

Applying Lemma A.2 to the first term on the right side, the first term $\overset{p}{\rightsquigarrow} \phi'_{\theta_0}(G_0)$ (and also $\rightsquigarrow \phi'_{\theta_0}(G_0)$ by van der Vaart and Wellner (1996) Theorem 1.11.1). The second term is $o_P(1)$ by van der Vaart and Wellner (1996) Theorem 1.11.1. Since $X_n \overset{p}{\rightsquigarrow} X$ and $Y_n = o_P(1)$ implies $X_n + Y_n \overset{p}{\rightsquigarrow} X$, summing the two terms on the right hand side leads to

$$ \frac{\phi(\hat{\theta}_n + \epsilon_n Z_n^*) - \phi(\hat{\theta}_n)}{\epsilon_n} \overset{p}{\rightsquigarrow} \phi'_{\theta_0}(G_0) $$

Convexity implies Subadditivity

Lemma A.1 When $\phi(\cdot)$ is convex and directionally differentiable at $\theta_0$, $\forall 0 \leq \lambda \leq 1$,

$$ \phi'_{\theta_0}(h_1 + h_2) \leq \phi'_{\theta_0}(h_1) + \phi'_{\theta_0}(h_2), \quad \phi'_{\theta_0}(\lambda h_1 + (1 - \lambda) h_2) \leq \lambda \phi'_{\theta_0}(h_1) + (1 - \lambda) \phi'_{\theta_0}(h_2). $$

(36)

Proof of Lemma A.1 For $h_1$ and $h_2$, it follows from the convexity of $\phi(\cdot)$ that

$$ \phi(\theta_0 + t(h_1 + h_2)) = \phi\left( \frac{1}{2} (\theta_0 + 2th_1) + \frac{1}{2} (\theta_0 + 2th_2) \right) \leq \frac{1}{2} \phi(\theta_0 + 2th_1) + \frac{1}{2} \phi(\theta_0 + 2th_2) $$

\[55\]
Hence \( \phi(\theta_0 + t(h_1 + h_2)) - \phi(\theta_0) \leq \frac{1}{2} (\phi(\theta_0 + 2th_1) - \phi(\theta_0)) + \frac{1}{2} (\phi(\theta_0 + 2th_2) - \phi(\theta_0)) \)
and
\[
\frac{\phi(\theta_0 + t(h_1 + h_2)) - \phi(\theta_0)}{t} \leq \frac{\phi(\theta_0 + 2th_1) - \phi(\theta_0)}{2t} + \frac{\phi(\theta_0 + 2th_2) - \phi(\theta_0)}{2t}.
\]

Taking \( t \to 0 \) on both sides we conclude that \( \phi'_{\theta_0}(h_1 + h_2) \leq \phi'_{\theta_0}(h_1) + \phi'_{\theta_0}(h_2) \).

**Proof of Theorem 3.2** We first note that the arguments in the proofs of Theorem 2.11 in Bhattacharya and Rao (1986) can be revised for a convergence in probability version: Let \( C \) be a class of convex sets such that \( \sup_{P \in \mathcal{P}} P(G_0 \in \partial C) = 0 \) for all \( C \in \mathcal{C} \). Then \( \forall \epsilon > 0, \)
\[
\lim_{n \to \infty} \sup_{P \in \mathcal{P}} \left( \sup_{C \in \mathcal{C}} |P(Z_n^* \in C|X_n) - P(G_0 \in C)| \geq \epsilon \right) \to 0,
\]
whenever \( \forall \epsilon > 0, \lim_{n \to \infty} \sup_{P \in \mathcal{P}} P(\rho_{BL_1}(Z_n^*, G_0) \geq \epsilon) \to 0 \). Under assumption 3.2, the key to invoking (37) is the fact that level sets of convex functions are convex.

Under assumption 3.1 part (i), \( \forall \epsilon > 0, \sup_{P \in \mathcal{P}} P \left( \left| \frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta(P) \right) \right| \geq \epsilon \right) = o(1) \). It can therefore be combined with part (ii) of assumption 3.1 to show that
\[
\sup_{P \in \mathcal{P}} P \left( \frac{\sum_{n \in \mathcal{P}} P(\hat{\theta}_n - \theta(P) \geq \epsilon)}{\epsilon_n} \right) = o(1).
\]

Next note that the set \( C = \{ g : \frac{1}{\epsilon_n} \left( \phi(\theta_0 + \epsilon_n g) - \phi(\hat{\theta}_n) \right) \leq x \} \) is convex whenever \( \phi(\cdot) \) is a convex function and is a member of the class specified in assumption 3.2. Therefore by (38) and (37),
\[
\sup_{P \in \mathcal{P}} P \left( \sup_{x} \left| P \left( \phi_n'(Z_n^*) \leq x \right) - P \left( \frac{1}{\epsilon_n} \left( \phi(\theta_0 + \epsilon_n G_0) - \phi(\hat{\theta}_n) \right) \leq x \right) \geq \epsilon \right) \right) = o(1).
\]

Finally, we use the last condition in the theorem statement to show that
\[
\sup_{P \in \mathcal{P}} \sup_{x \in \mathcal{A}} |P \left( \frac{1}{\epsilon_n} \left( \phi(\theta_0 + \epsilon_n G_0) - \phi(\hat{\theta}_n) \right) \leq x \right) - J_{\epsilon_n}(x, G_0) | = o(1).
\]
Note that we do not need the last equation or the last condition of the theorem if we replace \( \hat{\theta}_n \) in \( \hat{\phi}'_n(\hat{Z}_n^*) \) by a fixed \( \theta \) in hypothesis testing settings. Similarly, the level set

\[ C = \{ g : r_n \left( \phi(\theta(P) + r_n^{-1}g) - \phi(\theta(P)) \right) \leq x \} \]

is also convex. By Theorem 2.11 of Bhattacharya and Rao (1986), part (i) of A3.1, and A3.2,

\[
\sup_{P \in \mathcal{P}} \sup_{x \in A} |J_n(x, P) - J_n(x, G_0)| = o(1). \tag{41}
\]

The first conclusion of the theorem then follows from combining (39), (40), (41) and (9). The second conclusion follows from similar arguments to Lemma A.1(vi) in Romano and Shaikh (2012).

**Proof of Theorem 3.3** Consider any sequence \( \{P_n \in \mathcal{P} : n \geq 1\} \) that determines \( \theta_n = \theta(P_n) \) and the laws of \( r_n \left( \hat{\theta}_n - \theta(P_n) \right), G_0, \) and \( Z_n^* \). Note that assumptions 3.1 and 3.2 imply the following:

**Assumption A.1** Let the sequence \( \theta_n, P_n \) and \( G^0_n \) be such that

\[ \rho_{BL_1} \left( r_n \left( \hat{\theta}_n - \theta_n \right), G^0_n \right) = o(1) \quad \text{and} \quad \rho_{BL_1} \left( Z_n^*, G^0_n \right) = o_{P_n}(1). \]

**Assumption A.2** For all \( \epsilon \) small enough, and all \( x = J_n^{-1}(1 - \tau - \epsilon, P_n), x_n \) is a sequence of asymptotic equicontinuity points of \( J(x) \) being either \( J_{\epsilon_n}(x, G^0_n) \) or \( J_n(x, G^0_n) \):

\[
\lim_{\delta \to 0} \lim_{n \to \infty} \sup_{|x - x_n| \leq \delta} |J(x) - J(x_n)| = 0.
\]

First \( \rho_{BL_1}(Z_n^*, G_0) = o_{P_n}(1) \) in part (ii) of assumption A.1 implies that \( \rho_{BL_1}(Z_n^* + o_{P_n}(1), G_0) = o_{P_n}(1) \), which follows from

\[
\sup_{h \in BL_1} |E_M h (Z_n^* + o_{P_n}(1)) - Eh(G_0)| \leq \sup_{h \in BL_1} |E_M h (Z_n^*) - Eh(G_0)| + o_{P_n}(1).
\]
Since $r_n \epsilon_n \to \infty$, $\epsilon_n^{-1} \left( \hat{\theta}_n - \theta_n \right) = o_{P_n}(1)$, so that

$$\rho_{BL1} \left( \mathbb{E}_n^* + \frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_n \right) , \mathbb{G}_0 \right) = o_{P_n}(1).$$

Next note that the functions $g_n(Z) = \frac{1}{\epsilon_n} \left( \phi(\theta_n + \epsilon_n Z) - \phi(\theta_n) \right)$ are uniformly Lipschitz in $Z$ with the Lipschitz constant bounded by that of $\phi(\cdot)$. The same arguments in Proposition 10.7 of Kosorok (2007) adapted to a sequence of such functions $g_n(\cdot)$ show that

$$\rho_{BL1} \left( g_n \left( \mathbb{E}_n^* + \frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_n \right) \right) , g_n (\mathbb{G}_0) \right) = o_{P_n}(1).$$

Then we can write

$$\hat{\phi}_n'(Z_n^*) = g_n \left( \mathbb{E}_n^* + \frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_n \right) \right) - \frac{1}{\epsilon_n} \left( \phi(\hat{\theta}_n) - \phi(\theta_n) \right) = g_n \left( \mathbb{E}_n^* + \frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_n \right) \right) + o_{P_n}(1),$$

so that also $\rho_{BL1} \left( \hat{\phi}_n'(Z_n^*) , g_n (\mathbb{G}_0) \right) = o_{P_n}(1)$. Then using assumption A.2, similar arguments to those in Lemma 10.11 in Kosorok (2007) can be used to establish

$$J_{\epsilon_n}(x_n, P_n) - J_{\epsilon_n}(x_n, \mathbb{G}_0) = o_{P_n}(1). \quad (42)$$

Next using part (i) of assumption A.1 and applying a nonstochastic version of the arguments in Proposition 10.7 of Kosorok (2007), it can be shown that

$$\rho_{BL1} \left( J_{\epsilon_n}(\cdot, P_n) , J_{\epsilon_n}(\cdot, \mathbb{G}_0) \right) = o(1).$$

The $J_{\epsilon_n}(\cdot, \mathbb{G}_0)$ part of assumption A.2 in combination with modified arguments in Lemma 10.11 in Kosorok (2007) produces that

$$J_{\epsilon_n}(x_n, P_n) - J_{\epsilon_n}(x_n, \mathbb{G}_0) = o(1). \quad (43)$$

When $\phi(\cdot)$ satisfies assumption 3.3, equations (8),(9), (42), and (43) imply that for all $\epsilon, \eta > 0$ and $n$ large enough,

$$P_n \left( J_{\epsilon_n}(x_n, P_n) \leq J_n(x_n, P_n) + \epsilon \right) \geq 1 - \delta.$$
Next we consider arguments similar to Lemma A.1 parts (i) and (vi) and Theorem 2.4 in Romano and Shaikh (2012), If \( J_{\epsilon_n}(x_n, P_n) \leq J_n(x_n, P_n) + \epsilon \) at \( x_n = J_n^{-1}(1 - \alpha - \epsilon, P_n) \), then \( J_{\epsilon_n}^{-1}(1 - \alpha, P_n) \geq J_n^{-1}(1 - \alpha - \epsilon, P_n) \). Combining these inequalities,

\[
P_n \left( r_n \left( \phi \left( \hat{\theta}_n \right) - \phi (\theta_n) \right) \right) \leq J_{\epsilon_n}^{-1}(1 - \alpha, P_n)
\]
\[
\geq P_n \left( r_n \left( \phi \left( \hat{\theta}_n \right) - \phi (\theta_n) \right) \right) \leq J_{\epsilon_n}^{-1}(1 - \alpha, P_n) \cap J_{\epsilon_n}(x_n, P_n) \leq J_n(x_n, P_n) + \epsilon
\]
\[
\geq P_n \left( r_n \left( \phi \left( \hat{\theta}_n \right) - \phi (\theta_n) \right) \right) \leq J_{\epsilon_n}^{-1}(1 - \alpha - \epsilon, P_n) \cap J_{\epsilon_n}(x_n, P_n) \leq J_n(x_n, P_n) + \epsilon
\]
\[
\geq P_n \left( r_n \left( \phi \left( \hat{\theta}_n \right) - \phi (\theta_n) \right) \right) \leq J_n^{-1}(1 - \alpha - \epsilon, P_n)
\]
\[
\geq 1 - \alpha - \epsilon - \delta.
\]

Since both \( \epsilon \) and \( \delta \) can be arbitrarily small,

\[
\limsup_{n \to \infty} P_n \left( r_n \left( \phi \left( \hat{\theta}_n \right) - \phi (\theta_n) \right) \right) \geq \hat{c}_{1-\tau} \leq \tau.
\]

Now define

\[
\beta = \limsup_{n \to \infty} \sup_{P \in \mathcal{P}} P \left( r_n \left( \phi \left( \hat{\theta}_n \right) - \phi (\theta(P)) \right) \geq \hat{c}_{1-\alpha} \right).
\]

Then one can find a sequence of \( P_n \in \mathcal{P} \) such that, for \( \theta_n = \theta_{P_n} \),

\[
\beta = \lim_{n \to \infty} P_n \left( r_n \left( \phi \left( \hat{\theta}_n \right) - \phi (\theta_n) \right) \geq \hat{c}_{1-\alpha} \right).
\]

Find a subsequence \( \mu_n \) of \( n \) for which \( \theta_n \) converges, with its limit denoted \( \theta \). The previous arguments allow us to claim that

\[
\limsup_{\mu_n \to \infty} P_{\mu_n} \left( r_{\mu_n} \left( \phi \left( \hat{\theta}_n \right) - \phi (\theta_{\mu_n}) \right) \geq \hat{c}_{1-\alpha} \right) \leq \alpha.
\]

Since \( P_{\mu_n}, \theta_{\mu_n} \) is a subsequence of \( P_n, \theta_n \), it is also the case that

\[
\beta = \lim_{\mu_n \to \infty} P_{\mu_n} \left( r_{\mu_n} \left( \phi \left( \hat{\theta}_n \right) - \phi (\theta_{\mu_n}) \right) \geq \hat{c}_{1-\alpha} \right) \leq \alpha.
\]

Now suppose \( \phi (\cdot) \) does not satisfy assumption 3.3, but assumptions 3.4 and 3.5 hold. For \( t_n = r_n^{-1}, \epsilon_n \), define \( g_{t_n,n}(h) = \frac{1}{t_n} (\phi (\theta_n + t_n h) - \phi (\theta_n)) \). We first show that
\[ \rho_{BL} \left( \hat{\phi}_n^\prime (Z_n^*), g_{\epsilon_n,n} (G_0) \right) = o_P (1), \quad \text{and} \quad \rho_{BL} \left( r_n \left( \phi (\hat{\theta}_n) - \phi (\theta_n) \right), g_{r_n^{-1},n} (G_0) \right) = o (1). \]

Next we can use A3.5 to show that both
\[
\rho_{BL} \left( g_{\epsilon_n,n} (G_0), \varphi_n^\prime \left( \frac{\theta_n}{\epsilon_n} + G_0 \right) - \varphi_{0}^\prime \left( \frac{\theta_n}{\epsilon_n} \right) \right) = o (1)
\]
\[
\rho_{BL} \left( g_{r_n^{-1},n} (G_0), \varphi_n^\prime (\theta_n r_n + G_0) - \varphi_{0}^\prime (\theta_n r_n) \right) = o (1)
\]

Define then \( J_{\epsilon_n}^\prime (x, G_0) \) and \( J_n^\prime (x, G_0) \), respectively, as the CDFs of \( \varphi_{0}^\prime \left( \frac{\theta_n}{\epsilon_n} + G_0 \right) - \varphi_{0}^\prime \left( \frac{\theta_n}{\epsilon_n} \right) \) and \( \varphi_{0}^\prime (\theta_n r_n + G_0) - \varphi_{0}^\prime (\theta_n r_n) \). Using AA.2 and arguments analogous to Lemma 10.11 in Kosorok (2007) then shows that for each sequence \( x_n \) of asymptotic equicontinuity points,
\[
J_{\epsilon_n} (x_n, P_n) - J_{\epsilon_n} (x_n, G_0) = o_P (1), \quad J_n (x_n, P_n) - J_n (x_n, G_0) = o (1) ,
\]
\[
J_{\epsilon_n} (x_n, G_0) - J_{\epsilon_n}^\prime (x_n, G_0) = o (1) , \quad J_n (x_n, G_0) - J_n^\prime (x_n, G_0) = o (1) .
\]

So that \( J_{\epsilon_n} (x_n, P_n) - J_{\epsilon_n}^\prime (x_n, G_0) = o_P (1) \) and \( J_n (x_n, P_n) - J_n^\prime (x_n, G_0) = o (1) \). Then by A3.4, \( J_{\epsilon_n}^\prime (x_n, G_0) \leq J_n^\prime (x_n, G_0) \).

Then for each \( \epsilon > 0 \) and \( n \) sufficiently large, \( J_n (x_n, P_n) \geq J_n^\prime (x_n, G_0) - \frac{\epsilon}{2} \), and
\[
\lim_{n \to \infty} P_n \left( J_{\epsilon_n} (x_n, P_n) \leq J_n (x_n, P_n) + \epsilon \right)
\geq \lim_{n \to \infty} P_n \left( J_{\epsilon_n} (x_n, P_n) \leq J_n^\prime (x_n, G_0) + \frac{\epsilon}{2} \right) \to 1.
\]

**Proof for Theorem 4.1** The first part of the theorem is exactly Theorem 2 in Römisch (2005). The second part will be argued using Lemma A.2. Define \( g_n (h) = \frac{1}{\epsilon_n^2} (\phi (\theta_0 + \epsilon_n h) - \phi (\theta_0) - \phi_{\theta_0}^\prime (h)) \). By definition of (14), for \( h_n \to h, h \in \mathbb{D}_0, \theta + \epsilon_n h_n \in \mathbb{D}, g_n (h_n) \to g (h) = \frac{1}{2} \phi_{\theta_0}^\prime (h) \). Then write, noting that \( \phi_{\theta_0}^\prime (h) \equiv 0 \),
\[
\frac{\phi (\hat{\theta}_n + \epsilon_n Z_n^*) - \phi (\hat{\theta}_n)}{\epsilon_n^2} = \frac{\phi (\theta_0 + \epsilon_n (Z_n^* + \hat{\theta}_n - \theta_0))}{\epsilon_n^2} - \phi (\theta_0) = \frac{\phi (\theta_0 + \epsilon_n (\epsilon_n^{-1} (\hat{\theta}_n - \theta_0)))}{\epsilon_n^2} - \phi (\theta_0)
\]
\[
= g_n \left( Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) - g_n \left( \epsilon_n^{-1} (\hat{\theta}_n - \theta_0) \right).
\]
Since $Z_n \xrightarrow{P} G_0$, $\epsilon_n^{-1} \left( \hat{\theta}_n - \theta_0 \right) = o_P(1)$, $Z_n + \epsilon_n^{-1} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{w} G_0$ (also $\sim G_0$). Apply Lemma A.2 to the first term on the right side, the first term $\xrightarrow{w} \frac{1}{2} \phi''_{\theta_0}(G_0)$ (and also $\sim \frac{1}{2} \phi''_{\theta_0}(G_0)$ by van der Vaart and Wellner (1996) Theorem 1.11.1). The second term is $o_P(1)$ by van der Vaart and Wellner (1996) Theorem 1.11.1. Since $X_n \xrightarrow{W} X$ and $Y_n = o_P(1)$ implies $X_n + Y_n \xrightarrow{P} X$, summing the two terms on the right hand side leads to

$$\frac{\phi(\hat{\theta}_n + \epsilon_n Z_n^*) - \phi(\hat{\theta}_n)}{\epsilon_n^2} \xrightarrow{W} \frac{1}{2} \phi''_{\theta_0}(G_0).$$

**Proof for Theorem 4.2** Note $\frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{P} 0$, $\frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_0 + \epsilon_n Z_n^* \right) \xrightarrow{P} G_0$, and $\frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_0 + 2 \epsilon_n \hat{G}_n \right) \xrightarrow{w} G_0$. For $g_n(h) = \frac{1}{\epsilon_n} \left( \phi(\theta_0 + \epsilon_n h) - \phi(\theta_0) - \phi''_{\theta_0}(h) \right)$ and $g(h) = \frac{1}{\epsilon_n} \phi''_{\theta_0}(h)$, $g_n(h_n) \rightarrow g(h)$ when $h_n \rightarrow h$. Then by van der Vaart and Wellner (1996) Theorem 1.11.1 and Lemma A.2, jointly,

$$\frac{1}{\epsilon_n^2} \left[ \phi(\hat{\theta}_n + 2 \epsilon_n Z_n^*) - \phi(\theta_0) - \phi''_{\theta_0}(\hat{\theta}_n - \theta_0 + 2 \epsilon_n Z_n^*) \right] = g_n \left( 2 Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) \xrightarrow{w} 0 \phi''_{\theta_0}(2 G_0).$$

Furthermore, $\frac{1}{\epsilon_n^2} \left[ \phi(\hat{\theta}_n + \epsilon_n \hat{G}_n) - \phi(\theta_0) - \phi''_{\theta_0}(\hat{\theta}_n - \theta_0 + \epsilon_n \hat{G}_n) \right] = g_n \left( Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) \xrightarrow{w} \frac{1}{2} \phi''_{\theta_0}(G_0)$.

By linearity of $\phi''_{\theta_0}(h)$,

$$R(\hat{\theta}_n, \theta_0, h) = \phi''_{\theta_0} \left( \frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_0 + 2 \epsilon_n h \right) \right) - 2 \phi''_{\theta_0} \left( \frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_0 + \epsilon_n h \right) \right) + \phi''_{\theta_0} \left( \frac{1}{\epsilon_n} \left( \hat{\theta}_n - \theta_0 \right) \right) = 0.$$

Therefore by the above joint convergence and continuous mapping,

$$\hat{\phi''}(Z_n^*) = \frac{1}{\epsilon_n^2} \left[ \phi(\hat{\theta}_n + 2 \epsilon_n Z_n^*) - 2 \phi(\hat{\theta}_n + \epsilon_n Z_n^*) + \phi(\theta_0) \right] = g_n \left( 2 Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) - g_n \left( Z_n^* + \frac{\hat{\theta}_n - \theta_0}{\epsilon_n} \right) + g_n \left( \hat{\theta}_n - \theta_0 \right) + \frac{1}{\epsilon_n} R(\hat{\theta}_n, \theta_0, Z_n^*) \xrightarrow{w} \frac{1}{2} \phi''_{\theta_0}(2 h) - 2 \frac{1}{2} \phi''_{\theta_0}(h) + \frac{1}{2} \phi''_{\theta_0}(0) = \frac{1}{2} 4 \phi''_{\theta_0}(h) - 2 \frac{1}{2} \phi''_{\theta_0}(h) = \phi''_{\theta_0}(h).$$

**Lemma A.2 (Bootstrap extended continuous mapping theorem)** Under the conditions of van der Vaart and Wellner (1996) Theorem 1.11.1 and Kosorok (2007) Theorem 10.8, $g_n \left( \hat{X}_n \right) \xrightarrow{W} g(X)$.  

61
Proof: We use the notations in Kosorok (2007) Theorem 10.8, where it is argued that $\hat{X}_n \xrightarrow{w} X$ implies that unconditionally $\hat{X}_n \rightsquigarrow X$, so that $g_n(\hat{X}_n) \rightsquigarrow g(X)$ by van der Vaart and Wellner (1996) Thm 1.11.1. Let $E_M$ denote the expectation conditional on the data. Write

$$\sup_{h \in BL_1} \left| E_M h \left( g_n(\hat{X}_n) \right) - Eh \left( g(X) \right) \right| \leq \sup_{h \in BL_1} \left| E_M h \left( g_n(\hat{X}_n) \right) - E_M h \left( g(\hat{X}_n) \right) \right|$$

$$+ \sup_{h \in BL_1} \left| E_M h \left( g(\hat{X}_n) \right) - Eh \left( g(X) \right) \right|$$

Since $h \in BL_1$, $E_M h \left( g_n(\hat{X}_n) \right) - E_M h \left( g(\hat{X}_n) \right) \leq E_M \{ g_n(\hat{X}_n) - g(\hat{X}_n) \} \wedge 2$. Next by van der Vaart and Wellner (1996) Theorem 1.11.1, $(g_n(\hat{X}_n), g(\hat{X}_n)) \rightsquigarrow (g(X), g(X))$, $g_n(\hat{X}_n) - g(\hat{X}_n) \rightsquigarrow 0$, so that argue as before equation 10.8 on page 185 of Kosorok,

$$E^* E_M \{ g_n(\hat{X}_n) - g(\hat{X}_n) \} \wedge 2 \to 0 \implies E_M \{ g_n(\hat{X}_n) - g(\hat{X}_n) \} \wedge 2 = o_P(1).$$

By Kosorok (2007) Theorem 10.8, $\sup_{h \in BL_1} \left| E_M h \left( g(\hat{X}_n) \right) - Eh \left( g(X) \right) \right| = o_P(1)$. It then follows

$$\sup_{h \in BL_1} \left| E_M h \left( g_n(\hat{X}_n) \right) - Eh \left( g(X) \right) \right| = o_P(1).$$

Proof for Theorem 6.1 Our first step is to show that $\hat{\theta}_n$ and $\hat{\theta}_n^*$ are respectively $n^\gamma$-consistent and $\epsilon_n^{2\gamma}$-consistent for $\theta_0$. Assumptions (iv) and (viii) imply that the conditions of Lemma 4.1 of Kim and Pollard (1990) are satisfied. Therefore, for each $\eta > 0$, there exist random variables $\{M_n\} = O_p(1)$ such that $|P_n g(\cdot, \theta) - P g(\cdot, \theta)| \leq \eta |\theta - \theta_0|^2 + n^{-2\gamma} M_n^2$. Here, $R_0 > 0$ is the constant such that $P G_R = O(R^{2\rho})$ for all $R \leq R_0$. Assumptions (i) and (ii) imply that the conditions of Corollary 4.2 of Kim and Pollard (1990) are satisfied, which, in combination with Lemma 4.1, imply that $n^\gamma (\hat{\theta}_n - \theta_0) = O_p(1)$.

\footnote{The main revisions to Lemma 4.1 of Kim and Pollard (1990) are redefining $A(n, j) = (j - 1) n^{-\gamma} \leq |\theta| \leq j n^{-\gamma}$, bounding the $j$th summand in $P(M_n > m)$ by $n^{4\gamma} P \sup_{|\theta| < j n^{-\gamma}} |P_n g(\cdot, \theta) - P g(\cdot, \theta)|^2 / \left[ \eta (j - 1)^2 + m^2 \right]$, where the numerator is further bounded by $n^{4\gamma} (n^{-1} C' j n^{-\gamma}(2\rho)) = C' j$.}
Next by Lemma A.3, which generalizes Lemma 4.1 in Kim and Pollard (1990) to a bootstrap version with step size \( \epsilon \), \( \exists M^*_n = O^*_p(1) \) such that \( \epsilon_n|\hat{g}^*_n g(\cdot, \hat{\theta}^*_n)| \leq \eta|\theta - \theta_0|^2 + \epsilon^2_n M^2_n. \) Combine this with Kim and Pollard (1990) Lemma 4.1, and note that since \( O^*_p(e^4_n) + O_P(n^{-2\gamma}) = O^*_p(e^4_n) \), we obtain \( |Z^*_n g(\cdot, \hat{\theta}^*_n) - P g(\cdot, \theta_0)| \leq \eta|\hat{\theta}^*_n - \theta_0|^2 + O^*_p(e^4_n). \) Then choose \( \eta \) so that \( P g(\cdot, \theta) - P g(\cdot, \theta_0) \leq -2\eta|\theta - \theta_0|^2 \), and since \( g(\cdot, \theta_0) = 0, \)

\[
-O^*_p(e^4_n) = Z^*_n g(\cdot, \theta_0) - O^*_p(e^4_n) \leq Z^*_n g(\cdot, \hat{\theta}^*_n) \leq -\eta|\hat{\theta}^*_n - \theta_0|^2 + O^*_p(e^4_n),
\]

from which we conclude that \( |\hat{\theta}^*_n - \theta_0| = O^*_p(e^2_n) \) (and hence also \( O^*_p(e^4_n) \)).

Since \( \hat{\theta}_n \) converges at rate \( n^\gamma \), we can write \( n^\gamma(\hat{\theta}_n - \theta_0) = \arg \max_{h \in \mathbb{R}} n^\gamma \sqrt{n} P g(\cdot, \theta_0 + n^{-\gamma} h) \) converges in finite dimensional distribution to a Gaussian process \( Z_0(h) \) and that \( n^\gamma \sqrt{n} P g(\cdot, \theta_0 + n^{-\gamma} h) \to -\frac{1}{2} h' H h \) for each \( h \) as \( n \to \infty \). We then show that \( W_n \) is stochastically equicontinuous. Finally, we argue using the Argmax Continuous Mapping Theorem (Theorem 2.7 Kim and Pollard (1990)) that \( n^\gamma(\hat{\theta}_n - \theta_0) \sim \arg \max_{h \in \mathbb{R}} Z_0(h) - \frac{1}{2} h' H h. \)

Consider the first part. Assumption (vii) implies the Lindeberg condition is satisfied and that \( n^\gamma \sqrt{n} (P_n - P) g(\cdot; \theta_0 + n^{-\gamma} h) \) converges in finite dimensional distribution to a mean zero Gaussian process with covariance kernel:

\[
\Sigma_{\rho}(s, t) = \lim_{n \to \infty} \text{Cov}(n^\rho \sqrt{n} (P_n - P) g(\cdot; \theta_0 + n^{-\gamma} s), n^\rho \sqrt{n} (P_n - P) g(\cdot; \theta_0 + n^{-\gamma} t))
\]

\[
= \lim_{n \to \infty} n^{2\rho} P g(\cdot; \theta_0 + n^{-\gamma} s) g(\cdot; \theta_0 + n^{-\gamma} t) - n^{2\rho} P g(\cdot; \theta_0 + n^{-\gamma} s) P g(\cdot; \theta_0 + n^{-\gamma} t).
\]

Taking a second order Taylor expansion of \( n^\rho \sqrt{n} P g(\cdot; \theta_0 + n^{-\gamma} h) \) around \( \theta_0 \) and using \( g(\cdot; \theta_0) = 0 \) and \( \frac{\partial}{\partial \theta} P g(\cdot; \theta_0) = 0 \), \( n^\rho \sqrt{n} P g(\cdot; \theta_0 + n^{-\gamma} h) = -n^{\gamma} r + \frac{1}{2} n^{-2\gamma} h' H h + o(||h||^2) \to -\frac{1}{2} h' H h. \) To show that \( W_n \) is stochastically equicontinuous, it suffices to show that for every sequence of positive numbers \( \{\delta_n\} \) converging to zero,

\[
n^\rho \sqrt{n} \sup_{D(n)} |P_n d - P d| = o(1)
\]

where \( D(n) = \{d(\cdot; \theta_0, h_1, h_2) = g(\cdot; \theta_0 + n^{-\gamma} h_1) - g(\cdot; \theta_0 + n^{-\gamma} h_2) \text{ such that } \max(|h_1|, |h_2|) \leq M \text{ and } |h_1 - h_2| \leq \delta_n \}. \) Note that \( D(n) \) has envelope function \( D_n = 2G_{R(n)} \) where
\[ R(n) = Mn^{-\gamma}. \]

Using the Maximal Inequality in section 3.1 of Kim and Pollard (1990), for sufficiently large \( n \), splitting up the expectation according to whether \( n^{2\gamma} P_n D_n^2 \leq \eta \) for each \( \eta > 0 \), and applying the Cauchy-Schwarz inequality,

\[
\begin{align*}
n^{\gamma} \sqrt{\bar{\eta}} \sup_{\varphi(n)} |P_n d - Pd| &\leq \sqrt{\bar{\eta}} \left( \frac{n^{2\gamma} \sup_{\varphi(n)} P_n d^2}{\bar{\eta}} \right) \\
&\leq \sqrt{\bar{\eta}} J(1) + \sqrt{n^{2\gamma} P_n D_n^2} \left[ \sup_{\varphi(n)} \min \left( 1, \frac{1}{\eta} n^{2\gamma} \sup_{\varphi(n)} P_n d^2 \right) \right].
\end{align*}
\]

To show that this is \( o(1) \) for each fixed \( \eta > 0 \), first, note that by assumption (viii), \( \mathbb{P} n^{2\gamma} P_n D_n^2 = 4n^{2\gamma} \mathbb{P} G_{R(n)}^2 = O(n^{2\gamma} R(n)^{2\rho}) = O(1) \) since \( R(n) = Mn^{-\gamma} \). The proof will then be complete if \( n^{2\gamma} \sup_{\varphi(n)} P_n d^2 = o(1) \). Next, for each \( K > 0 \) write

\[
\mathbb{P} \sup_{\varphi(n)} P_n d^2 \leq \mathbb{P} \sup_{\varphi(n)} P_n d^2 \{ D_n > K \} + K \mathbb{P} \sup_{\varphi(n)} P_n d \leq \mathbb{P} P_n D_n^2 \{ D_n > K \} + K \sup_{\varphi(n)} \mathbb{P} d + K \sup_{\varphi(n)} |P_n d| - P |d| \].
\]
By assumption (xi), \( \mathbb{P} P_n D_n^2 \{ D_n > K \} < \eta n^{-2\gamma} \) for large enough \( K \). By assumption (xi) and the definition of \( \varphi(n) \), \( K \sup_{\varphi(n)} |P_n d| - P |d| \ll \eta n^{-2\gamma} \). By assumption (viii) and the maximal inequality, \( K \sup_{\varphi(n)} |P_n d| - P |d| \ll \eta n^{-2\gamma} \).

We have therefore shown that \( n^{\gamma} \sqrt{\eta n} P_n g(\cdot, \theta_0 + n^{-\gamma} h) \Rightarrow Z_0(h) - \frac{1}{2} h' H h \). It follows from the Argmax Continuous Mapping Theorem and \( n^{\gamma} \left( \hat{\theta}_n - \theta_0 \right) = O_P(1) \) that

\[
n^{\gamma} (\hat{\theta}_n - \theta_0) = \arg \max_h n^{\gamma} \sqrt{\eta n} P_n g(\cdot, \theta_0 + n^{-\gamma} h) \Rightarrow \mathbb{H} = \arg \max_h Z_0(h) - \frac{1}{2} h' H h \quad (46)
\]

Similarly, since \( |\hat{\theta}_n - \theta_0| = O_P(\epsilon_n^{2\gamma}) \), we can write

\[
\epsilon_n^{-2\gamma} (\hat{\theta}_n - \theta_0) = \arg \max_h \epsilon_n^{-4\gamma} Z_n^* g(\cdot, \theta_0 + \epsilon_n^{2\gamma} h).
\]
The goal is to show that \( \epsilon_n^{2\gamma} \left( \hat{\theta}_n^* - \hat{\theta}_0 \right) \overset{p}{\sim} \mathbb{H} \) in (46). Note that

\[
\epsilon_n^{-2\gamma} \left( \hat{\theta}_n^* - \hat{\theta}_0 \right) = \epsilon_n^{-2\gamma} \left( \hat{\theta}_n^* - \theta_0 \right) - \epsilon_n^{-2\gamma} \left( \hat{\theta}_n - \theta_0 \right).
\]

By Assumption (ix) and \( n^{\gamma} \) consistency of \( \hat{\theta}_n^* \), \( \epsilon_n^{-2\gamma} \left( \hat{\theta}_n - \theta_0 \right) = \frac{1}{(1+\sqrt{n})^{2\gamma}} n^{\gamma} \left( \hat{\theta}_n - \theta_0 \right) = o_P(1) \). It therefore suffices to show that \( \epsilon_n^{-2\gamma} \left( \hat{\theta}_n^* - \theta_0 \right) \overset{p}{\sim} \mathbb{H} \), since \( \mathbb{H}_n + o_P(1) \overset{p}{\sim} \mathbb{H} \) whenever \( \mathbb{H}_n \overset{p}{\sim} \mathbb{H} \).

To analyze \( \epsilon_n^{-2\gamma} \left( \hat{\theta}_n^* - \theta_0 \right) \), we use Lemma A.4 which extends the Arg Max Theorem (Theorem 3.2.2 in van der Vaart and Wellner (1996)) to a bootstrap version. It therefore suffices to show that (i)

\[
W_n^*(h) \equiv \epsilon_n^{-(1+2\gamma)} (Z_n^* - P) g(\cdot; \theta_0 + \epsilon_n^{2\gamma} h) \overset{p}{\sim} Z_0(h) \quad \text{in} \quad \ell_{\infty}(K) \tag{47}
\]

for any compact \( K \) and that (ii) \( \epsilon_n^{-(1+2\gamma)} P g(\cdot; \theta_0 + \epsilon_n^{2\gamma} h) \to -\frac{1}{2} h' H h \) for each \( h \).

We show (ii) by a second order Taylor expansion around \( \theta_0 \) using \( g(\cdot; \theta_0) = 0 \) and \( \frac{\partial}{\partial \theta} P g(\cdot; \theta_0) = 0 \).

\[
\epsilon_n^{-(1+2\gamma)} P g(\cdot; \theta_0 + \epsilon_n^{2\gamma} h) = -\epsilon_n^{-(1+2\gamma)} \frac{1}{2} \epsilon_n^{4\gamma} h' H h + o(||h||^2) \to -\frac{1}{2} h' H h.
\]

To show (47), we first show its unconditional version \( W_n^*(h) \sim \mathbb{Z}_0(h) \), and then use an almost sure conditional finite dimensional CLT to convert it to \( W_n^*(h) \overset{p}{\sim} \mathbb{Z}_0(h) \) using arguments analogous to Theorem 2.9.6 in van der Vaart and Wellner (1996). Unconditional convergence \( W_n^*(h) \sim \mathbb{Z}_0(h) \) is in turn shown by invoking the Lindeberg finite dimensional CLT and verifying stochastic equicontinuity. Note that

\[
W_n^*(h) = \epsilon_n^{-(1+2\gamma)} (P_n - P) g(\cdot; \theta_0 + \epsilon_n^{2\gamma} h) + \epsilon_n^{-2\gamma} \hat{\theta}_n^* g(\cdot; \theta_0 + \epsilon_n^{2\gamma} h)
\]

\[
= \frac{1}{\sqrt{n} \epsilon_n} \hat{\theta}_n^* \epsilon_n^{-2\gamma} g(\cdot; \theta_0 + \epsilon_n^{2\gamma} h) + \hat{\theta}_n^* \epsilon_n^{-2\gamma} g(\cdot; \theta_0 + \epsilon_n^{2\gamma} h).
\]

For the first part, by assumption (vi), the covariance kernel of \( \hat{\theta}_n g(\cdot; \theta_0 + \epsilon_n^{2\gamma} h) \) converges to the limit

\[
\Sigma_{\rho}(s, t) = \lim_{n \to \infty} \text{Cov} \left( \epsilon_n^{-2\gamma} \hat{\theta}_n g(\cdot; \theta_0 + \epsilon_n^{2\gamma} s), \epsilon_n^{-2\gamma} \hat{\theta}_n g(\cdot; \theta_0 + \epsilon_n^{2\gamma} t) \right)
\]

65
\[\lim_{n \to \infty} n^{-\gamma \rho} \tilde{P}_n g(\cdot; \theta_0 + n^{-\gamma} \sqrt{\theta}) - \lim_{n \to \infty} n^{-\gamma \rho} \tilde{P}_n g(\cdot; \theta_0 + n^{-\gamma} \sqrt{\theta})\]

The Lindeberg condition also holds for \(\hat{G}_n g(\cdot; \theta_0 + n^{-\gamma} \sqrt{\theta})\) by assumptions (vii) and (ix):

\[\lim_{n \to \infty} n^{-\gamma \rho} P \{ n^{-\gamma} \sqrt{\theta} \leq \eta \}
\]

Therefore by (xi), \(\frac{1}{\sqrt{n}} n^{-\gamma \rho} g(\cdot; \theta_0 + n^{-\gamma} \sqrt{\theta}) = o_P(1)\). Alternatively, we can also impose a strong condition, such as \(\sqrt{n} \epsilon_n / \log n \to \infty\), and invoke a version of the law of iterated logarithm, so that we can replace \(o_P(1)\) with \(o_a.s. (1)\). Using similar arguments as those after (45), it can also be shown that for \(\mathcal{D}(n) = \{ d(\cdot, \theta_0, h_1, h_2) = g(\cdot; \theta_0 + n^{-\gamma} h_1) - g(\cdot; \theta_0 + n^{-\gamma} h_2) \} \),

\[\sup_{\mathcal{D}(n)} |\hat{G}_n d| = o(1)\]  

To see (48), as before bound, with \(P \epsilon_n^{-\gamma \rho} P_n D_n^2 = O(1)\), and split according to \(\eta\),

\[P \epsilon_n^{-\gamma \rho} \sup_{\mathcal{D}(n)} |\hat{G}_n d| \leq P \sqrt{\epsilon_n^{-\gamma \rho} P_n D_n^2 J} \left( \epsilon_n^{-\gamma \rho} \sup_{\mathcal{D}(n)} P_n d^2 \right) \]

\[= \sqrt{\eta} J (1) + \sqrt{P \epsilon_n^{-\gamma \rho} P_n D_n^2 J} \left( \epsilon_n^{-\gamma \rho} \sup_{\mathcal{D}(n)} P_n d^2 \right) \]

Finally to show that \(P \epsilon_n^{-\gamma \rho} \sup_{\mathcal{D}(n)} P_n d^2 = o(1)\), split using large \(K\),

\[P \sup_{\mathcal{D}(n)} P_n d^2 \leq P \sup_{\mathcal{D}(n)} P_n D_n^2 \{ D_n > K \} + K \sup_{\mathcal{D}(n)} P_n d^2 \leq K \sup_{\mathcal{D}(n)} P_n d^2 \leq K P \sup_{\mathcal{D}(n)} P_n |d| - P |d| \]

By (x), \(P \sup_{\mathcal{D}(n)} P_n D_n^2 \{ D_n > K \} < \eta n^{-\gamma \rho} \). By (xi), \(K \sup_{\mathcal{D}(n)} P_n d^2 = O(\epsilon_n^{-\gamma \rho} \delta_n^p) = o(\epsilon_1^{-\gamma \rho})\). By (viii) and Kim and Pollard (1990) 3.1, \(K \sup_{\mathcal{D}(n)} |P_n - P|^2 \leq K n^{-1/2} J (1) \sqrt{P \sup_{\mathcal{D}(n)} P_n d^2} = O(n^{-1/2} \epsilon_n^{-\gamma \rho}) = O(\epsilon_n^{-\gamma \rho})\).

Hence we verified (48) to conclude that \(\hat{G}_n \epsilon_n^{-\gamma \rho} g(\cdot; \theta_0 + n^{-\gamma} \sqrt{\theta}) \sim \mathcal{D}(n, h)\), and
\[ \frac{1}{\sqrt{n} \epsilon_n} \hat{G}_n g(\cdot; \theta_0 + \epsilon_n^2 h) = o_P(1), \] both as a process indexed by \( h \) in \( \ell_\infty(K) \) for any compact \( K \).

We remark that while condition (vii) is modeled after (iv) in Lemma 4.5 of Kim and Pollard (1990), neither seems to be needed for the Lindeberg condition. In fact they should all be implied by (vi). Under the integrability condition (vi), for any \( \kappa_n \to \infty \),

\[
\lim_{\alpha \to \infty} \alpha^2 P g\left( \cdot, \theta_0 + \frac{t}{\alpha} \right)^2 \left\{ \alpha^\rho \left| g\left( \cdot, \theta_0 + \frac{t}{\alpha} \right) \right| \geq \kappa_n \epsilon \right\} \to 0.
\]

Then (vii) corresponds to \( \kappa_n = \alpha^2 \rho \) and the Lindeberg condition to \( \alpha = \epsilon_n^{-2\gamma} \) and \( \kappa_n = \sqrt{n} \).

Next, the conditional (given the sample) covariance kernel of \( \epsilon_n^{-2\gamma} \hat{G}_n g(\cdot; \theta_0 + \epsilon_n^2 h) \)

satisfies

\[
\hat{\Sigma}(s, t) = \frac{1}{n} \sum_{i=1}^{n} \epsilon_n^{-4\gamma} g\left( \cdot, \theta_0 + \epsilon_n^{2\gamma} t \right) g\left( \cdot, \theta_0 + \epsilon_n^{2\gamma} s \right) \to \Sigma(s, t), \tag{49}
\]

almost surely by a strong law of large numbers for both the Wild and multinomial bootstrap. The conditional (in \( \xi \)) Lindeberg condition is satisfied if almost surely,

\[
\frac{1}{n} \sum_{i=1}^{n} \epsilon_n^{-4\gamma} g^2\left( z_i, \theta_0 + \epsilon_n^{2\gamma} \right) E\xi^2 \left\{ \left| \epsilon_n^{-2\gamma} \right| g\left( z_i, \theta_0 + \epsilon_n^{2\gamma} \right) \left| \geq \sqrt{n} \epsilon \right\} \right) \leq E\xi^2 \left\{ \left| \max \left| g\left( z_i, \theta_0 + \epsilon_n^{2\gamma} \right) \right| \geq \epsilon \right\} \frac{1}{n} \sum_{i=1}^{n} \epsilon_n^{-4\gamma} g^2\left( z_i, \theta_0 + \epsilon_n^{2\gamma} \right) \to 0. \tag{50}
\]

This holds by the strong LLN and that almost surely, \( \max \left| g\left( z_i, \theta_0 + \epsilon_n^{2\gamma} \right) \right| \to 0 \) using (vi).

Therefore almost surely in finite dimension,

\[ \epsilon_n^{-2\gamma} \hat{G}_n^* g(\cdot; \theta_0 + \epsilon_n^{2\gamma} h) \xrightarrow{\mathcal{L}} \mathcal{N}(0), \quad h = \{h_1, \ldots, h_J\}. \]

With multinomial bootstrap, (50) is replaced by

\[
E^* \epsilon_n^{-4\gamma} g^2\left( z_i^*, \theta_0 + \epsilon_n^{2\gamma} h \right) \left\{ \frac{1}{\sqrt{n}} \epsilon_n^{-2\gamma} \right| g\left( z_i^*, \theta_0 + \epsilon_n^{2\gamma} h \right) \left| \geq \epsilon \right\} \right) \to 0 \]

\[
= \frac{1}{n} \sum_{i=1}^{n} \epsilon_n^{-4\gamma} g^2\left( z_i, \theta_0 + \epsilon_n^{2\gamma} h \right) \left\{ \frac{1}{\sqrt{n}} \epsilon_n^{-2\gamma} \right| g\left( z_i, \theta_0 + \epsilon_n^{2\gamma} h \right) \left| \geq \epsilon \right\} \to 0
\]

67
almost surely by Strong LLN and conditions (vi) and (vii).

Finally we show (unconditional stochastic equicontinuity) of the \( \epsilon_n^{-2\gamma \rho} \hat{G}_n^* g(\cdot; \theta_0 + \epsilon_n^2\gamma h) \) part of \( W_n^* \), separately for the wild and multinomial bootstrap. In the Wild bootstrap case,

\[
\epsilon_n^{-2\gamma \rho} \hat{G}_n^* g(\cdot; \theta_0 + \epsilon_n^2\gamma h) = \epsilon_n^{-2\gamma \rho} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (\delta_i - P) g (\cdot; \theta_0 + \epsilon_n^2\gamma h) + \bar{\xi}_i \hat{G}_n \epsilon_n^{-2\gamma \rho} g(\cdot; \theta_0 + \epsilon_n^2\gamma h).
\]

Since \( \bar{\xi} \overset{a.s.}{\to} 0 \), the second term is \( o_{a.s.} (1) \) in \( \ell_\infty (K) \). The first term is handled in van der Vaart and Wellner (1996) Lemmas 2.3.6 and 2.9.1, and is stochastically equicontinuous whenever \( \hat{G}_n \epsilon_n^{-2\gamma \rho} g(\cdot; \theta_0 + \epsilon_n^2\gamma h) \) is. Combined with the unconditional versions of (49) and (50),

\[
\epsilon_n^{-2\gamma \rho} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (\delta_i - P) g (\cdot; \theta_0 + \epsilon_n^2\gamma h) \overset{\text{a.s}}{\to} Z_0(h) \text{ in } \ell_\infty (K)
\]

Next using the approximation scheme in van der Vaart and Wellner (1996) Theorem 2.9.6, (51), (49) and (50),

\[
\epsilon_n^{-2\gamma \rho} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (\delta_i - P) g (\cdot; \theta_0 + \epsilon_n^2\gamma h) \overset{P}{\Rightarrow} Z_0(h) \text{ in } \ell_\infty (K)
\]

Finally, as the sum between (52) and \( o_P (1) \) terms, \( W_n^* (h) \overset{P}{\Rightarrow} Z_0 (h) \) in \( \ell_\infty (K) \).

In the multinomial bootstrap case, for \( \mathcal{D}(n) \) in (48), we use a bootstrap version of Kim and Pollard (1990) to show \( \epsilon_n^{-2\gamma \rho} P_n \sup_{\mathcal{D}(n)} \hat{G}_n^* d = o(1) \). As before bound, with \( P_n \epsilon_n^{-4\gamma \rho} P_n^* D_n^2 = O (1) \),

\[
\epsilon_n^{-2\gamma \rho} P_n \sup_{\mathcal{D}(n)} \hat{G}_n^* d \leq P_n \epsilon_n^{-4\gamma \rho} P_n^* D_n^2 \left( \frac{\epsilon_n^{-4\gamma \rho} \sup_{\mathcal{D}(n)} P_n^* d^2}{\epsilon_n^{-4\gamma \rho} P_n^* D_n^2} \right) \leq \sqrt{n} J (1) + \sqrt{P_n \epsilon_n^{-4\gamma \rho} P_n^* D_n^2} \left( \min (1, \frac{1}{\eta} \epsilon_n^{-4\gamma \rho} \sup_{\mathcal{D}(n)} P_n^* d^2) \right)
\]

68
Finally to show that \( P_n^{-\gamma\rho} \sup_{n} P_n^* d^2 = o(1) \), split using large \( K \),

\[
\mathbb{P} \sup_{n} P_n^* d^2 \leq \mathbb{P} P_n^* D_n^2 \{ D_n > K \} + K \sup_{n} P d
\]

\[
+ K \mathbb{P} \sup_{n} |P_n d| - P d\| + K \mathbb{P} \sup_{n} |P_n^* d| - P_n |d|\|.
\]

The first three terms are handled after (48). The last term is handled by a bootstrap version of Kim and Pollard (1990) maximal inequality 3.1, with \( P_n D_n^2 = P D_n^2 + O(n^{-1/2}) \),

\[
\mathbb{P} P_n \sup_{n} |P_n^* d| - P_n |d|\| \leq J(1) \mathbb{P} n^{-1/2} \sqrt{P_n D_n^2} = O(n^{-1/2} (\epsilon^{2\gamma\rho} + n^{-1/4})) .
\]

This is \( o(\epsilon^{2\gamma\rho}) \) when \( \gamma = 1/3, \rho = 1/2 \). For larger values of \( \rho > 2/3 \) and \( \gamma \) we will impose the additional condition that \( n^{3/4} \epsilon^{2\gamma\rho} \rightarrow \infty \) to achieve \( o(\epsilon^{2\gamma\rho}) \). Alternatively, we can impose Holder continuity to argue directly that \( \mathbb{P} P_n \sup_{n} |P_n^* d| - P_n |d|\| = O(\delta_n) = o(1) \).

It is also clear from the arguments in the proof that unconditionally \( \epsilon_n^{-2\gamma} (\hat{\theta}_n^* - \hat{\theta}_n) \rightarrow \mathcal{J} \). This is because also unconditionally, \( \epsilon_n^{-2\gamma} (\hat{\theta}_n^* - \hat{\theta}_n) = O_P(1) \),

\[
\epsilon_n^{1+2\gamma\rho_0} Z_n^* g(\cdot, \theta_0 + \epsilon_n^{2\gamma} h) \rightarrow Z_0(h) - \frac{1}{2} h' H h \quad \text{in} \quad \ell_\infty(K),
\]

so that the unconditional arg max theorem van der Vaart and Wellner (1996) 3.2.2 can be applied.

\textbf{Lemma A.3} Under conditions (viii) and (ix) of Theorem 6.1, for each \( \eta \) there exists random variables \( M_n^* = O_P(1) \) such that for all \( \theta \) close to \( \theta_0 \),

\[
\epsilon_n |\hat{G}_n^* g(\cdot, \theta)| \leq \eta |\theta - \theta_0|^2 + \epsilon_n^{4\gamma} M_n^*.
\]

\textbf{Proof:} We first consider the Wild Bootstrap where \( \hat{G}_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\xi_i - \bar{\xi}) \delta_i \). Also WLOG let \( \theta_0 = 0 \). Define \( A(n, j) = \{ \theta : (j - 1) \epsilon_n^{2\gamma} \leq |\theta| \leq j \epsilon_n^{2\gamma} \} \). Then

\[
P_\xi(M_n^* > m) \leq P_\xi \left( \exists \theta : \epsilon_n |\hat{G}_n^* g(\cdot, \theta)| > \eta |\theta|^2 + \epsilon_n^{4\gamma} m^2 \right)
\]

69
\[
\leq \sum_{j=1}^{\infty} P_n \{ \exists \theta \in A(n, j) : \epsilon_n^{-8\gamma} \epsilon_n |\hat{G}_n^* g (\cdot, \theta) | > \eta (j - 1)^2 + m^2 \}.
\]

The \( j \)th summand is then bounded by,
\[
\epsilon_n^{-8\gamma} \epsilon_n^2 P_{\xi} \sup_{|\theta| < j \epsilon_n^{2\gamma}} |\hat{G}_n^* g (\cdot, \theta) |^2 / [(j - 1)^2 + m^2]^2
\]

Note that \( \hat{G}_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (\delta_i - P) - \bar{\xi} \hat{G}_n \). We bound the expectation in the numerator in (53) by
\[
P_{\xi} \sup_{|\theta| < j \epsilon_n^{2\gamma}} |\hat{G}_n^* g (\cdot, \theta) |^2 \leq P_{\xi} \sup_{|\theta| < j \epsilon_n^{2\gamma}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (\delta_i - P) (\cdot, \theta) \right|^2 + \epsilon^2 \sup_{|\theta| < j \epsilon_n^{2\gamma}} |\hat{G}_n g (\cdot, \theta) |^2.
\]

By Maximal Inequality 3.1 in Kim and Pollard (1990) and also \( \bar{\xi} = o_P (1) \), \( \bar{\xi} \sup_{|\theta| < j \epsilon_n^{2\gamma}} |\hat{G}_n g (\cdot, \theta) |^2 = o_P \left((\epsilon_n^{2\gamma})^{2\rho}\right) \). Next by Lemmas 2.3.6 and (a square version of) 2.9.1 of van der Vaart and Wellner (1996), for large \( n_0 \) and \( n \),
\[
P \sup_{|\theta| < j \epsilon_n^{2\gamma}} \left| \sum_{i=1}^{n} \xi_i (\delta_i - P) (\cdot, \theta) \right|^2 \leq O \left( P \sup_{|\theta| < j \epsilon_n^{2\gamma}} |g (\cdot, \theta) |^2 + \max_{n_0 \leq k \geq n} P \sup_{|\theta| < j \epsilon_n^{2\gamma}} |\hat{G}_k g (\cdot, \theta) |^2 \right).
\]

Both terms are \( O \left((\epsilon_n^{2\gamma})^{2\rho}\right) \) by (viii) and again Kim and Pollard (1990) Maximal Inequality 3.1. Therefore
\[
P_{\xi} \sup_{|\theta| < j \epsilon_n^{2\gamma}} |\hat{G}_n^* g (\cdot, \theta) |^2 = O_p \left((\epsilon_n^{2\gamma})^{2\rho}\right) .
\]

The numerator of (53) is thus further bounded by, for \( C_n' = O_p (1) \), \( \epsilon_n^{-8\gamma} \epsilon_n^2 C_n' (\epsilon_n^{2\gamma})^{2\rho} = C_n' \). Then \( M_n^* = O_p (1) \) since by choosing \( m \), the following can be made asymptotically small. \( \forall \epsilon > 0, \)
\[
P (P_n (M_n^* > m) > \epsilon) \leq P \left( C_n' \sum_{j=1}^{\infty} 1 / [(j - 1)^2 + m^2]^2 > \epsilon \right).
\]
Next consider the multinomial bootstrap $\hat{G}_n^* = \sqrt{n} (P_n^* - P_n)$. Note a bootstrap version of part (ii) of Maximal Inequality 3.1 in Kim and Pollard (1990) holds with:

$$nP_n \sup_{f} |P_n^* f - P_n f|^2 \leq J(1)^2 P_n F^2 = J(1)^2 P F^2 + O_P \left( \frac{1}{\sqrt{n}} \right). \quad (54)$$

This can be used to similarly bound the numerator in (53) by $\epsilon_n^{-\gamma} \epsilon_n^2 C_n^\gamma \implies C_n^\gamma$ for $C_n^\gamma = O_P(1)$. Strictly speaking, (54) is $O_P \left( (\epsilon_n^2)^{2\rho} \right)$ only when $n^{-1/2} = O \left( (\epsilon_n^2)^{2\rho} \right)$, which holds when $\rho = 1/2, \gamma = 1/3$ but not when $\rho = 1, \gamma = 1/2$. However, under the additional assumption that $g(\cdot, \theta)$ is Holder with index $\rho$, it can be directly verified that $P_n F^2 = O_P \left( (\epsilon_n^2)^{2\rho} \right)$.

**Lemma A.4** Let $M_n \xrightarrow{p} M$ in $\ell_\infty(K)$ for every compact $K$. If there exists a tight $\hat{h}$ such that for every open $G$ containing $\hat{h}$, $M \left( \hat{h} \right) > \sup_{\hat{g} \in G, \hat{h} \in K} M(\hat{h})$, and $M_n \left( \hat{h}_n \right) \geq \sup_{\hat{h}} M_n(\hat{h}) - \sigma_{\hat{F}}(1)$, where $\hat{h}_n = O^*_p(1)$, then $\hat{h}_n \xrightarrow{p} \hat{h}$.

**Proof:** First a bootstrap version of the Portmanteau theorem can be shown. The following are equivalent (TFAE): (i) $X_n \xrightarrow{p} X$; (ii) For every open $G$ and $\forall \epsilon > 0$, $P \left( P_n (X_n \in G) \geq P (X \in G) - \epsilon \right) \to 1$; (iii) For every closed $F$ and $\forall \epsilon > 0$, $P \left( P_n (X_n \in F) \leq P (X \in F) + \epsilon \right) \to 1$. By the bootstrap CMT (Theorem 10.8 Kosorok (2007)), $\sup_{\hat{h} \in F \cap K} M_n(\hat{h}) - \sup_{\hat{h} \in K} M_n(\hat{h}) \xrightarrow{p} \sup_{\hat{h} \in F \cap K} M(\hat{h}) - \sup_{\hat{h} \in K} M(\hat{h})$. Then by bootstrap Portmanteau, with probability converging to 1 (w.p.c.1), $\forall \epsilon > 0$,

$$P_n \left( \hat{h}_n \in F \cap K \right) \leq P_n \left( \sup_{\hat{h} \in F \cap K} M_n(\hat{h}) \geq \sup_{\hat{h} \in K} M_n(\hat{h}) - \sigma^*_p(1) \right)
\leq P \left( \sup_{\hat{h} \in F \cap K} M(\hat{h}) \geq \sup_{\hat{h} \in K} M(\hat{h}) \right) + \epsilon \leq P \left( \hat{h} \in F \right) + P \left( \hat{h} \notin K \right) + \epsilon.$$

Next split up $P_n \left( \hat{h}_n \in F \right) \leq P_n \left( \hat{h}_n \in F \cap K \right) + P_n \left( \hat{h} \notin K \right)$. Choose $K$ large so that $P_n \left( \hat{h} \notin K \right) \leq \epsilon$ has probability larger than $1 - \delta$ for large $n$. Conclude that with probability larger than $1 - 2\delta$ for large $n$, $P_n \left( \hat{h}_n \in F \right) \leq P \left( \hat{h} \in F \right) + 2\epsilon$. ■

**Proof for Theorem 6.2** Part 1: Define $w_n(h) = +\infty 1(h \notin n^\gamma (C - \theta_0))$ and $w(h) = +\infty 1(h \notin T_C(\theta_0))$. By (23), $w_n(\cdot) \xrightarrow{e} w(\cdot)$, as a sequence of nonrandom
functions. Next define
\[
\mathcal{H}_n(h) = n^{-\gamma} \sqrt{n} P_n g (\cdot, \theta_0 + n^{-\gamma} h) + w_n(h).
\]
Similarly define \( \mathcal{H}(h) = Z_0(h) + \frac{1}{2} h' H h + w(h) \). Then for \( \hat{h} = n^{-\gamma} (\hat{\theta} - \theta_0) \), \( \mathcal{H}_n(\hat{h}) = \inf h \mathcal{H}_n(h) + o_P(1) \). Almost the same arguments as in the proof of Theorem 6.1 can be applied to show that \( \hat{h} = O_P(1) \). As in the proof of Theorem 4.3 of Geyer (1994), we confine our attention to compact sets.

In Theorem 6.1 it has been shown that \( n^{-\gamma} \sqrt{n} P_n g (\cdot, \theta_0 + n^{-\gamma} h) \rightsquigarrow Z_0(h) + \frac{1}{2} h' H h \) in \( \ell_\infty(K) \) in the sense of finite dimensional convergence and stochastic equicontinuity. Furthermore, \( Z_0(h) + \frac{1}{2} h' H h \) has a continuous sample path. Then according to page 5 in Knight (1999), \( n^{-\gamma} \sqrt{n} P_n g (\cdot, \theta_0 + n^{-\gamma} h) \rightarrow_{u-d} Z_0(h) + \frac{1}{2} h' H h \). Next by Theorem 4 in Knight (1999), \( \mathcal{H}_n(\cdot) \rightarrow_{e-d} \mathcal{H}(\cdot) \). The remaining arguments are the same as in the second part of the proof of Theorem 4.4 in Geyer (1994). In other words, the arg min functional is continuous with respect to the metric of epi-convergence on the space of functions embedding \( \mathcal{H}_n(\cdot) \) and \( \mathcal{H}(\cdot) \), which allows for the application of a continuous mapping theorem as in Theorem 1 of Knight (1999).

Part 2: As before, for \( \hat{h}^* = \epsilon_n^{-2\gamma} (\hat{\theta}^* - \theta_0) \), it suffices to show that \( \hat{h}^* \xrightarrow{\mathcal{W}} \mathcal{J} \) and \( \hat{h}^* \rightsquigarrow \mathcal{J} \). Define \( w_n^*(h) = +\infty 1 (h \notin \epsilon_n^{-2\gamma} (C - \theta_0)) \), so that by (23), \( w_n^*(\cdot) \xrightarrow{\mathcal{W}} w(\cdot) \) as a nonrandom sequence. Next let \( \mathcal{H}_n^*(h) = \epsilon_n^{-4\gamma} Z_n^* g (\cdot, \theta_0 + \epsilon_n^\gamma h) + w_n(h) \). Then for \( \hat{h}^* = \epsilon_n^{-2\gamma} (\hat{\theta}^* - \theta_0) \), by assumption, \( \mathcal{H}_n^*(\hat{h}^*) = \inf h \mathcal{H}_n^*(h) + o_P(1) \). Almost the same arguments as in the proof of Theorem 6.1 can be applied to show that \( \hat{h}_n^* \) is \( O_P(1) \) and \( O_P(1) \), so we can confine attention to compact sets. Theorem 6.1 has shown that \( \epsilon_n^{-4\gamma} Z_n^* g (\cdot, \theta_0 + \epsilon_n^\gamma h) \xrightarrow{\mathcal{W}} (\rightsquigarrow) Z_0(h) \) in \( \ell_\infty(K) \). A bootstrap in probability version of Theorem 4 Knight (1999) can be stated to show that \( \mathcal{H}_n^*(\cdot) \rightarrow_{e-d} \mathcal{H}(\cdot) \) conditionally in probability, which can be equivalently stated as \( \rho_{BL_1}(\mathcal{H}_n^*(\cdot), \mathcal{H}(\cdot)) = o_p(1) \) where \( BL_1 \) now represents the class of Lipschitz norm 1 functions with respect to the metric of epi-convergence (see last equation on page 4 of Knight (1999)), and \( \mathcal{H}_n^*(\cdot) \) is understood to be the conditional law given the data. Finally, by revising the bootstrap argmax continuous mapping Lemma A.4 to replace weak convergence by epi-convergence after incorporating Theorem 1 in Knight (1999), we can show that \( \hat{h}_n^* \xrightarrow{\mathcal{W}} \hat{h} \).
Proof for Theorem 6.3  The same arguments as in the proof of Theorem 6.2 (see also Sherman (1993) and Newey and McFadden (1994)) show that in \( \ell_\infty (K) \),

\[
n \left( \hat{Q}_n (\theta_0 + h/\sqrt{n}) - \hat{Q}_n (\theta_0) \right) \\
\sim \Delta_0 h + \frac{1}{2} h' H h = (h + H^{-1} \Delta_0)' H (h + H^{-1} \Delta_0) - \frac{1}{2} \Delta_0 H^{-1} \Delta_0.
\]

Then argue as in the previous proof that \( \hat{h} = \sqrt{n} (\hat{\theta}_n - \theta_0) = O_P (1) \) and that

\[
n \left( \hat{Q}_n (\theta_0 + h/\sqrt{n}) - \hat{Q}_n (\theta_0) \right) + w_n (h) \\
\rightarrow_{e-d} (h + H^{-1} \Delta_0)' H (h + H^{-1} \Delta_0) - \frac{1}{2} \Delta_0 H^{-1} \Delta_0 + w (h).
\]

for \( w_n (h) = \infty 1 (h \notin \sqrt{n} (C - \theta_0)) \) and \( w (h) = \infty 1 (h \notin T_C (\theta_0)) \). Therefore \( \hat{J}_n \sim J \). Next note that \( \hat{J}_n^* = \hat{J}_n^* - c_n^{-1} (\hat{\theta}_n - \theta_0) \) where

\[
\hat{J}_n^* = \arg \min_{\hat{h} \in \epsilon_n^{-1} (C - \theta_0)} \hat{Q}_n^* (\hat{h}) = \left( \hat{h} + \hat{H}^{-1} \hat{\Delta}_n - \frac{\hat{\theta} - \theta_0}{\epsilon_n} \right)' \hat{H} \left( \hat{h} + \hat{H}^{-1} \hat{\Delta}_n - \frac{\hat{\theta} - \theta_0}{\epsilon_n} \right).
\]

Note that \( \frac{\hat{\theta} - \theta_0}{\epsilon_n} = o_P (1) \), \( \hat{H}^{-1} \hat{\Delta}_n \sim H^{-1} \Delta_0 \), therefore \( \hat{H}^{-1} \hat{\Delta}_n - \frac{\hat{\theta} - \theta_0}{\epsilon_n} \sim H^{-1} \Delta_0 \). Therefore by the bootstrap CMT (Proposition 10.7 in Kosorok (2007)), in \( \ell_\infty (K) \),

\[
\hat{Q}_n^* (\hat{h}) \xrightarrow{P} \hat{Q}_\infty (\hat{h}) = (\hat{h} + H^{-1} \Delta_0)' H (\hat{h} + H^{-1} \Delta_0)
\]

Therefore the same arguments as in part 2 of the proof of Theorem 6.2 apply. \( \hat{Q}_n^* (\hat{h}) + w_n (\hat{h}) \rightarrow_{e-d} \hat{Q}_\infty (\hat{h}) + w (\hat{h}) \) conditionally in probability, \( \hat{J}_n^* \xrightarrow{P} J \), and \( \hat{J}_n^* \xrightarrow{P} J \).

Proof for Theorem 7.1  Given consistency, first we show that \( n^{1/3} (\hat{\theta}_n - \theta_0) = O_P (1) \). Define \( \hat{G}_n (\theta) = \sqrt{n} (P_n - P) g (z_i, \theta) \). Then \( \hat{\pi} (\theta) = g (\theta) + \hat{\pi} (\theta_0) + \hat{\eta}_n (\theta) \), where \( \hat{\eta}_n (\theta) = \frac{1}{\sqrt{n}} \hat{G}_n (\theta) \). Recall that \( \hat{Q}_n (\theta) = \hat{\pi} (\theta)' W \hat{\pi} (\theta) \). Let \( \hat{g} (\theta) = P_n g (\cdot, \theta) \) and \( g (\theta) = Pg (\cdot, \theta) \). Write \( \hat{Q}_n (\theta) - \hat{Q}_n (\theta_0) = Q_1 (\theta) + \hat{Q}_2 (\theta) + \hat{Q}_3 (\theta) \), where

\[
Q_1 (\theta) = g (\theta)' W g (\theta) + g (\theta)' W \hat{\pi} (\theta_0) \; \; \; \; \; \hat{Q}_2 (\theta) = \hat{\pi} (\theta_0)' W \hat{\eta}_n (\theta) \; \; \; \; \; \hat{Q}_3 (\theta) = \hat{\pi} (\theta)' W \hat{\pi} (\theta_0) - \pi (\theta_0)) + g (\theta)' W \hat{\eta}_n (\theta) + \hat{g} (\theta_0)' W \hat{\eta}_n (\theta).
\]

73
Then we can write \(Q_1(\theta) = \frac{1}{2} (\theta - \theta_0) \cdot (\bar{H} + o(1)) (\theta - \theta_0)\). Next apply Kim and Pollard (1990) Lemma 4.1 to \(\hat{\theta}_n(\theta)\), and in turn \(\hat{Q}_3(\theta)\): \(\forall \epsilon > 0, \exists M_n = O_P(1)\),

\[
|\hat{Q}_3(\theta)| \leq \epsilon |\theta - \theta_0| + n^{-2/3}M_n^2.
\]

The 1st, 3rd and 4th terms in \(\hat{Q}_2(\theta)\) are all of the form \(o_P(1) \hat{\theta}_n(\theta)\), hence are also bounded by \(\epsilon |\theta - \theta_0| + n^{-2/3}M_n^2\). The 2nd term in \(\hat{Q}_2(\theta)\) can also be bounded by, for \(n\) large enough,

\[
|g(\theta)' W(\hat{\pi}(\theta_0) - \pi(\theta_0))| = O_P\left(\frac{|\theta - \theta_0|}{\sqrt{n}}\right) \leq \epsilon |\theta - \theta_0|^2 + \frac{1}{\epsilon^2 n} \leq \epsilon |\theta - \theta_0|^2 + n^{-2/3}M_n^2.
\]

Therefore \(|\hat{Q}_2(\theta) + \hat{Q}_3(\theta)| \leq \epsilon |\theta - \theta_0| + n^{-2/3}M_n^2\). The same arguments as in Kim and Pollard (1990) corollary 4.2 can then be applied to show that \(|\hat{\theta}_n - \theta_0| = O_P(n^{-1/3})\).

Next note \(\hat{\theta} = n^{1/3} (\hat{\theta}_n - \theta_0)\) where \(\hat{\theta} = \arg \min_h n^{1/6} \hat{Q}(\theta_0 + n^{-1/3}h)\). It will follow from van der Vaart and Wellner (1996) Theorem 3.2.2 that \(\hat{\theta} \rightsquigarrow \arg \min_h \bar{Z}_0(h) + \frac{1}{2} h' \bar{H} h\) if we can show that

\[
n^{2/3} \left(\hat{Q}(\theta_0 + n^{-1/3}h) - \hat{Q}_n(\theta_0)\right) \rightsquigarrow \bar{Z}_0(h) + \frac{1}{2} h' \bar{H} h
\]

over compact sets. Since \(Q_1(\theta_0 + n^{-1/3}h) = \frac{1}{2} h' \bar{H} h + o(1)\), it remains to show that \(\left(\hat{Q}_2 + \hat{Q}_3\right)(\theta_0 + n^{-1/3}h) \rightsquigarrow \bar{Z}_0(h)\).

From the proof of Theorem 6.1, \(n^{2/3} \hat{\theta}_n(\theta_0 + n^{-1/3}h) = n^{1/6} \hat{\theta}_n(\theta_0 + n^{-1/3}h) \rightsquigarrow \bar{Z}_0(h)\). Therefore \(n^{2/3} \hat{Q}_3(\theta_0 + n^{-1/3}h) \rightsquigarrow \bar{Z}_0(h)\). Since the 1st, 3rd and 4th terms in \(n^{2/3} \hat{Q}_2(\theta_0 + n^{-1/3}h)\) are all of the form \(o_P(1) n^{2/3} \hat{\theta}_n(\theta_0 + n^{-1/3}h)\), they all converge weakly to 0. For the 2nd term there,

\[
n^{2/3} |g(\theta_0 + n^{-1/3}h)' W(\hat{\pi}(\theta_0) - \pi(\theta_0))| = n^{2/3} O_P\left(\frac{|n^{-1/3}h|}{\sqrt{n}}\right) \leq O_P(h n^{-1/6}) = o_P(1).
\]

Therefore \(n^{2/3} \hat{Q}_2(\theta_0 + n^{-1/3}h) = o_P(1)\). By Slutsky (van der Vaart and Wellner (1996) Theorem 1.4.7),

\[
n^{2/3} \left(\hat{Q}_2 + \hat{Q}_3\right)(\theta_0 + n^{-1/3}h) \rightsquigarrow \bar{Z}_0(h).
\]

The rest of the arguments are the same as in Theorem 6.1.
Remark: Assumption 7.1 (5) defines misspecification (under overidentification). In a correctly specified model, \( \pi(\theta_0) = 0 \) so that \( \hat{Q}_3(\theta) \equiv 0 \). In this case \( \sqrt{n} \) convergence is driven by the second term in \( \hat{Q}_2(\theta) \):

\[
\hat{Q}_2(\theta) = (\theta - \theta_0)' G' W (\hat{\pi}(\theta_0) - \pi(\theta_0)) + o_P \left( \frac{1}{n} \right) + o_P \left( \frac{1}{\sqrt{n}} \right) |\theta - \theta_0|,
\]

so that \( \sqrt{n} \hat{Q}_2 \left( \theta_0 + \frac{h}{\sqrt{n}} \right) \sim h' G' W N (0, \Omega) \). Furthermore, \( \bar{H} \) simplifies to \( (G' W G) \) (Hall and Inoue (2003)). In misspecified models where the data and parameters are linearly separable, \( \hat{\eta}_n(\theta) \equiv 0 \). Then misspecification is only manifested in the difference between \( \bar{H} \) and \( G' W G \).

\( \hat{Q}_3(\theta) \) becomes important in nonlinear misspecified models. In nonsmooth (simulated) models \( \hat{Q}_3(\theta) \) dominates \( \hat{Q}_2(\theta) \) and drives the convergence rate and asymptotic distribution. In smooth models in Hall and Inoue (2003), both \( \hat{Q}_3(\theta) \) and (the 2nd term in) \( \hat{Q}_2(\theta) \) contribute the same order asymptotically. Suppose \( \hat{\pi}(\cdot, \theta) \) is four times continuously differentiable with bounded analytic derivatives. Then

\[
\hat{Q}_3(\theta) = \pi(\theta_0)' W \left( \hat{G}(\theta_0) - G \right) (\theta - \theta_0) + O_P \left( \frac{1}{\sqrt{n}} \right) |\theta - \theta_0|^2 + O \left( |\theta - \theta_0|^3 \right).
\]

This implies \( |\hat{\theta}_n - \theta_0| = O_P \left( \frac{1}{\sqrt{n}} \right) \) and subsequently,

\[
\sqrt{n} \left( \hat{Q}_2 + \hat{Q}_3 \right) \left( \theta_0 + \frac{h}{\sqrt{n}} \right) \sim h' G' W \sqrt{n} (\hat{\pi}(\theta_0) - \pi(\theta_0)) + o_P (1).
\]

In nonsmooth misspecified models where \( \hat{G}(\theta) \) is not analytically available, \( n^{1/3} \) asymptotics apply to accurate nongradient based optimizers. Gradient based optimization methods based on numerically differentiating \( \hat{\pi}(\theta) \) will generate the non-parametric rate and asymptotic distribution in Hong et al. (2015).

**Proof for Theorem 7.2** The logic is similar to that of Theorem 6.1 so we highlight the key steps, and continue with the notations in the proof of Theorem 7.1. As before, since \( \epsilon_n^{-2/3} \left( \hat{\theta}_n - \theta_0 \right) = o_P (1) \), it suffices to show that \( \epsilon_n^{-2/3} \left( \hat{\theta}_n^* - \theta_0 \right) \xrightarrow{p} \mathcal{J} \). Define

\[
\hat{Q}_n^*(\theta) = Z_n^* \pi(\cdot, \theta)' W Z_n^* \pi(\cdot, \theta). \quad \text{Define} \quad \hat{\eta}_n^*(\theta) = (Z_n^* - P) g(\cdot, \theta).
\]

Then \( Z_n^* \pi(\cdot, \theta) = \)
\[ \hat{\eta}_n^*(\theta) + g(\theta) + Z_n^* \pi(\theta_0). \]

Decompose \( Q_n^*(\theta) - \hat{Q}_n^*(\theta_0) = Q_1(\theta) + \hat{Q}_2(\theta) + \hat{Q}_3(\theta), \)

where
\[
Q_1(\theta) = g(\theta)' W g(\theta) + g(\theta)' W \pi(\theta_0), \quad \hat{Q}_2(\theta) = \pi(\theta_0)' W \hat{\eta}_n^*(\theta) \]
\[
\hat{Q}_3(\theta) = \hat{\eta}_n^*(\theta)' W \hat{\eta}_n^*(\theta) + g(\theta)' W (Z_n^* - P) \pi(\cdot, \theta_0) \]
\[
+ g(\theta)' W \hat{\eta}_n^*(\theta) + (Z_n^* - P) \pi(\cdot, \theta_0)' W \hat{\eta}_n^*(\theta). \]

As before \( Q_1(\theta) = \frac{1}{2} (\theta - \theta_0)' (\bar{H} + o(1)) (\theta - \theta_0). \) Next by Lemma A.3, for all \( \epsilon > 0, \)
\( \exists \epsilon_0 = O_P(1) \) such that \( \epsilon_0 |\hat{\eta}_n^* (\cdot, \theta)| \leq \epsilon |\theta - \theta_0|^2 + \epsilon_0^2 M_n^2. \) Then as \( O_P(\epsilon_n^{4/3}) + O_P(n^{-2/3}) = O_P(\epsilon_n^{4/3}), |\hat{\eta}_n^*(\theta)| = |Z_n^* g(\cdot, \theta) - P g(\cdot, \theta)| \leq \epsilon |\theta - \theta_0|^2 + O_P(\epsilon_n^{4/3}). \) So that \( |\hat{Q}_3(\theta)| \leq \epsilon |\theta - \theta_0|^2 + O_P(\epsilon_n^{4/3}). \) The 1st, 3rd and 4th terms in \( \hat{Q}_2(\theta) \) are either \( o_P(1) \hat{\eta}_n^*(\theta) \) or \( o(1) \hat{\eta}_n^*(\theta). \) Hence there is also \( |\hat{Q}_3(\theta) + \hat{Q}_2(\theta)| \leq \epsilon |\theta - \theta_0|^2 + O_P(\epsilon_n^{4/3}). \) The same arguments as in the proof of Theorem 6.1 can be used to conclude that \( |\hat{\theta}_n^* - \theta_0| = O_P(\epsilon_n^{4/3}). \)

Given the above derived rate of convergence, to obtain the limiting distribution, define \( \hat{h}^* = \epsilon_n^{-2/3} (\hat{\theta}_n^* - \theta_0). \) Then \( \hat{h}^* = \arg \min_h \hat{L}_n^*(h), \) where we decompose
\[
\hat{L}_n^*(h) = \epsilon_n^{-4/3} \left(\hat{Q}_n^*(\theta_0 + \epsilon_n^{-2/3} h) - \hat{Q}_n^*(\theta_0)\right)
\]
\[
= \epsilon_n^{-4/3} Q_1(\theta_0 + \epsilon_n^{-2/3} h) + \epsilon_n^{-4/3} \hat{Q}_2(\theta_0 + \epsilon_n^{-2/3} h) + \epsilon_n^{-4/3} \hat{Q}_3(\theta_0 + \epsilon_n^{-2/3} h) \]

Consider each of these three terms separately. First, \( \epsilon_n^{-4/3} Q_1(\theta_0 + \epsilon_n^{-2/3} h) \to^{P} \frac{1}{2} h' \bar{H} h. \) Next, by the same arguments as in Theorem 6.1, with \( \rho = 1/2 \) and \( \gamma = 1/3, \)
\( \epsilon_n^{-4/3} \hat{\eta}_n^*(\theta_0 + \epsilon_n^{-2/3} h) \sim \mathcal{Z}(h). \) Then it also holds that \( \epsilon_n^{-4/3} \hat{Q}_3(\theta_0 + \epsilon_n^{-2/3} h) \sim \mathcal{Z}(h). \)

Finally consider \( \epsilon_n^{-4/3} \hat{Q}_2(\theta_0 + \epsilon_n^{-2/3} h) \). The 1st, 3rd and 4th terms in \( \hat{Q}_2(\theta) \) are either \( o_P(1) \epsilon_n^{-4/3} \hat{\eta}_n^*(\theta_0 + \epsilon_n^{-2/3} h) \) or \( o(1) \epsilon_n^{-4/3} \hat{\eta}_n^*(\theta_0 + \epsilon_n^{-2/3} h), \) and are therefore all \( o_P(1). \) The 2nd term is also \( o_P(1) \) since
\[
\epsilon_n^{-4/3} g(\theta_0 + \epsilon_n^{-2/3} h)' W (P_n - P + \epsilon_n g_n^*) \pi(\cdot, \theta_0) \]
\[
= \epsilon_n^{-2/3} (G + o(1)) h \left( O_P \left( \frac{1}{\sqrt{n}} \right) + O_P(\epsilon_n) \right) = O_P(\epsilon_n^{1/3}). \]

Then we use Slutsky (van der Vaart and Wellner (1996) Theorem 1.4.7) to conclude
that

$$
\varepsilon_n^{-4/3} \left( \hat{Q}_2^* (\theta_0 + \varepsilon_n^{2/3} h) + \hat{Q}_3^* (\theta_0 + \varepsilon_n^{2/3} h) \right) \overset{p}{\rightarrow} \hat{Z}_0 (h).
$$

Hence \( \hat{c}_n (h) \overset{p}{\sim} \hat{Z}_0 (h) + \frac{1}{3} h' \hat{H} h \) over compact sets. The bootstrap arg max theorem can then be applied to conclude that \( h^* \overset{P}{\rightarrow} \arg \min_n \hat{Z}_0 (h) + \frac{1}{3} h' \hat{H} h \).

**Proof for Theorem 8.1** When \( \lambda_0 < \infty \), the conditions in Theorem 6.1 hold for classes of functions changing with \( n \) and can be verified here. The Arg max theorem can also possibly be extended to allow for sample dependent \( \lambda_n \to \infty \). Instead we take a more direct approach following Knight and Fu (2000). When \( \lambda_0 < \infty \), Knight and Fu (2000) already proved that \( \rho_{BL1} = (\hat{G}_n, \mathbb{G}^0_n) = o_P (1) \). So consider \( \lambda_n \to \infty \). First note, as in Knight and Fu (2000), that for \( C_n = \frac{1}{n} X'X \) and \( W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \epsilon_i, \sqrt{n} \left( \hat{\beta}_n - \beta_0 \right) \equiv u \equiv \arg \min_u V_n (u) \) where

$$
V_n (u) \equiv (u - C_n^{-1} W_n)' C_n (u - C_n^{-1} W_n) + \lambda_n \sum_{\beta_{0j} \neq 0} |\sqrt{n} \beta_{0j} + u_j| + \lambda_n \sum_{\beta_{0j} = 0} |u_j| \quad (55)
$$

Define \( \hat{u} \equiv \arg \min_u \hat{V} (u) \) for

$$
\hat{V} (u) = (u - C_n^{-1} W_n)' C_n (u - C_n^{-1} W_n) + \lambda_n \sum_{\beta_{0j} \neq 0} \text{sgn} (\beta_{0j}) u_j + \lambda_n \sum_{\beta_{0j} = 0} |u_j| \quad (56)
$$

We argue that \( \bar{u} = \hat{u} \) w.p.c 1. In order to do this, we need to show that \( \text{sign}(\sqrt{n} \beta_0 + \bar{u}) = \text{sign}(\beta_0) \). To see this, suppose to the contrary that \( \beta_0 \) and \( \sqrt{n} \beta_{0j} + \bar{u}_j \) have different signs for some \( j \) for which \( \beta_{0j} \neq 0 \). This implies that \( |\bar{u}_j| \geq \sqrt{n} \delta_1 \) for some \( j \) and some \( \delta_1 > 0 \). Since \( W_n = O_P (1) \) and \( C_n \Rightarrow C \) for nonsingular \( C \), on this event for some \( \delta_2 > 0 \),

$$
(\bar{u} - C_n^{-1} W_n)' C_n (\bar{u} - C_n^{-1} W_n) \geq \lambda_{\min} (C_n) |\bar{u} - C_n^{-1} W_n|^2 \geq \delta_2 |\bar{u}_j|^2 \geq \delta_2 \epsilon_2.
$$

Noting that \( \lambda_n = o (\sqrt{n}) \), this implies that \( V_n (\bar{u}) \geq \delta_3 |\bar{u}_j|^2 \geq \delta_3 n^2 \) for some \( \delta_3 > 0 \), so that \( V_n (\bar{u}) > V_n (C_n^{-1} W_n) \), which contradicts the fact that \( \bar{u} = \arg \min_u V_n (u) \). We conclude that \( \text{sgn} (\beta_{0j}) = \text{sgn} (\sqrt{n} \beta_{0j} + \bar{u}_j) \) when \( \beta_{0j} \neq 0 \). In other words, \( \hat{\beta}_{nj} \) is sign consistent (in the sense of Zhao and Yu (2006)) whenever \( \beta_{0j} \neq 0 \).
Denote \( \Lambda = \{ j : \beta_{0j} \neq 0 \} \). For clarity we impose the strong irrepresentable condition (IC) \( |C_2|C_1^{-1}\text{sgn}(\beta_\Lambda) | < 1 \) in Zhao and Yu (2006) which restricts the dependence between relevant and irrelevant regressors, although the results hold more generally. When \( \lambda_n \to \infty \), w.p.c.1, \( \hat{\beta}_{nj} = 0 \) for \( j \in \Lambda^c \) under (IC) arguments in Zhao and Yu (2006). Note that the FOC of (55) is given by

\[
C_{n,\Lambda} \bar{u}_\Lambda - C_{n,\Lambda} \left[ C_{n}^{-1} W_n \right]_\Lambda - C_{n,\Lambda,\Lambda^c} \left[ C_{n}^{-1} W_n \right]_{\Lambda^c} + \frac{1}{2} \lambda_n \text{sgn} \left( \sqrt{n} \hat{\beta}_\Lambda \right) = 0,
\]

which implies that

\[
\hat{u}_\Lambda = \left[ C_{n}^{-1} W_n \right]_\Lambda + C_{n,\Lambda}^{-1} \left( C_{n,\Lambda,\Lambda^c} \left[ C_{n}^{-1} W_n \right]_{\Lambda^c} - \frac{1}{2} \lambda_n \text{sgn} \left( \beta_{0\Lambda} \right) \right).
\]

The FOC of (56) can be written explicitly as

\[
C_{n,\Lambda} \hat{u}_\Lambda - C_{n,\Lambda} \left[ C_{n}^{-1} W_n \right]_\Lambda - C_{n,\Lambda,\Lambda^c} \left[ C_{n}^{-1} W_n \right]_{\Lambda^c} + \frac{1}{2} \lambda_n \text{sgn} \left( \beta_{0\Lambda} \right) = 0,
\]

which implies that

\[
\hat{u}_\Lambda = \left[ C_{n}^{-1} W_n \right]_\Lambda + C_{n,\Lambda}^{-1} \left( C_{n,\Lambda,\Lambda^c} \left[ C_{n}^{-1} W_n \right]_{\Lambda^c} - \frac{1}{2} \lambda_n \text{sgn} \left( \beta_{0\Lambda} \right) \right).
\]

If we define \( \mathbb{G}_{n\Lambda}^0 \) as the law of \( u_\Lambda \equiv \left[ C^{-1} W \right]_\Lambda + C_{\Lambda,\Lambda^c}^{-1} \left( C_{\Lambda,\Lambda^c}^{-1} W_{\Lambda^c} - \lambda_n/2 \text{sgn} \left( \beta_{0\Lambda} \right) \right) \), then to show that \( \rho_{BL_1} \left( \hat{\mathbb{G}}_n, \mathbb{G}_{n}^0 \right) = o_P(1) \) since \( \hat{W}_n \overset{p}{\to} W \) and w.p.c.1, \( \hat{\beta}_{\Lambda^c} = \hat{u}_{\Lambda^c} = 0 \).

In the case of \( C_n \), note that \( \lambda_n \left( C_n^{-1} - C^{-1} \right) = \lambda_n C_n^{-1} (C - C_n) C^{-1} = o_P(1) \) since \( \lambda_n = o \left( \sqrt{n} \right) \) and \( C - C_n = O_P \left( \frac{1}{\sqrt{n}} \right) \), and both \( C_n^{-1} \) and \( C^{-1} \) are \( O_P(1) \). It is then the case that \( \rho_{BL_1} \left( \hat{\mathbb{G}}_n, \mathbb{G}_{n}^0 \right) = o(1) \).

In addition, it is clear that \( \mathbb{G}_{n\Lambda}^0 \) can be consistently estimated by, using \( \hat{W}_n \overset{p}{\to} W \), \( \hat{\Lambda} = \{ j : \hat{\beta}_{nj} \neq 0 \} \), \( \text{sgn} \left( \beta_{\hat{\Lambda}} \right) = \{ 1 \left( \hat{\beta}_{nj} > 0 \right) - 1 \left( \hat{\beta}_{nj} < 0 \right), j \in \hat{\Lambda} \} \), \( \hat{\mathbb{G}}_{n\Lambda}^* = 0 \),

\[
\hat{\mathbb{G}}_{n\Lambda}^* = \left[ C_{n}^{-1} W_n \right]_\Lambda + C_{n,\Lambda}^{-1} \left( C_{n,\Lambda,\hat{\Lambda}^c} \left[ C_{n}^{-1} \hat{W}_n \right]_{\hat{\Lambda}^c} - \frac{1}{2} \lambda_n \text{sgn} \left( \beta_{\hat{\Lambda}} \right) \right).
\]

Note that even if \( \lambda_n/\sqrt{n} \to \infty \), it is still possible to have \( \rho_{BL_1} \left( \hat{\mathbb{G}}_n, \mathbb{G}_{n}^0 \right) = o(1) \) if \( C_n \) is used in the definition of \( u_\Lambda \). Next we investigate the consistency of the numerical bootstrap.

Recall the definition of \( \hat{\beta}_n^* = \arg \min_{\beta} \mathbb{Z}_n^* (e - x' (\beta - \beta_0))^2 + \lambda_n \epsilon_n \sum_{j=1}^p |\beta_j| \) for
\[ e = y - x' \beta_0. \] Define \( \hat{u}^* = \epsilon_n^{-1} (\hat{\beta}^* - \beta_0) \). Then \( \hat{u}^* = \arg \min_u \mathcal{Z}_n^* \left( \frac{\epsilon}{\epsilon_n} - x'u \right) \] 
\[ + \lambda_n \sum_{j=1}^p |u_j + \frac{\beta_{0j}}{\epsilon_n}|, \]
which can be rewritten as
\[ \hat{u}^* = \arg \min_u V_n^* (u) \equiv (u - (B_n^*)^{-1} A_n^*) B_n^* (u - (B_n^*)^{-1} A_n^*) + \lambda_n \sum_{j=1}^p |u_j + \frac{\beta_{0j}}{\epsilon_n}| \]
for \( B_n^* = \mathcal{Z}_n^*xx' \) and \( A_n^* = \frac{1}{\epsilon_n} \mathcal{Z}_n^*xe \). Note first \( B_n^* = C_n + \epsilon_n \sqrt{n} (C_n^* - C_n) = C + o_P (1) \).
Next \( A_n^* = \frac{1}{\epsilon_n} \mathcal{Z}_n^*xe = \frac{1}{\sqrt{n} \epsilon_n} \mathcal{G}_nxe + \mathcal{G}_nxe = W_n^* + o_P (1) \frac{p}{\sqrt{w}} W \). Argue as before by contradiction that given \( \lambda_n \epsilon_n \to 0 \), w.p.c.1, \( \hat{u}^* = \arg \min_u \tilde{V}_n^* (u) \) where
\[ \tilde{V}_n^* (u) = (u - (B_n^*)^{-1} A_n^*) B_n^* (u - (B_n^*)^{-1} A_n^*) + \lambda_n \sum_{j=1}^p \text{sgn} (\beta_{0j}) u_j + \lambda_n \sum_{j=1}^p |u_j| \]
when \( \text{sgn} (\beta_{0j}) = \text{sgn} \left( \frac{\beta_{0j}}{\epsilon_n} + \hat{u}_j^* \right) \) whenever \( \beta_{0j} \neq 0 \). On its complement, \( |\hat{u}_j^*| \geq \delta_1/\epsilon_n \) for some \( \delta_1 > 0 \), and then for some \( \delta_2, \delta_3 > 0 \),
\[ \hat{Q} (\hat{u}^*) \geq \lambda_{\min} (B_n^*) |\hat{u}^* - (B_n^*)^{-1} A_n^*|^2 - \lambda_n \epsilon_n^{-1} \delta_1 \geq \delta_3 \epsilon_n^{-2} \hat{Q} (\hat{u}^*) \geq O_P (\lambda_n \epsilon_n^{-1}) \],
where \( \bar{u}^* = (B_n^*)^{-1} A_n^* \), which contradicts the definition of \( \hat{u}^* \). For \( \lambda_0 < \infty \), argue as in Knight and Fu (2000) that \( \hat{u}^* = O_P (1) \). Otherwise \( \exists M_n \to \infty \) and \( \epsilon > 0 \) such that \( P_n (|\hat{u}^*| \geq M_n) > \epsilon \) infinitely often. But then \( \tilde{V}_n^* (\hat{u}^*) \geq \delta M_n^2 \) for some \( \delta > 0 \) while \( \tilde{V}_n^* (\bar{u}^*) = O_P (1) \), a contradiction. Using the convexity arguments in Pollard (1985) and the pointwise convergence (in probability) of \( \tilde{V}_n (u) \), we deduce that \( \hat{u}^* \overset{p}{\sim} \bar{u}^* \), where \( \bar{u}^* = \arg \min_u \bar{V}^* (u) \) for
\[ \bar{V}^* (u) = (u - C^{-1}W) C (u - C^{-1}W) + \lambda_0 \sum_{j=1, \beta_{0j} \neq 0}^p \text{sgn} (\beta_{0j}) u_j + \lambda_0 \sum_{j=1, \beta_{0j} = 0}^p |u_j| \]
Next we turn to the case of \( \lambda_n \to \infty \). In this case, note that \( B_n^* \) satisfies (IC) almost surely by bootstrap LLN. Then invoke Zhao and Yu (2006) to argue that w.p.c.1, \( \hat{u}_j^* = 0 \) for \( j \in \Lambda^c \). The FOC for \( \bar{V}_n^* (u) \) can be analytically written as
\[ \frac{B_{n,A}^* u_{A} - B_{n,A}^* [B_{n,A}^* - C_{n,A}^*] A_n^* - C_{n,A}^* [C_{n,A}^* - A_n^*] A_n^* \lambda_n/2 \text{sgn} (\beta_{0A}) = 0, \]
which implies that \( \hat{u}_A^* = [B_{n,A}^* - C_{n,A}^*] A_n^* + B_{n,A}^* [B_{n,A}^* - C_{n,A}^*] A_n^* \lambda_n/2 \text{sgn} (\beta_{0A}) \). Since \( B_{n,A}^* - C_{n,A} = O_P (1/\sqrt{n}) \) and \( \lambda = o (\sqrt{n}) \), \( \lambda_n (B_{n,A}^* - C_{n,A}) = o_P (1) \). Then by replacing \( B_{n}^* \) with \( C \) and \( A_n^* \) with \( W \) in \( \hat{u}_A^* \), it follows that \( \rho_{BL1} \left( \mathcal{G}_n^*, \mathcal{G}_n^{0,\Lambda} \right) = o_P (1) \). Finally, w.p.c.1,
\[ \hat{u}_{\Lambda^c} = \tilde{u}_{\Lambda^c} = 0. \]

To complete the proof we now show \( \epsilon_n^{-1} (\hat{\beta}_n - \beta_0) = o_P(1) \). This follows from \( \hat{\beta}_n - \beta_0 = O_P(\max(1, \lambda_n) / \sqrt{n}) \) and the assumption that \( \max(1, \lambda_n) / \sqrt{n}\epsilon_n \to 0 \), which requires \( \lambda_n = o(n^{1/4}) \). If we use \( \epsilon_n^{-1} (\hat{\beta}_n^* - \bar{\beta}) \), where \( \bar{\beta} \) is the OLS estimate, to form the numerical bootstrap distribution, then since \( \bar{\beta} - \beta_0 = O_P(1/\sqrt{n}) \), we only need \( \sqrt{n}\epsilon_n \to \infty \) without requiring \( \lambda_n = o(n^{1/4}) \).

Next we relax the (IC) condition in Zhao and Yu (2006). We focus on the case when \( |C_21C_{11}^{-1}\gamma| \neq 1 \) for all partitions of \( C \) into 1, 2 blocks and for all \( \gamma \) out of \( 2^p \) permutations of \( \pm 1 \). (This case can also be handled but with more complex notations.) Consider the first simulation example 1 in Zhao and Yu (2006) with a scalar irrelevant regressor \( \beta_3 \). When (IC) does not hold, the second KKT condition on Zhao and Yu (2006) page 2554 will be violated in either direction, resulting in all interior solutions. In this case, w.p.c.1, an interior solution is obtained. If \( C_21C_{11}^{-1}\sgn (\beta_1) > 1 \), then w.p.c.1, \( \hat{u}_3 > 0 \), in which case \( \hat{u} = C_1^{-1} (W_n - \frac{\lambda_n}{2} (\sgn (\beta_\Lambda), 1))' \). Likewise, if \( C_21C_{11}^{-1}\sgn (\beta_1) < -1 \), then w.p.c.1, \( \hat{u}_3 < 0 \), in which case \( \hat{u} = C_1^{-1} (W_n - \frac{\lambda_n}{2} (\sgn (\beta_\Lambda), -1))' \). Indeed, in Figure 1(a) in Zhao and Yu (2006), \( \hat{\beta}_{n3} \) is always positive.

More generally, we assume a partition of \( \Lambda^c = (\Lambda^c_\Lambda, \Lambda^c_\Lambda, \Lambda^c_0) \) of dimensions \( d^c_+, d^c_-, d^c_0 \), such that there is a corresponding partition of \( C = ((C_11, C_12)', (C_21, C_22)') \), with \( C_{11} \) of dimension \( (d_\Lambda + d^c_+ + d^c_-)^2 \), and \( C_{22} \) of \( (d^c_0)^2 \), so that \( |C_21C_{11}^{-1}(\sgn (\beta_\Lambda), 1_{d^c_+}, 1_{d^c_-})| < 1 \), (where the relation is understood elementwise) and that \( C_{11}^{-1} \) is sign preserving in the \( (\Lambda^c_\Lambda, \Lambda^c_\Lambda) \) components:

\[
\sgn \left( C_{11}^{-1} \left( \sgn (\beta_\Lambda), 1_{d^c_+}, -1_{d^c_-} \right) \right)_{\Lambda^c_\Lambda, \Lambda^c_\Lambda} = \left( \sgn (\beta_\Lambda), 1_{d^c_+}, -1_{d^c_-} \right)_{\Lambda^c_\Lambda, \Lambda^c_\Lambda}'.
\]

Under (57), it is easy to see that w.p.c.1,

\[
\sgn \left( C_{11}^{-1} \left( W_1 - \frac{\lambda_n}{2} \sgn (\beta_\Lambda), 1_{d^c_+}, -1_{d^c_-} \right) \right)_{\Lambda^c_\Lambda, \Lambda^c_\Lambda} = - \left( \sgn (\beta_\Lambda), 1_{d^c_+}, -1_{d^c_-} \right)_{\Lambda^c_\Lambda, \Lambda^c_\Lambda}'.
\]

which implies that w.p.c.1, \( \hat{u}_{\Lambda^c_0} \equiv 0 \), and

\[
\hat{u}_{\Lambda^c_\Lambda + d^c_+ + d^c_-} = C_{11}^{-1} \left( W_1 - \frac{\lambda_n}{2} \left( \sgn (\beta_\Lambda), 1_{d^c_+}, -1_{d^c_-} \right) \right)_{\Lambda^c_\Lambda, \Lambda^c_\Lambda} + o_P(1).
\]

Carefully going over the above arguments also shows that the numerical bootstrap
satisfies
\[ \hat{u}_{d_{\beta}} = C_{11}^{-1} \left( W_{1}^* - \frac{\lambda n}{2} \left( \text{sgn} (\beta) , 1 d_{\beta}, -1 d_{\beta} \right) \right) + o_P(1). \]
which thus provides valid inference for \( \hat{u} \) as long as not all \( \beta_j \equiv 0 \). □

**Proof for Theorem 9.1**
For \( \hat{u} = \sqrt{n} (\hat{\beta}_n - \beta_0) \), \( \hat{u} = \arg \min_u L_n(u) \) where
\[ L_n(u) \equiv n \left( \hat{Q}_n (\beta_0 + \frac{u}{\sqrt{n}}) - \hat{Q}_n (\beta_0) \right). \]
The goal is to show that \( \hat{u} - \bar{u} = o_P(1) \), where \( \bar{u} = \arg \min_u \bar{L}_n(u) \) for
\[ \bar{L}_n(u) \equiv \frac{1}{2} (u + H^{-1} \Delta_n)' H (u + H^{-1} \Delta_n) + \lambda n \sum_{j=1}^{k} \left( |\sqrt{n}\beta_0j + u_j| - |\sqrt{n}\beta_0j| \right) \]
By the same arguments as in the LASSO case, w.p.c.1, \( \text{sgn} \left( \sqrt{n}\beta_j^0 + \hat{u}_j \right) = \text{sgn} \left( \beta_j^0 \right) \) for those \( j \) such that \( \beta_j^0 \neq 0 \), since otherwise, w.p.c.1, \( \exists \delta > 0, \bar{L}_n(\bar{u}) > \delta n > \bar{L}_n(-H^{-1}\Delta_n) \). Let \( \bar{u} = \arg \min_u \bar{L}_n(u) \) for
\[ \bar{L}_n(u) \equiv \frac{1}{2} (u + H^{-1} \Delta)' H (u + H^{-1} \Delta) + \lambda n \sum_{j=1}^{k} (\text{sgn}(\beta_{0j}) 1 (\beta_{0j} \neq 0) u_j + 1 (\beta_{0j} = 0) |u_j|). \]
The OLS analysis in Theorem 8.1 shows that \( \rho_{BL_1}(L(\bar{u}), L(\bar{u})) = o(1) \), so we focus on \( \hat{u} - \bar{u} \). For \( \lambda_0 < \infty \), the above \( \bar{L}_n(u) \) then implies that \( \bar{u} = O_P(1) \), as argued before by contradiction. Next by the same arguments as in Pollard (1985), for any compact \( K \), \( \sup_{u \in K} | L_n(u) - \bar{L}_n(u) | = o_P(1) \) through pointwise convergence and convexity. Finally convexity is used to argue that \( \hat{u} - \bar{u} = o_P(1) \). Let \( \beta = |\hat{u} - \bar{u}| \).
For any \( \delta > 0 \), define \( \tilde{u} = \left( 1 - \frac{\delta}{\beta} \right) \hat{u} + \frac{\delta}{\beta} \bar{u} \). By convexity and uniform convergence over \( K \) and the definition of \( \bar{u} \), \( \exists \lambda > 0 \) such that
\[ \left( 1 - \frac{\delta}{\beta} \right) L_n(\hat{u}) + \frac{\delta}{\beta} L_n(\bar{u}) \geq L_n(\tilde{u}) \geq \bar{L}_n(\tilde{u}) + o_P(1) \]
\[ \geq \bar{L}_n(\tilde{u}) + \lambda \delta^2 + o_P(1) \geq L_n(\bar{u}) + \lambda \delta^2 + o_P(1). \]
which can be rewritten as \( \mathcal{L}_n(\check{u}) \geq \mathcal{L}_n(\check{u}) + \lambda \delta + o_P(1) \) since \( 1 - \delta^2 \geq 1 \). By choosing \( o_P(1) \) smaller than, e.g. \( \frac{1}{2} \delta \), \( P(|\check{u} - \check{u}| \geq \delta) \to 0 \). Note that \( \check{L}_n(\check{u}) > \check{L}_n(\check{u}) \) follows from \( H \) having full rank when \( \check{u} \) and \( \check{u} \) lie in the same quadrant. When they lie in different quadrants, a change of the directional derivative only increases \( \check{L}_n(\check{u}) \) further.

Next consider \( \lambda_n \to \infty \) but \( \lambda_n = o\left(n^{1/6}\right) \) (which is stronger than OLS). First we note that the uniform convergence over compact sets result in Pollard (1985) can be extended to \( \sup_{|u| = o(n^{1/6})} |\mathcal{L}_n(u) - \mathcal{L}_n(u)| = o_P(1) \), since (up to constant)

\[
E\mathcal{L}_n(u) - E\check{L}_n(u) = nO\left(\left(\frac{|u|}{\sqrt{n}}\right)^3\right) = o(1) \text{ when } |u| = o\left(n^{1/6}\right).
\]

Also assume \( x_i \) is bounded, so \( \max_{i \leq n} x_i / \sqrt{n} = O\left(1/\sqrt{n}\right) \). The arguments on the bottom of page 192 of Pollard (1991) then hold for \( \theta = o\left(n^{1/6}\right) \). Therefore (59) also applies and again \( \check{u} - \check{u} = o_P(1) \). A more detailed characterization of the asymptotic distribution of \( \check{u} \) can then proceed as in Zhao and Yu (2006) and the proof of Theorem 8.1, with \( H \) replacing \( C \). In particular, if \( H \) satisfies the (IC) condition in Zhao and Yu (2006), then w.p.c.1, \( \check{u}_j = \check{u}_j = 0 \) for \( j \in \Lambda^c \). But (IC) is not needed as long as more complex notations can be tolerated.

The analysis of numerical bootstrap is facilitated by the bootstrap convexity lemma in Hahn (1995). Note that \( \hat{\beta}_n = \arg \min_{b \in B} \mathcal{Z}_n^*(\rho_r(y - x'b) - \kappa)^+ + \lambda_n \epsilon_n |b||1| \). Consider first \( \lambda_0 < \infty \). If we renormalize using \( \check{u}^* = \epsilon_n^{-1}\left(\hat{\beta}_n - \beta_0\right) \), then for \( e = y - x'\beta_0 \), \( \rho_r^e(e) = (\rho_r(e) - \kappa)^+ \) and

\[
\mathcal{L}_n^*(u) = \frac{\mathcal{Z}_n^*(\rho_r(e - \epsilon_n x'u) - \rho_r^e(e))}{\epsilon_n^2} + \lambda_n \sum_{j=1}^p |u_j + \frac{\beta_j}{\epsilon_n}|.
\]

Define \( \hat{u}^* = \arg \min_u \mathcal{L}_n^*(u) \). As long as \( \hat{\beta}^* - \beta_0 = o_P(1) \), w.p.c.1, up to constant, \( |u_j + \frac{\beta_j}{\epsilon_n}| = \text{sgn}\left(\beta_j^0\right) u_j + 1 (\beta_j = 0) |u_j| \), but convexity allows for the direct derivation of the asymptotic distribution without having to show consistency first. It is easy to see that \( \epsilon_n^{-2} P(\rho_r(e - \epsilon_n x'u) - \rho_r^e(e)) = \frac{1}{2} u'Hu + o(1) \). Next \( \epsilon_n^{-2} (\mathcal{Z}_n^* - P) = \frac{1}{\sqrt{n\epsilon_n}} \mathcal{G}_n + \epsilon_n^{-1} \mathcal{G}_n^* \). Denote

\[
\Delta_i = x_i \tau (e_i \geq \kappa/\tau) - x_i (1 - \tau) (e_i \leq -\kappa / (1 - \tau))
\]

\[
R_i(u) = \rho_r(e_i - \epsilon_n x_i'u) - \rho_r^e(e_i) / \epsilon_n = \Delta_i u,
\]

82
\[ R_n(u) = R_n^1(u) + R_n^2(u) = \frac{1}{\sqrt{n} \epsilon_n} \mathcal{G}_n R_i(u) + \mathcal{G}_n^* R_i(u). \]

Since \( E R_i(u) = 0 \) and \(|R_i(u)| \leq |x_i u| 1(|e_i| \leq \epsilon_n |x_i u|)\), the variance calculations in Pollard (1985) and Hahn (1995) imply that \( R_n^1(u) = o_P^*(1) \) and \( R_n^2(u) = o_P^*(1) \). Consequently, sup_{u \in R} |\mathcal{L}_n^*(u) - \tilde{\mathcal{L}}_n^*(u)| = o_P^*(1), where now \( \Delta_n^* = \frac{1}{\sqrt{n} \epsilon_n} \mathcal{G}_n \Delta_i + \mathcal{G}_n^* \Delta_i \), so that \( \Delta_n^* \xrightarrow{p} \Delta \) and \( \tilde{\mathcal{L}}_n^*(u) \equiv \frac{1}{2} (u + H^{-1} \Delta_n^*)' H (u + H^{-1} \Delta_n^*) + \lambda_n \sum_{j=1}^k |\epsilon_n^{-1} \beta_j^0 + u_j| \).

A bootstrap version of (59) thus shows that \( \hat{u}^* - \bar{u}^* = o_P^*(1) \) with \( \bar{u}^* = \text{arg min}_u \tilde{\mathcal{L}}_n^*(u) \). Finally, the OLS analysis in Theorem 8.1 shows that \( \rho_{BL_1} (\mathcal{L} (\bar{u}^*), \mathcal{L} (\hat{u})) = o_P^*(1) \), which entails that \( \rho_{BL_1} (\mathcal{L} (\bar{u}^*), \mathcal{L} (\hat{u})) = o_P^*(1) \).

Now consider \( \lambda_n \to \infty \). We will require that \( \lambda_n^2 \epsilon_n \to 0 \). This is more stringent than \( \lambda_n = o \left( n^{1/6} \right) \) and further constrains the range of \( \epsilon_n \). Under this condition, Taylor expanding to the 3rd order shows that \( \sup_{|u| = O(\lambda_n)} |\epsilon_n^{-2} P (\rho_T (e - \epsilon_n x_i u) - \rho_T^* (e)) - \frac{1}{2} u' H u| = O (\epsilon_n \lambda_n^3) = o (1) \). Analogously, \( \sup_{|u| = O(\lambda_n)} Var (R_i(u)) = O (\epsilon_n \lambda_n^3) = o (1) \) pointwise. Then argue by convexity that \( \sup_{|u| = O(\lambda_n)} |R_n(u)| = o_P^*(1) \) and \( \hat{u} - \bar{u} = o_P^*(1) \).

As seen in the OLS case, the model selection consistency of the 1-norm SVM depends on \( H \) (analogous to \( C \) for OLS). The subset of \( \hat{u}_\lambda \) corresponding to nonzero \( \beta_\lambda \) will be substantially biased. As long as \( |H_{21} H_{11}^{-1} \gamma| \neq 1 \) for all partitions of \( H \) and for all permutations \( \gamma \) of \( \pm 1 \), then w.p.c.1, each component of \( \hat{u}_\lambda \) will either be perfectly selected to zero or be substantially biased as in \( \hat{u}_\lambda \). Only when the Zhao and Yu (2006) (IC) condition holds will there be \( \hat{u}_\lambda = 0 \) w.p.c.1. However, regardless of the degree of selection, it always holds that \( \rho_{BL_1} (\mathcal{L} (\bar{u}^*), \mathcal{L} (\hat{u})) = o_P^*(1) \). The reason is that each component of \( \bar{u}^* \) and \( \hat{u} \) is either identically zero w.p.c.1, or they have normal distributions that are shifted by a large bias term of order \( \lambda_n \), but the two bias terms are close up to \( o_P^*(1) \). Therefore \( \rho_{BL_1} (\mathcal{L} (\bar{u}^*), \mathcal{L} (\hat{u})) = o_P^*(1) \). \( \blacksquare \)