University of Toronto
Department of Economics

Working Paper 764

A Simple Specification Test for Models with Many Conditional Moment Inequalities

By Mathieu Marcoux, Thomas Russell and Yuanyuan Wan

November 14, 2023

A Simple Specification Test for Models with Many Conditional Moment Inequalities

Mathieu Marcoux* *Université de Montréal*

Thomas M. Russell† *Carleton University*

Yuanyuan Wan‡ *University of Toronto*

November 8, 2023

Abstract

This paper proposes a simple specification test for partially identified models with a large or possibly uncountably infinite number of conditional moment (in)equalities. The approach is valid under weak assumptions, allowing for both weak identification and non-differentiable moment conditions. Computational simplifications are obtained by reusing certain expensive-to-compute components of the test statistic when constructing the critical values. Because of the weak assumptions, the procedure faces a new set of interesting theoretical issues which we show can be addressed by an unconventional sample-splitting procedure that runs multiple tests of the same null hypothesis. The resulting specification test controls size uniformly over a large class of data generating processes, has power tending to 1 for fixed alternatives, and has power against certain local alternatives which we characterize. Finally, the testing procedure is demonstrated in three simulation exercises.

Keywords: Misspecification, Moment Inequality, Partial Identification, Specification Testing

We are grateful to audiences at Academia Sinica, the African Meeting of the Econometric Society, La Société Canadienne de Science Économique, the Canadian Econometric Study Group, the International Association for Applied Econometrics, the Penn State Alumni Conference, and Simon Fraser University. Thomas Russell gratefully acknowledges support from the Social Sciences and Humanities Research Council of Canada (No. 435-2022-1016), and Yuanyuan Wan gratefully acknowledges support from the Social Sciences and Humanities Research Council of Canada (No. 435190500) and the National Natural Science Foundation of China (No. 72073078). All errors are our own.

*Département de Sciences Économiques, Université de Montréal, C.P. 6128, succ. Centre-Ville, Montreal, Quebec, H3C 3J7, Canada. Email: mathieu.marcoux@umontreal.ca.

†Department of Economics, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada. Email: thomas.russell3@carleton.ca.

‡Department of Economics, University of Toronto, 150 St. George Street, Toronto, Ontario, M5S 3G7, Canada. Email: yuanyuan.wan@utoronto.ca.

1 Introduction

Despite requiring weaker assumptions than their point-identified counterparts, partially identified models are not immune to misspecification. For partially identified models defined by moment inequalities, misspecification occurs when no parameter vector satisfies all the inequalities simultaneously, leading to an identified set that is empty. In practice empty estimated identified sets are common. While this may be a result of model misspecification, it can also occur due to sampling uncertainty, especially when the model is close to being point-identified. These two alternatives have very different implications, making specification testing an especially important topic in the context of partially identified models.

This paper proposes a simple specification test for models defined by moment inequalities, and makes three main contributions relative to the existing literature. First, we extend the existing literature on specification testing in moment inequality models—which has focused primarily on the case of a finite number of unconditional moment inequalities—by presenting a specification test applicable to models with a continuum of conditional moment inequalities. For instance, this makes our test applicable to moment inequalities derived from support-function based representations of the identified set, a general and empirically relevant class of models. Second, we show our specification test is able to control size uniformly over a large class of data generating processes (DGPs) under weaker assumptions on the moment conditions compared to the existing literature; in particular, we do not require the existence of a polynomial minorant, and allow for moment functions that are not everywhere differentiable. After introducing our main results, we discuss these conditions at length and provide simple examples where they can fail. Third, our testing procedure is developed with computational concerns in mind, and so is computationally simpler than existing moment inequality specification tests. The result is a specification test that is many times faster to perform than it is to estimate the identified set or to construct a confidence set.

The proposed procedure uses a MinMax test statistic, with the minimum taken over the parameter space, and the maximum taken over a (possibly uncountable) collection of moment conditions. Finding the parameter vector that minimizes the MinMax test statistic is the most computationally intensive component of the procedure, a difficulty also shared by other specification tests in this literature. However, our method obtains computational gains relative to existing procedures by reusing the minimizing parameter vector when bootstrapping to compute the critical value. Because of our weak assumptions, reusing the minimizer in the bootstrap procedure introduces a new set of interesting theoretical issues which we show can be addressed by an unconventional sample-splitting procedure that runs multiple tests of the same null hypothesis. For this reason, we refer to our testing procedure as the *split-sample multiple test* (SSMT) procedure. We show that the SSMT procedure controls size uniformly over a large class DGPs, has power tending to 1 for

fixed alternatives, and has power against certain local alternatives.

The SSMT procedure requires selecting an appropriate subsample size in the sample splitting step, which has a strong influence on the finite sample performance of the test. We discuss the effect of the subsample size and other tuning parameters at length, illustrate the practical performance of our method in three simulation examples, and provide some practical advice for potential users.

1.1 Previous Literature

A number of papers have explored specification testing in partially identified models, including Guggenberger, Hahn, and Kim (2008) for linear moment inequalities, Santos (2012) for nonparametric instrumental variable models, as well as Romano and Shaikh (2008), Andrews and Guggenberger (2009), Galichon and Henry (2009), Andrews and Soares (2010), and Bugni, Canay, and Shi (2015) for more general models.[1] Much of the previous literature has focused on specification testing as a by-product of confidence set construction. A notable exception is Bugni, Canay, and Shi (2015), who deal explicitly with specification testing in a general class of moment inequality models, and use a test statistic similar to ours. They propose two tests: a re-sampling (RS) test, and a recycling (RC) test. They also show that these tests have favorable power properties relative to a conventional by-product test which is based on confidence set construction.

We build on the approach of Bugni, Canay, and Shi (2015) in three main ways. First, our focus is on a continuum of conditional moment inequalities, where the focus of Bugni, Canay, and Shi (2015) is on a finite number of unconditional moment inequalities.

Second, we show how to relax certain assumptions required by Bugni, Canay, and Shi (2015). Specifically, the results in Bugni, Canay, and Shi (2015) rely on their Assumption A6. They show that this assumption is implied by their Assumption A8, which posits the existence of a polynomial minorant on the criterion function used to define the identified set, and requires a uniform equicontinuity condition on the gradients of the moment functions. The polynomial minorant condition is common in the literature on inference for partially identified models, although it rules out a form of weak identification of the identified set, and is especially strong when applied to conditional moment inequalities. We discuss this condition in Section 3.4, and show that it can be violated in simple examples. Furthermore, the assumptions of Bugni, Canay, and Shi (2015) rule out moment conditions that are not differentiable. This is often the case, for instance, for moment conditions derived from support-function based characterizations of the identified set, a leading example with an uncountable number of moments. In contrast, the SSMT procedure does not require the existence of a polynomial minorant and does not assume the moment conditions are everywhere differentiable.

[1] See Section 5 of Molinari (2020) for a review and discussion.

Finally, the SSMT procedure has some computational advantages. For the RS test of Bugni, Canay, and Shi (2015), the researcher must compute the infimum of a bootstrap test statistic for each bootstrap sample, where the feasible region is given by the argmin set of a criterion function.[2] In the absence of any special structure of the moment conditions (e.g. linearity, convexity, or high-order differentiability and closed-form gradients), this problem can be difficult, and reaching a global optimum at each iteration can be expensive.[3] The problem is magnified in our setting with a continuum of conditional moment inequalities, where even evaluating the test statistic at a single parameter vector can be expensive.[4] In contrast, we show that the parameter vector that minimizes the test statistic can be reused when computing the critical value, saving substantial computation time for the models we consider.

The theoretical techniques in this paper are most similar to Andrews and Shi (2013) and Andrews and Shi (2017), who provide a method of inference for models defined by an uncountably infinite number of conditional moment inequalities. They construct a test of a null hypothesis that *a fixed parameter vector* satisfies all moment conditions, and invert this test to construct confidence sets for the true vector of model parameters. In contrast, we construct a test of the null hypothesis that *there exists a parameter vector* that satisfies all moment conditions, with a rejection of the test signalling model misspecification. We use an MinMax test statistic, and the uniform asymptotic analysis requires the consideration of a different set of drifting sequences than in Andrews and Shi (2017). Our resulting testing procedure is also very different, relying on a new sample-splitting procedure that runs multiple tests of a common null hypothesis.

There has been a longstanding interest in the misspecification of partially identified models. Earlier papers on the topic include Ponomareva and Tamer (2011) and Kaido and White (2013). More recently Kédagni, Li, and Mourifié (2020) study the problem of estimating outer sets—a computationally convenient alternative to estimating the identified set—and show that outer sets can provide misleading results when the underlying model is misspecified. Both Kédagni, Li, and Mourifié (2020) and Masten and Poirier (2021) also study the problem of salvaging falsified models; that is, salvaging models which have failed a test of correct specification. However, neither of these papers focus on the topic of specification testing. In that sense, our test is complementary to this literature. Other recent papers have also explored confidence sets that are robust to spurious

[2]This argmin set will *typically* be a singleton when the identified set is empty. When the researcher *knows* the argmin set is a singleton, there is no computational benefit of the SSMT procedure relative to Bugni, Canay, and Shi (2015) (at least in the case of a finite number of unconditional moments), since both procedures will reuse the minimizer in the bootstrap procedure. However, with a complex model (e.g. non-convex moment inequalities), it is typically not possible for the researcher to *know* if the argmin set is a singleton, in which case they must solve a nonlinear constrained optimization problem either way.

[3]A similar comment applies to the RC test of Bugni, Canay, and Shi (2015), which has the same feasible region as the RS test but instead computes the infimum of the GMS critical value.

[4]For instance, in our first simulation example in Section 4.1 with a continuum of moment inequalities and a three-dimensional parameter vector, it takes over 22 hours just to evaluate the test statistic on a sparse uniform grid with only 20 points in each dimension (or 8000 points total).

precision under model misspecification, including Andrews and Kwon (2019) and Stoye (2020).[5] These procedures deliver valid inference for the true vector of model parameters when the identified set is empty but the model is correctly specified, and provide valid inference for a pseudo-true vector of model parameters when the model is misspecified. Recently, Andrews and Kwon (2021) also provide a specification test for the case of a finite number of unconditional moment inequalities, but where the *null hypothesis* is that *the identified set is empty* (that is, with the typical null and alternative reversed). Their confidence set is constructed by inverting a level $1-\alpha$ upper confidence set for a "misspecification index," which is very close to our test statistic in the case of a finite number of unconditional moments. A specification test similar to the one considered here (that is, when the *alternative hypothesis* is that *the identified set is empty*) can be completed using the $1-\alpha$ lower confidence set for their misspecification index, but similar to Bugni, Canay, and Shi (2015) the validity of this procedure relies on a polynomial minorant condition and a uniform equicontinuity condition on the gradients of the moment functions.[6]

1.2 Roadmap

The remainder of the paper proceeds as follows. Section 2 introduces the main environment and motivating examples and then provides a high-level overview, including a discussion of the main theoretical challenges and how they are addressed by the SSMT procedure. Section 3 focuses on a formal presentation of the theoretical properties of the procedure, stating the main assumptions and providing the results on size control and power against fixed and local alternatives. Section 3 also compares our assumptions and local power to existing procedures, and discusses computation and the choice of tuning parameters. Section 4 presents three simulation examples to illustrate the performance of our method, and compares our method to Bugni, Canay, and Shi (2015) and Andrews and Shi (2017). Section 5 concludes. The proofs of the main results are provided in Appendix A, and additional details for the proofs, background for the simulation examples, and some additional simulation evidence is provided in the Online Supplementary Material.

2 Overview and Examples

In this section we begin with an overview of the testing environment, and introduce some motivating examples. The main purpose of this section is to explain how the SSMT procedure is able to obtain a computational advantage under weak assumptions, the new theoretical challenges that arise as a result, and how the proposed procedure is able to overcome those challenges. To begin, let $\mathcal{P}$

[5] A spuriously precise confidence set in this context is one that does not cover any parameter vector with the desired coverage probability.

[6] See Theorem 7.1(b) in Andrews and Kwon (2021), which requires their Assumption A.8. Note that their Assumption A.8 is nearly identical to Assumption A.8 in Bugni, Canay, and Shi (2015).

denote a class of DGPs (defined formally in Assumption 3.1), suppose that $W_i \sim P \in \mathcal{P}$, and let X_i denote a subvector of W_i. Consider the following collection of conditional moment inequalities:

$$E_P[m(W_i, \theta, \tau) \mid X_i \,] \leq 0 \text{ a.s., } \forall \tau \in \mathcal{T} \subset \mathbb{R}^{d_\tau}, \tag{2.1}$$

where $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ is a vector of model parameters, and $\tau \in \mathcal{T}$ is an index. Here the index set $\mathcal{T}$ is arbitrary and may be uncountably infinite, leading to a model defined by an infinite number of conditional moment inequalities. The parameter space $\Theta \subset \mathbb{R}^{d_\theta}$ is also unrestricted, and may be non-compact and non-convex. Note that this framework also accommodates moment equalities, which can be expressed as two moment inequalities.[7] The following examples illustrate applications that fit this environment.

Example 1. *Consider the environment in Beresteanu, Molchanov, and Molinari (2011). There is a vector of observable random variables W_i and a vector of unobservable random variables ε_i defined on a non-atomic probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. The econometric model maps each realization (w, ε) to a nonempty closed set $Q_\theta(w, \varepsilon)$. It is assumed that the econometric model can be augmented with a selection mechanism, giving rise to a measurable $d_\tau-$dimensional map $\psi(W_i, \varepsilon_i, \theta)$ satisfying $\psi(W_i, \varepsilon_i, \theta) \in Q_\theta(W_i, \varepsilon_i)$ a.s. Let $Sel(Q_\theta)$ be the collection of all such measurable selections. Beresteanu, Molchanov, and Molinari (2011) define the identified set as:*

$$\Theta_I(P) := \{\theta \in \Theta : \exists \psi(W_i, \varepsilon_i, \theta) \in Q_\theta(W_i, \varepsilon_i) \text{ s.t. } E_P[h(W_i) \mid X_i] = E_P[\psi(W_i, \varepsilon_i, \theta) \mid X_i] \text{ a.s. } \},$$

where $h(\,\cdot\,)$ is a known function mapping W_i into $\mathbb{R}^{d_\tau}$, and where X_i is a subvector of W_i. Theorem 2.1 in Beresteanu, Molchanov, and Molinari (2011) shows that, under some additional assumptions, this identified set can be equivalently characterized as:

$$\Theta_I(P) := \left\{ \theta \in \Theta : \max_{\tau \in \mathbb{B}^{d_\tau}} \left(\tau^\top E_P[h(W_i) \mid X_i \,] - E_P\left[\sup_{q \in Q_\theta(W_i, \varepsilon_i)} \tau^\top q \;\middle|\; X_i \right] \right) = 0 \text{ a.s. } \right\},$$

where $\mathbb{B}^{d_\tau}$ is the $d_\tau-$dimensional unit ball, and where the second term is the expected support function of the random set $Q_\theta(W_i, \varepsilon_i)$. Thus, the set Θ_I can instead be written as the set of all $\theta \in \Theta$ satisfying an uncountable collection of conditional moment inequalities of the form:

$$E_P\left[\tau^\top h(W_i) - E_P\left[\sup_{q \in Q_\theta(W_i, \varepsilon_i)} \tau^\top q \;\middle|\; W_i \right] \;\middle|\; X_i \right] \leq 0 \text{ a.s. } \forall \tau \in \mathbb{B}^{d_\tau}.$$

Furthermore, each moment condition is a concave function of τ.[8] *In this context, the model is misspecified if $\Theta_I = \varnothing$. This support function characterization of the identified set is general, and includes examples like best linear prediction (Beresteanu, Molchanov, and Molinari (2011)),*

[7] In particular, the equality $x = 0$ can be written as $x \leq 0$ and $-x \leq 0$.

[8] This follows since $\tau \mapsto \sup_{q \in Q_\theta(\omega)} \tau' q$ is convex in τ.

static games with varying information structures (Magnolfi and Roncoroni (2023)) and solution concepts (Beresteanu, Molchanov, and Molinari (2011)), discrete choice with heterogeneous choice sets (Barseghyan, Coughlin, Molinari, and Teitelbaum (2021)), and others.

The previous example is our leading example of a model with a continuum of conditional moment inequalities. Our approach is also applicable to models with finite but large number of conditional moment inequalities, as illustrated in the next example.

Example 2. *Consider the generalized instrumental variable (GIV) framework of Chesher and Rosen (2017) and Chesher and Rosen (2020). In this framework, $W_i := (Y_i, Z_i)$, where Y_i is the observed vector of endogenous variables and Z_i is the observed vector of exogenous variables. There is also a vector of latent variables U_i that satisfies the selection relation $U_i \in \mathcal{U}(Y_i, Z_i; h)$ a.s., where:*

$$\mathcal{U}(y, z; h) := \{u \in \mathcal{U} : h(y, z, u) = 0\},$$

and where $h : \mathcal{Y} \times \mathcal{Z} \times \mathcal{U} \to \mathbb{R}$ is a known structural function. Let $\mathcal{P}_{U|Z}$ denote the collection of all conditional distributions of the latent variables U_i given Z_i. The researcher's model imposes the constraints on the pair $(h, \mathcal{P}_{U|Z})$. For instance, the researcher may impose that both h and $\mathcal{P}_{U|Z}$ are parametrically specified up to some finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$, and that $U \perp\!\!\!\perp Z$. In this case, Chesher and Rosen (2017) show that the identified set of model parameters is given by:

$$\Theta_I = \left\{\theta \in \Theta : P_{Y|Z}(\mathcal{U}(Y_i, Z_i; h) \subseteq S \mid Z_i) \leq P_U(S) \text{ a.s., for all } S \in \mathcal{S}\right\},$$

where $\mathcal{S}$ is some appropriate collection of "test sets." When (Y_i, Z_i) is discrete, $\mathcal{S}$ can be taken as a finite collection, in which case this provides a characterization of the model in terms of a finite number of conditional moment inequalities. Applications of this framework include discrete choice with heterogeneous choice sets, auctions, entry games, treatment effect estimation, among many others.

Even when a model is not partially identified, conditional moment inequalities can be useful when testing certain modelling assumptions, as the next example illustrates.

Example 3. *Consider testing the local average treatment effect (LATE) assumptions, as in Kitagawa (2015) and Mourifié and Wan (2017). Let $D_i \in \{0, 1\}$ and $Z_i \in \{0, 1\}$ be the binary treatment and instrument, respectively. Let Y_{i1} and Y_{i0} be two potential outcomes, let D_{i0} and D_{i1} be two potential treatments, and let X_i be a vector of covariates. The LATE assumptions are: (i) $(Y_{i1}, Y_{i0}, D_{i0}, D_{i1}) \perp\!\!\!\perp Z_i \mid X_i$, (ii) $P(D_i = 1 \mid Z_i = 1, X_i) \neq P(D_i = 1 \mid Z_i = 0, X_i)$ a.s., and (iii) $D_{i1} \geq D_{i0}$ or $D_{i0} \geq D_{i1}$ a.s. Under these assumptions the conditional LATE is identified by the Wald estimand. However, in the applied literature it is common for researchers to model*

the propensity score with a parametric model when the vector X_i is large.[9] *For instance, one can assume (iv)* $P(Z_i = 1 \mid X_i = x) = \Lambda(x, \theta_0) := \frac{\exp(x'\theta_0)}{1+\exp(x'\theta_0)}$ *for some unknown finite-dimensional parameter* $\theta_0 \in \mathbb{R}^{d_\theta}$. *We can then formulate the testable implications of Assumptions (i)-(iv) as a set of conditional moment inequalities. Define* $\mathcal{S} = \{[y, y'] : y < y',\ y, y' \in \mathcal{Y}\}$. *Then the LATE assumptions imply:*

$$
\begin{aligned}
&E\left[1\{Y_i \in S\}\left(\Lambda(X_i,\theta)D_i(1-Z_i) - (1-\Lambda(X_i,\theta_0))D_iZ_i\right) \mid X_i\right] \leq 0, && \forall S \in \mathcal{S},\\
&E[1\{Y_i \in S\}\left((1-\Lambda(X_i,\theta))(1-D_i)Z_i - \Lambda(X_i,\theta_0)(1-D_i)(1-Z_i)\right) \mid X_i] \leq 0, && \forall S \in \mathcal{S},\\
&E[\Lambda(X_i,\theta) - Z_i \mid X_i] \leq 0,\\
&E[Z_i - \Lambda(X_i,\theta) \mid X_i] \leq 0,
\end{aligned}
$$

X_i-a.s. for some vector $\theta \in \Theta$. When $\mathcal{Y}$ is finite, we have a finite number of conditional moment inequalities; otherwise, we have a continuum of conditional moment inequalities.

Inference for the true but partially identified vector $\theta_0 \in \Theta$ for models defined by moment inequalities of the form (2.1) has been addressed by Andrews and Shi (2017). In contrast, we develop a computationally simple method of specification testing for these models. In particular, we focus on testing the null hypothesis:

$$
H_0 : P \in \mathcal{P}_0 \text{ versus } H_1 : P \in \mathcal{P} \setminus \mathcal{P}_0, \tag{2.2}
$$

where $\mathcal{P}_0$ is the collection of null DGPs:

$$
\mathcal{P}_0 := \{P \in \mathcal{P} :\ \exists \theta \in \Theta \text{ s.t. } (\theta, P) \text{ satisfies (2.1)}\}.
$$

Here $\mathcal{P}_0 \subset \mathcal{P}$ is the subset of DGPs for which there exists at least one vector $\theta \in \Theta$ satisfying the collection of moment inequalities in (2.1).[10] In order to test the null hypothesis (2.2), our approach first converts the conditional moment inequalities in (2.1) into the following collection of unconditional moment inequalities:

$$
E_P[m(W_i, \theta, \tau)g(X_i)] \leq 0, \text{for all } \tau \in \mathcal{T}, \text{ and } g \in \mathcal{G}, \tag{2.3}
$$

where $g \in \mathcal{G}$ is an instrument function. If the collection of instrument functions $\mathcal{G}$ is suitably rich, converting the conditional moments in (2.1) to the unconditional moments in (2.3) is without any loss of identifying information.[11] This is the case, for instance, if $\mathcal{G}$ is a countable collection of hypercubes (see Andrews and Shi (2013) p.621 or Andrews and Shi (2017) p.279) or boxes (see

[9] See discussions in Słoczyński, Uysal, and Wooldridge (2022), among many others.

[10] In a partially identified model, it is possible for the identified set to be nonempty and for the model to be misspecified. These cases are not detectable by any test, but they suggest some care must be taken when interpreting the null hypothesis in (2.2) as "correct specification."

[11] See Andrews and Shi (2013) Lemma 2.

Andrews and Shi (2013) p.622). Continuing from (2.3), let $\{W_i\}_{i=1}^n$ be an i.i.d. sample from P, and define the sample analog unconditional moments as:

$$\bar{m}_n(\theta, \tau, g) = \frac{1}{n}\sum_{i=1}^{n} m(W_i, \theta, \tau, g), \qquad \text{where} \qquad m(W_i, \theta, \tau, g) = m(W_i, \theta, \tau)g(X_i). \tag{2.4}$$

To test the null hypothesis in (2.2), consider the following function:

$$T_n(\theta) := \sup_{\tau \in \mathcal{T}} \max_{g \in \mathcal{G}_n} \frac{\sqrt{n}\bar{m}_n(\theta, \tau, g)}{\hat{\varsigma}_n(\theta, \tau, g)}, \tag{2.5}$$

where $\hat{\varsigma}_n(\theta, \tau, g)$ is a positive data-dependent re-weighting of the moment function, and $\mathcal{G}_n \subset \mathcal{G}$ is a (growing) subset of the set of instrument functions. For example, $\hat{\varsigma}_n(\theta, \tau, g)$ could be taken as the sample standard deviation of the moment conditions:

$$\hat{\varsigma}_n(\theta, \tau, g) := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m(W_i, \theta, \tau, g) - \bar{m}_n(\theta, \tau, g))^2}.$$

Now for any sequence $\epsilon_n = o(1)$, let $\hat{\theta}_n \in \Theta$ be any (measurable) vector of parameters satisfying:[12]

$$T_n(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} T_n(\theta) + \epsilon_n. \tag{2.6}$$

Note that the quantity $T_n(\hat{\theta}_n)$ serves as a natural MinMax test statistic of the null hypothesis in (2.2). In particular, for sufficiently small values of ϵ_n, a large and positive value of $T_n(\hat{\theta}_n)$ indicates at least one moment inequality in (2.3) is violated for every $\theta \in \Theta$. On the other hand, negative values of $T_n(\hat{\theta}_n)$ indicate that there exists at least one $\theta \in \Theta$ (namely, $\hat{\theta}_n$) satisfying all of the moment inequalities in (2.3).

Given the test statistic $T_n(\hat{\theta}_n)$, we can specify a test function ϕ_n by comparing $T_n(\hat{\theta}_n)$ to an appropriate critical value. The conventional approach to obtain critical values is to approximate the distribution $T_n(\hat{\theta}_n)$ under the null using a resampling procedure, which typically involves repeatedly minimizing a bootstrap version of $T_n(\theta)$ over some approximation of the identified set. This is the approach taken, for instance, by Bugni, Canay, and Shi (2015) in their re-sampling (RS) critical value.[13] However, minimizing $T_n(\theta)$ over any subset $S \subset \Theta$ is a MinMax problem, and so is computationally expensive.[14]

Rather than repeatedly minimizing $T_n(\theta)$ when constructing a critical value, we propose reusing the approximate minimizer $\hat{\theta}_n$ in the resampling procedure. This avoids the most computationally

[12] In particular, the infimum in (2.5) may not be obtained under our assumptions.

[13] See also Andrews and Kwon (2021). In the related context of subvector inference, a similar approach is also taken by Bugni, Canay, and Shi (2017) and Belloni, Bugni, and Chernozhukov (2019).

[14] Even if $T_n(\theta)$ is replaced by another test statistic (e.g. as in Bugni, Canay, and Shi (2015)), or if an approach based on subsampling is used, the need to repeatedly minimize (a version) of $T_n(\theta)$ remains the most computationally burdensome aspect of constructing a critical value using a resampling procedure.

burdensome component of constructing the critical value, although potentially at the cost of some power loss in finite sample. However, without introducing strong assumptions, reusing $\hat{\theta}_n$ in the construction of the critical value introduces some additional complications. To understand why, suppose for simplicity that $\hat{\varsigma}_n(\theta, \tau, g) = 1$. Then $T_n(\hat{\theta}_n)$ can be rewritten as:

$$T_n(\hat{\theta}_n) = \sup_{\tau \in \mathcal{T}} \max_{g \in \mathcal{G}_n} \left\{ \underbrace{\sqrt{n}(\bar{m}_n(\hat{\theta}_n, \tau, g) - E_P[m(W_i, \hat{\theta}_n, \tau, g)])}_{\text{Empirical Process}} + \underbrace{\sqrt{n} E_P[m(W_i, \hat{\theta}_n, \tau, g)]}_{\text{Recentering}} \right\}. \quad (2.7)$$

That is, $T_n(\hat{\theta}_n)$ is the supremum of the sum of an empirical process and a recentering term evaluated at $\hat{\theta}_n$. The distribution of the empirical process can be approximated under standard assumptions using a variety of resampling procedures. However, the recentering term cannot be consistently estimated in a uniform sense. Furthermore, since the recentering term is evaluated at the data-dependent vector $\hat{\theta}_n$, without stronger assumptions it is possible that the recentering term converges to a value that is above, below, or equal to zero under the null hypothesis.[15] Combined with our desire to reuse $\hat{\theta}_n$ when resampling to compute the critical value, these features make an approach based on generalized moment selection (GMS) difficult to apply.

Nevertheless, constructing an appropriate critical value requires some understanding of the behavior of the recentering term under the null hypothesis. Lemma 3.1 provides the required result. Under some additional assumptions, the recentering term satisfies:

$$\max \left\{ \sup_{\tau \in \mathcal{T}} \max_{g \in \mathcal{G}_n} \sqrt{n} E_P[m(W_i, \hat{\theta}_n, \tau, g)], 0 \right\} = O_{\mathcal{P}_0}(1), \quad (2.8)$$

where $O_{\mathcal{P}_0}(1)$ indicates stochastic boundedness, uniformly in $P \in \mathcal{P}_0$.[16] As a result, for any sequence $\{q_n\}_{n=1}^{\infty}$ satisfying $q_n/n = o(1)$ we have:

$$\max \left\{ \sup_{\tau \in \mathcal{T}} \max_{g \in \mathcal{G}_n} \sqrt{q_n} E_P[m(W_i, \hat{\theta}_n, \tau, g)], 0 \right\} = o_{\mathcal{P}_0}(1), \quad (2.9)$$

where $o_{\mathcal{P}_0}(1)$ denotes convergence in probability to zero, uniformly in $P \in \mathcal{P}_0$. Intuitively, (2.8) suggests that, even under weak assumptions on the moment conditions, the ϵ_n-minimizer $\hat{\theta}_n$ is of sufficiently "high quality" that it prevents the recentering term from diverging to $+\infty$. To take advantage of the result in (2.9), consider the following modified test statistic:

$$T_{q_n}(\hat{\theta}_n) = \sup_{\tau \in \mathcal{T}} \max_{g \in \mathcal{G}_n} \left\{ \underbrace{\sqrt{q_n}(\bar{m}_{q_n}(\hat{\theta}_n, \tau, g) - E_P[m(W_i, \hat{\theta}_n, \tau, g)])}_{\text{Modified Empirical Process}} + \underbrace{\sqrt{q_n} E_P[m(W_i, \hat{\theta}_n, \tau, g)]}_{\text{Modified Recentering}} \right\},$$

where q_n satisfies $q_n/n \to 0$ and the mean function $\bar{m}_{q_n}(\hat{\theta}_n, \tau, g)$ is computed using a random subsample $\{W_{i_k}\}_{k=1}^{q_n}$ from the original sample $\{W_i\}_{i=1}^{n}$. Here the distribution of the modified empirical

[15] See Section S.4.2 of the Online Supplementary Material for a simple example where the recentering term is positive under the null.

[16] See Lemma 3.1 for a precise definition.

process can still be approximated by a resampling procedure, using the subsample $\{W_{i_k}\}_{k=1}^{q_n}$ rather than the original sample $\{W_i\}_{i=1}^n$. Furthermore, the modified recentering term will converge in probability uniformly over $\mathcal{P}_0$ to a value bounded above by zero. It follows that under the null the quantiles of the distribution of $T_{q_n}(\hat{\theta}_n)$ will be asymptotically (possibly over-)approximated by the quantiles of the bootstrap distribution of the quantity:

$$T_{q_n}^\sharp(\hat{\theta}_n) = \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{\sqrt{q_n}(\bar{m}_{q_n}^\sharp(\hat{\theta}_n,\tau,g) - \bar{m}_{q_n}(\hat{\theta}_n,\tau,g))}{\hat{\varsigma}_n(\theta,\tau,g)},$$

where $\bar{m}_{q_n}^\sharp(\hat{\theta}_n,\tau,g)$ is the bootstrap moment function computed by sampling i.i.d. draws from $\{W_{i_k}\}_{k=1}^{q_n}$ with replacement. Using the $1-\alpha$ quantile of the bootstrap distribution of $T_{q_n}^\sharp(\hat{\theta}_n)$, we can construct a critical value $c_{q_n}(1-\alpha)$ such that $T_{q_n}(\hat{\theta}_n) > c_{q_n}(1-\alpha)$ with probability at most α uniformly over $P\in\mathcal{P}_0$.

A final issue to address is the use of the subsample $\{W_{i_k}\}_{k=1}^{q_n}$ when constructing the modified test statistic. While the full sample is needed to construct $\hat{\theta}_n$, our final test statistic uses only $q_n \ll n$ observations, even though the remaining observations not used in constructing our test statistic $T_{q_n}(\hat{\theta}_n)$ can still be informative. To make use of the remaining observations, in practice we suggest dividing the sample into r_n samples of size q_n (with $r_n\cdot q_n \approx n$), and repeating the procedure described above for the $r=1,\ldots,r_n$ subsamples. The final SSMT procedure is as follows:

Step 1: Compute $\bar{m}_n(\theta,\tau,g)$ and $\hat{\varsigma}_n(\theta,\tau,g)$, and find an approximate minimizer $\hat{\theta}_n$ (in the sense of (2.6)) of the function $T_n(\theta)$ from (2.5).

Step 2: Fix q_n and r_n satisfying $q_n/n\to 0$ and $r_n\cdot q_n \le n$. Divide the sample $\{W_i\}_{i=1}^n$ into r_n non-overlapping subsamples of size q_n, and compute $T_{q_n}^{(1)}(\hat{\theta}_n),\ldots,T_{q_n}^{(r_n)}(\hat{\theta}_n)$ using each subsample, where:

$$T_{q_n}^{(r)}(\theta) := \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{\sqrt{q_n}\bar{m}_{q_n}^{(r)}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)}.$$

Step 3: Fix some large integer B. For $r=1,\ldots,r_n$ and each of the B bootstrap samples, compute the bootstrap test statistic $T_{q_n}^{(r)\sharp}(\hat{\theta}_n)$ given by:

$$T_{q_n}^{(r)\sharp}(\hat{\theta}_n) := \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{\sqrt{q_n}(\bar{m}_{q_n}^{(r)\sharp}(\hat{\theta}_n,\tau,g) - \bar{m}_{q_n}^{(r)}(\hat{\theta}_n,\tau,g))}{\hat{\varsigma}_n(\hat{\theta}_n,\tau,g)},$$

where:

$$\bar{m}_{q_n}^{(r)\sharp}(\theta,\tau,g) = \frac{1}{q_n}\sum_{i=1}^{q_n} m(W_i^{(r)\sharp},\theta,\tau)g(X_i^{(r)\sharp}).$$

Step 4: Fix some infinitesimal $\eta>0$ and some constant or decreasing sequence $\rho_n\in[1,\infty)$, and

for $r = 1, \ldots, r_n$ choose $c_n^{(r)\sharp}(1-\alpha/\rho_n+\eta)$ as the $1-\alpha/\rho_n+\eta$ quantile of the bootstrap distribution of $T_{q_n}^{(r)\sharp}(\hat{\theta}_n)$.

Step 5: Reject the null hypothesis in (2.2) at level α if $\phi_n(\rho_n, \alpha) = 1$, where:

$$\phi_n(\rho_n, \alpha) := \mathbb{1}\left\{\frac{1}{r_n}\sum_{r=1}^{r_n} \mathbb{1}\{T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho_n+\eta)+\eta\} \geq \frac{1}{\rho_n}\right\}. \tag{2.10}$$

The final step of the proposed procedure aggregates the results of the r_n "sub-tests" on each subsample to make a final rejection decision. The test function in (2.5) is designed to control the probability of rejection under the null when aggregating multiple sub-tests while also allowing the number of tests to grow indefinitely with the sample size. The test function also provides flexibility on the method used to aggregate the multiple tests through the choice of ρ_n. For instance, setting $r_n = \rho_n = \bar{r}$ for some fixed $\bar{r}$, we obtain:

$$\phi_n(\rho_n, \alpha) = \bigvee_{r=1}^{\bar{r}} 1\{T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\bar{r}+\eta)+\eta\},$$

which is precisely the Bonferroni method of aggregating multiple tests, rejecting (approximately) if at least one of $\bar{r}$ tests rejects at level $\alpha/\bar{r}$. Furthermore, setting $\rho_n = 1$ for any r_n we obtain:

$$\phi_n(\rho_n, \alpha) = \bigwedge_{r=1}^{\bar{r}} 1\{T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1-\alpha+\eta)+\eta\}.$$

which rejects (approximately) if all of the $\bar{r}$ tests reject at level α. Other intermediate cases are also possible. For instance, setting $\rho_n = 2$ for any r_n we obtain:

$$\phi_n(\rho_n, \alpha) = \mathbb{1}\left\{\frac{1}{r_n}\sum_{r=1}^{r_n} \mathbb{1}\{T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1-\alpha/2+\eta)+\eta\} \geq \frac{1}{2}\right\},$$

which rejects if half of the tests reject at the level $\alpha/2$. A similar method of aggregating multiple tests of a common null has been considered by Rüschendorf (1982) and Meng (1994). An extensive discussion on the value of ρ_n is given in Section 3.6 where we argue that ρ_n should be set so that $r_n/\rho_n \in \mathbb{N} \setminus \{1\}$. In Section 3.6 we also discuss the choice of q_n, r_n, and η, and recommend $q_n = \lfloor n^{4/5} \rfloor$, $r_n = \lfloor n^{1/5} \rfloor$, and $\eta = 10^{-6}$.

Note the use of multiple test statistics means that the SSMT procedure has some superficial connections to the literature on multiple testing. However, the procedure deviates from the classic multiple testing problem by it using many test statistics to test a common null hypothesis, rather than many test statistics to test distinct null hypotheses. This makes most of the methods from the literature on multiple testing inapplicable in this setting.[17] However, our method of aggregating

[17]For instance, Holm's method is equivalent to the Bonferroni method when all the null hypotheses are identical.

tests statistics using the test function in (2.10) typically offers power improvements over the use of a single test statistic, and also allows us to avoid the computationally expensive task of estimating the joint distribution of the test statistics.

In summary, the SSMT procedure offers computational advantages relative to existing specification tests by reusing the ϵ_n-minimizer of the MinMax test statistic when computing the critical value. Under our weak assumptions, reusing the minimizer introduces a novel theoretical complication, as it means that the recentering term can converge to a positive value under the null hypothesis. To account for this aberrant behavior of the recentering term, we propose a sample-splitting procedure and use multiple test statistics to test the null hypothesis, controlling size using the test function from (2.10).

The purpose of this section was to introduce the intuition behind the SSMT procedure, describing the main theoretical challenges and how the procedure overcomes those challenges. The next section focuses on the theoretical properties of the SSMT procedure, stating the formal assumptions required for the approach, as well as the main results on size control and power against fixed and local alternatives.

3 Methodology

In this section we first introduce the formal assumptions and main results on size control and power against fixed and local alternatives. We then compare the assumptions and power results to other approaches, and we provide some guidance on the tuning parameters and computation.

3.1 Main Assumptions

We begin by formally stating the main assumptions. The first assumption constrains the moment conditions, the class of DGPs $\mathcal{P}$, the parameter space Θ, the instrument functions $\mathcal{G}$, and the index set $\mathcal{T}$. In the statement of the first assumption, $\varsigma_P(\theta, \tau, g)$ represents a weight function. The relation between $\varsigma_P(\theta, \tau, g)$ and the data-dependent weight function $\hat{\varsigma}_n(\theta, \tau, g)$ from the previous section is clarified in Assumption 3.2 ahead.

Assumption 3.1. *The moment functions $m(w, \theta, \tau, g) := m(w, \theta, \tau)g(x)$, the parameter space $(\mathcal{P}, \Theta)$, and the instrument functions $g \in \mathcal{G}$ satisfy the following conditions:*

(i) $\{W_i : 1 \leq i \leq n\}$ *are i.i.d. under some* $P \in \mathcal{P}$.

(ii) $\varsigma_P(\theta, \tau, g) > 0$ *for all* $\theta \in \Theta$, $\tau \in \mathcal{T}$, $g \in \mathcal{G}$.

(iii) $|m(w, \theta, \tau, g)/\varsigma_P(\theta, \tau, g)| \leq \boldsymbol{M}(w)$, $\forall w \in \mathbb{R}^{d_w}$, $\forall \theta \in \Theta$, $\forall \tau \in \mathcal{T}$, $\forall g \in \mathcal{G}$, *for some measurable, finite envelope function* $\boldsymbol{M} : \mathbb{R}^{d_w} \to [0, \infty)$ *satisfying* $\sup_{P \in \mathcal{P}} E_P[\boldsymbol{M}(W_i)^{2+\delta}] \leq C$ *for some*

$C < \infty$ *and* $\delta > 0$.

(iv) The class of functions $\mathcal{M}_P := \{m(\cdot, \theta, \tau, g)/\varsigma_P(\theta, \tau, g) : (\theta, \tau, g) \in \Theta \times \mathcal{T} \times \mathcal{G}\}$ *is pointwise measurable for every* $P \in \mathcal{P}$ *and satisfies Dudley's entropy condition for the envelope* $\boldsymbol{M}$*, uniformly in* $P \in \mathcal{P}$.

Condition (i) in Assumption 3.1 assumes that the data is drawn i.i.d. according to some $P \in \mathcal{P}$, and condition (ii) imposes that each weight $\varsigma_P(\theta, \tau, g)$ is strictly positive. Condition (iii) assumes the existence of a measurable envelope function for all moment conditions which has a bounded $2 + \delta$ moment uniformly over $\mathcal{P}$, which is required for the application of a uniform central limit theorem. Condition (iv) requires a weak measurability condition to hold for the moment functions and restricts the complexity of the class of moment functions. Both pointwise measurability and Dudley's entropy condition are defined formally in the Online Supplementary Material. Importantly, condition (iv) is satisfied by many examples in the moment inequalities literature.[18]

To accommodate the conditional moment inequalities in (2.1), we follow Andrews and Shi (2013) and Andrews and Shi (2017) and convert the conditional moment inequalities to an equivalent set of unconditional moment inequalities using a collection of instrument functions. Assumption 3.1 is stated for a class of instrument functions $\mathcal{G}$, which should satisfy the conditions of Andrews and Shi (2013) and Andrews and Shi (2017) to exhaust all identifying information from the conditional moment inequalities.[19] This is not necessary for the validity of the SSMT procedure, but is important for power. For instance, $\mathcal{G}$ could be the class of countable hyperrectangles or boxes.[20] In practice, it is computationally infeasible to work with most countable classes $\mathcal{G}$, so all results ahead are instead stated for any finite nested sequence $\mathcal{G}_n \subset \mathcal{G}$ with the understanding that $\mathcal{G}_n \uparrow \mathcal{G}$.

While the SSMT procedure is applicable to a continuum of moment inequalities, it also covers a number of other special cases. For instance, taking $\mathcal{T}$ as a finite set, the framework covers a finite number of moment inequalities. Furthermore, taking $\mathcal{T} = \{1, \ldots, K\} \times \mathcal{T}'$, the framework covers moment functions of the form $m(w, \theta, \tau, g)/\varsigma_P(\theta, \tau, g) = m_k(w, \theta, \tau')g_k(x)/\varsigma_{P,k}(\theta)$, as considered in Andrews and Shi (2017). This latter example is important, since it restricts $\varsigma_P(\theta, \tau, g)$ to depend on a finite index. In general, some care must be taken when $\varsigma_P(\theta, \tau, g)$ is allowed to depend on $g \in \mathcal{G}$ and $\tau \in \mathcal{T}$, since it can be zero or (arbitrarily close to zero) for some (g, τ), causing Assumption 3.1(iii) and (iv) to fail. Using a constant weight function, adding a small positive non-vanishing constant to the weight function, or using a weight function that depends only on θ (as in Andrews and Shi (2017)) can help guard against these failures.

[18] For instance, a sufficient (but not necessary) condition is that the weight function $\varsigma_P(\theta, \tau, g)$ is uniformly bounded away from zero, and that the class of moment functions are VC-type (see Giné and Nickl (2021) Definition 3.6.10).

[19] See Andrews and Shi (2013) Lemma 2.

[20] See Andrews and Shi (2013) p.621 - 622 and Andrews and Shi (2017) p.279.

Some additional structure on the moment functions may be helpful in verifying Assumption 3.1 and in deriving other primitive conditions for the uniform size control result (see Remark 3.2). However, the main results do not require the existence of a polynomial minorant, and do not require that the moment functions be everywhere differentiable.[21] These conditions are discussed further in Section 3.4 after introducing the main theoretical results.

The second main assumption we require relates the data-dependent weights $\hat{\varsigma}_n(\theta, \tau, g)$ discussed in the previous section to the weight function $\varsigma_P(\theta, \tau, g)$ in Assumption 3.1.

Assumption 3.2. *For any $\varepsilon > 0$:*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} Pr_P \left(\sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \sup_{g\in\mathcal{G}} \left| \frac{\varsigma_P(\theta, \tau, g)}{\hat{\varsigma}_n(\theta, \tau, g)} - 1 \right| > \varepsilon \right) = 0.$$

Assumption 3.2 requires that the quantities $\hat{\varsigma}_n(\theta, \tau, g)$ converge uniformly in probability to the weight function $\varsigma_P(\theta, \tau, g)$ satisfying Assumption 3.1. This requirement is relatively weak, and allows for a variety of natural weighting schemes. For instance, a possible choice is to set $\hat{\varsigma}_n(\theta, \tau, g)$ as a constant, as the sample standard deviation of the moment functions, or as an upper bound on the moment functions. Reweighting the moment functions is not required for any of the theoretical properties of the SSMT procedure to hold, although different weights may imply different finite-sample properties.

3.2 The Testing Procedure and Size Control

To test the null hypothesis in (2.2), recall the function from (2.5):

$$T_n(\theta) := \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{\sqrt{n}\bar{m}_n(\theta, \tau, g)}{\hat{\varsigma}_n(\theta, \tau, g)},$$

where $\bar{m}_n(\theta, \tau, g)$ is the sample average unconditional moment defined in (2.4). Let $\hat{\theta}_n$ denote any ϵ_n-minimizer of $T_n(\theta)$, as in (2.6). The computational gains from the SSMT procedure come from the fact that it reuses $\hat{\theta}_n$ when computing the critical value. However, as discussed in the previous section, controlling size under weak assumptions while reusing $\hat{\theta}_n$ requires some understanding of the recentering term from (2.7). The following Lemma formalizes the result stated in (2.8) in the previous section.

Lemma 3.1. *Suppose that Assumptions 3.1 and 3.2 hold, let $\mathcal{G}_n \subset \mathcal{G}$ be a nested sequence, and suppose that $\epsilon_n = o(1)$, where $\hat{\theta}_n$ and ϵ_n are from (2.6). Then for any nested sequence $\mathcal{G}_n \subset \mathcal{G}$:*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0} Pr_P \left(\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{\sqrt{n}E_P[m(W_i, \hat{\theta}_n, \tau, g)]}{\varsigma_P(\hat{\theta}_n, \tau, g)} > M \right) = 0.$$

[21] See Bugni, Canay, and Shi (2015) Assumption A.8, Bugni, Canay, and Shi (2017) Assumption A.3, Andrews and Kwon (2021) Assumption A.8.

Lemma 3.1 shows that the recentering term is stochastically bounded, uniformly over $\mathcal{P}_0$. As an immediate consequence of this result we obtain (2.9) in the previous section, which can be formally stated as:

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0} \Pr_P\left(\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{\sqrt{q_n}E_P[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_P(\hat{\theta}_n,\tau,g)} > M\right) = 0,$$

for any $M > 0$, where $\{q_n\}_{n\geq 1} \subset \mathbb{N}$ is a diverging sequence satisfying $q_n = o(n)$. The SSMT procedure described in Section 2 takes advantage of this result by running multiple tests of a common null, before aggregating the results of each test using the test function from (2.10).

The following assumption collects the conditions discussed in Section 2 that are required for the results ahead.

Assumption 3.3. *(i) $\hat{\theta}_n$ is a measurable sequence satisfying (2.6) for some sequence $\epsilon_n = o(1)$, (ii) $\mathcal{G}_n \subset \mathcal{G}$ is a nested sequence, (iii) $\rho_n \in [1,\infty)$ is a (weakly) decreasing sequence, (iv) $\{q_n\}_{n\geq 1} \subset \mathbb{N}$ is a diverging sequence satisfying $q_n = o(n)$, (v) $\{r_n\}_{n\geq 1} \subset \mathbb{N}$ is a (possibly diverging) sequence satisfying $q_n \cdot r_n \leq n$, (v) $\alpha \in (0, 1/2)$, (vi) $\eta \in (0, \alpha/\rho_1)$ is some infinitesimal constant, where ρ_1 is the first element of the sequence $\{\rho_n\}_{n\geq 1}$, (vii) the non-overlapping subsamples $\{W_i^{(r)}\}_{i=1}^{q_n}$ are constructed by sampling i.i.d. uniformly without replacement from $\{W_i\}_{i=1}^n$, (viii) for $r = 1,\ldots,r_n$, the bootstrap sample $\{W_i^{(r)\sharp}\}_{i=1}^{q_n}$ is constructed by sampling i.i.d. uniformly with replacement from $\{W_i^{(r)}\}_{i=1}^{q_n}$.*

With this assumption in hand, the first main result shows that the SSMT procedure described in Section 2 controls size uniformly over $P \in \mathcal{P}_0$. To understand the statement, define:

$$h_{2,P} := h_{2,P}(\theta,\tau,g,\theta^\dagger,\tau^\dagger,g^\dagger) := Cov(m(W_i,\theta,\tau,g), m(W_i,\theta^\dagger,\tau^\dagger,g^\dagger)),$$
$$\text{and } h_{3,P} := h_{3,P}(\theta,\tau,g) = \varsigma_P(\theta,\tau,g).$$

Now let $\mathscr{H} := \{(h_{2,P}, h_{3,P}) : P \in \mathcal{P}\}$, equipped with the sup-norm.

Theorem 3.1. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold, and consider the test function $\phi_n(\rho_n,\alpha)$ from (2.10). Then for any compact subset $\mathscr{H}_{cpt} \subset \mathscr{H}$:*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} (Pr_P \times \mathbb{P}^\sharp)(\phi_n(\rho_n,\alpha) = 1) \leq \alpha. \tag{3.1}$$

Remark 3.1. *Here $Pr_P \times \mathbb{P}^\sharp$ is the product measure over the sample space and the bootstrap sample space, where Pr_P is the $n-$fold product distribution (i.e. the sampling distribution) and $\mathbb{P}^\sharp$ defines the distribution of the bootstrap samples. See the end of Appendix A.1 for a formal description of these objects.*

Remark 3.2. *We follow Andrews and Shi (2013) and Andrews and Shi (2017) by restricting the pair* $(h_{2,P}, h_{3,P})$ *to a compact subset of* $\mathscr{H}$*. This is a technical requirement which ensures we can always extract a uniformly converging subsequence from any sequence* $\{(h_{2,P_n}, h_{3,P_n})\}_{n=1}^{\infty}$*.*[22] *However, any other conditions ensuring the existence of a uniformly convergent subsequence can be used; for instance, in the case of continuous (but possibly non-differentiable) moment functions, the Arzela-Ascoli Theorem can be used to derive primitive conditions. Note the condition is also trivially satisfied when considering pointwise asymptotics.*

Theorem 3.1 represents one of the main theoretical results, showing that the SSMT procedure for specification testing controls size uniformly over a large class of DGPs under weak assumptions. In Section 4 ahead we demonstrate the practical size control properties of the procedure in three simulation examples.

3.3 Power Results

Although the SSMT procedure described in Section 2 is valid under weak assumptions and offers computational advantages, researchers may be concerned about the potential loss of power that arises from reusing $\hat{\theta}_n$ to compute the critical value. We now formally present our results on power against both fixed and local alternatives. We then compare the power of the SSMT procedure to other approaches in Section 3.5. For the first result, we require the following assumption on the sequence of alternative distributions.

Assumption 3.4. *The sequence of alternative distributions* $\{P_n \in \mathcal{P} \setminus \mathcal{P}_0\}_{n\geq 1}$ *satisfy the following: (i)* $h_{2,P_n} \xrightarrow{u} h_{2,P_0}$ *and* $h_{3,P_n} \xrightarrow{u} h_{3,P_0}$ *for some* $P_0 \in \mathcal{P}$*, (ii)* $\varsigma_{P_n}(\theta, \tau, g) \leq \Delta_\varsigma < \infty$ *for all* n *and all* $(\theta, \tau, g) \in \Theta \times \mathcal{T} \times \mathcal{G}$*, (iii) there is some* $N \geq 1$ *such that for every* $n \geq N$ *there exists a (possibly* θ*-dependent) index* $\tau_{n,\theta} \in \mathcal{T}$*, and a subset* $\mathcal{X}_n(\theta) \subseteq \mathcal{X}$ *such that* $Pr_{P_n}(\mathcal{X}_n(\theta)) > \eta_v \ \forall \theta \in \Theta$*, and:*

$$\mu_n := \inf_{\theta \in \Theta} \inf_{x \in \mathcal{X}_n(\theta)} E_{P_n}[m(W_i, \theta, \tau_{n,\theta}) \mid X_i = x], \tag{3.2}$$

for some positive constant $\eta_v > 0$ *and some positive sequence* $\mu_n > 0$*, and (iv) there exists a* $g_{n,\theta} \in \mathcal{G}$ *and a positive constant* $\varepsilon_v > 0$ *such that* $g_{n,\theta}(x) \geq \varepsilon_v$ *for all* $x \in \mathcal{X}_n(\theta)$*.*

Assumption 3.4 places some restrictions on the sequence of alternative distributions, which are required to determine the local power properties of the SSMT procedure. Parts (i) and (ii) are mild regularity conditions, and are trivially satisfied (for instance) under our other maintained assumptions for any fixed alternative. Parts (iii) and (iv) formalize some properties of the violated conditional moments under the alternative sequence, and introduce some constants (η_v and ε_v) that appear in the proofs of the power results. Under this assumption, we have the following result.

[22] This is also closely related to Assumption (ii) in Theorem 3.1 in Sheehy and Wellner (1992).

Theorem 3.2. *Suppose Assumptions 3.1, 3.2, 3.3 and 3.4 are satisfied. If μ_n from (3.2) satisfies $\sqrt{q_n}\mu_n \to \infty$, then:*

$$\liminf_{n\to\infty}(Pr_{P_n} \times \mathbb{P}^\sharp)(\phi_n(\rho_n, \alpha) = 1) = 1.$$

Theorem 3.2 shows that the worst-case power for the SSMT procedure is asymptotically bounded below by 1 along any sequence of local alternatives satisfying Assumption 3.4 with $\sqrt{q_n}\mu_n \to \infty$. Note that if $\mu_n > 0$ is a fixed constant then $\sqrt{q_n}\mu_n \to \infty$, so Theorem 3.2 demonstrates that the SSMT procedure has power tending to 1 for any fixed alternative. Furthermore, $\sqrt{q_n}\mu_n \to \infty$ for any sequence of local alternatives where μ_n converges to zero slower than $q_n^{-1/2}$. For these local alternatives, Theorem 3.2 also shows the SSMT procedure has power tending to 1.

To understand the source of power, suppose for simplicity that $\hat{\varsigma}_n(\theta, \tau, g) = 1$ and consider the test statistic $T_{q_n}^{(r)}(\hat{\theta}_n)$:

$$T_{q_n}^{(r)}(\hat{\theta}_n) = \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left\{ \underbrace{\sqrt{q_n}(\bar{m}_{q_n}^{(r)}(\hat{\theta}_n, \tau, g) - E_P[m(W_i, \hat{\theta}_n, \tau, g)])}_{\text{Empirical Process}} + \underbrace{\sqrt{q_n}E_P[m(W_i, \hat{\theta}_n, \tau, g)]}_{\text{Recentering}} \right\}.$$

The bootstrap is able to consistently approximate the distribution of the empirical process, so the recentering term is ultimately responsible for the power of the test. The proof of Theorem 3.2 proceeds by showing that for suitable local alternatives $T_{q_n}^{(r)}(\hat{\theta}_n)$ diverges due to a diverging recentering term. This occurs while the empirical process remains stochastically bounded, so that the bootstrap critical value remains asymptotically small relative to the test statistic. For any sequence ρ_n satisfying Assumption 3.3, the proof of Theorem 3.2 shows that this will ultimately force a rejection of the null hypothesis in (2.2) using the SSMT procedure.

While Theorem 3.2 provides a lower bound on the power for certain local alternatives, it is silent on the power of the SSMT procedure for sequences that approach the null at rates faster than $q_n^{-1/2}$. This case is addressed by Theorem 3.3 below. In place of Assumption 3.4, Theorem 3.3 requires the following alternative assumption.

Assumption 3.5. *The sequence of alternative distributions $\{P_n \in \mathcal{P}\setminus\mathcal{P}_0\}_{n\geq 1}$ satisfy the following: (i) $h_{2,P_n} \overset{u}{\to} h_{2,P_0}$ and $h_{3,P_n} \overset{u}{\to} h_{3,P_0}$ for some $P_0 \in \mathcal{P}$, (ii) $\varsigma_{P_n}(\theta, \tau, g) \geq \underline{\Delta}_\varsigma > 0$ for all n and all $(\theta, \tau, g) \in \Theta \times \mathcal{T} \times \mathcal{G}$, (iii) there is some $N \geq 1$ such that for every $n \geq N$ there exists a θ_n and a (possibly θ_n-dependent) subset $\mathcal{X}_n(\theta_n) \subseteq \mathcal{X}$ such that $Pr_{P_n}(\mathcal{X}_n(\theta_n)) = 1$, and:*

$$\mu_n \geq \sup_{\tau\in\mathcal{T}} \sup_{x\in\mathcal{X}_n(\theta_n)} E_{P_n}[m(W_i, \theta_n, \tau) \mid X_i = x], \tag{3.3}$$

for some positive sequence $\mu_n > 0$, and (iv) the class of functions $\mathcal{G}$ is uniformly bounded by some $\bar{G} < \infty$.

While Assumption 3.4 imposes that at least one conditional moment is sufficiently large (but

possible violated), Assumption 3.5 requires that all conditional moments are sufficiently small (but possibly violated). Intuitively, this is because Theorem 3.2 provides a lower bound on local power for alternatives that are sufficiently distant from the null, where the following result provides an upper bound on local power for alternatives that are close to the null.

Theorem 3.3. *Suppose Assumptions 3.1, 3.2, 3.3 and 3.5 are satisfied. If μ_n from (3.3) satisfies $\sqrt{q_n}\mu_n \to 0$, then:*

$$\limsup_{n\to\infty}(Pr_{P_n} \times \mathbb{P}^\sharp)(\phi_n(\rho_n, \alpha) = 1) \leq \alpha.$$

Theorem 3.3 shows that the SSMT procedure has no non-trivial local power against alternative sequences that approach the null faster than $q_n^{-1/2}$. In contrast, in Section 3.5 we show that competing methods have non-trivial power against some $n^{-1/2}-$local alternatives. In this sense, the result makes explicit the power loss associated with our proposed sample splitting method, and quantifies the cost of using a procedure that is both computationally simple and valid under weaker assumptions.

Note the power analysis makes it clear that the finite sample properties of the SSMT procedure under both fixed and local alternative sequences depends on the magnitude of the term $E_P[m(W_i, \hat{\theta}_n, \tau, g)]$, and suggests that power against fixed and local alternatives is higher when q_n is larger. However, as discussed in Section 2, there can also exist null DGPs $\{P_n\}_{n=1}^\infty$ along which $E_{P_n}[m(W_i, \hat{\theta}_n, \tau, g)]$ is always positive.[23] For these DGPs a larger sequence q_n may cause the SSMT procedure to have poor finite sample size control properties. The tension between large and small values of the subsample size q_n means it is an important tuning parameter in our framework. We discuss possible choices of q_n in Section 3.6, and we explore the impact of various values for this tuning parameter in the simulation exercises in the next section, as well as in Section S.4 of the Online Supplementary Material.

3.4 Discussion of the Assumptions

One of the main advantages of the SSMT procedure for specification testing is that it is valid under weaker assumptions than those required by other specification tests in the literature. In this section we review two main assumptions in the existing literature that are *not* required by the SSMT procedure.

3.4.1 The Existence of a Polynomial Minorant

One of the main assumptions that is *not* required by the SSMT procedure is the existence of a polynomial minorant. This assumption is required in both Bugni, Canay, and Shi (2015) and

[23] See Section S.4.2 of the Online Supplementary Material.

Andrews and Kwon (2021), among many others.[24] To match with these papers, consider the case of a finite number of unconditional moment inequalities, and let $\sigma_{P,k}(\theta)$ denote the standard deviation of the k^{th} moment. Furthermore, define:

$$\Theta_I(P) := \left\{ \theta \in \Theta : \max_{k=1,\ldots,K} \left[\frac{E_P[m_k(W_i,\theta)]}{\sigma_{P,k}(\theta)} \right]_+ = 0 \right\},$$

where $[\,\cdot\,]_+ = \max\{0, \cdot\}$. The polynomial minorant condition in Bugni, Canay, and Shi (2015) is given by the following.

Assumption 3.6 (Polynomial Minorant)**.** *There exists constants $c, \delta > 0$ such that for all $(\theta, P) \in \Theta \times \mathcal{P}_0$:*

$$\max_{k=1,\ldots,K} \left[\frac{E_P[m_k(W_i,\theta)]}{\sigma_{P,k}(\theta)} \right]_+ \geq c \min \left\{ \delta, \inf_{\bar{\theta} \in \Theta_I(P)} ||\theta - \bar{\theta}|| \right\}.$$

To appreciate the strength of this condition, consider the special case when $d_\theta = 1$, $K = 1$, and $\sigma_{P,k}(\theta) = 1$. In this case Assumption 3.6 requires:

$$\max\{E_P[m_k(W_i,\theta)], 0\} \geq c \min \left\{ \delta, \inf_{\bar{\theta} \in \Theta_I(P)} |\theta - \bar{\theta}| \right\}.$$

Intuitively, the condition says that the moment $E_P[m_k(W_i,\theta)]$ must "lift off" the set $\Theta_I(P)$ at a rate that is locally bounded below by a linear function in $\theta \in \Theta$. Furthermore, this must hold for all $(\theta, P) \in \Theta \times \mathcal{P}_0$.

For a simple practical example when this fails, consider the linear regression model $Y_i = X_i\theta + \varepsilon_i$ where Y_i is an outcome variable of interest, X_i is an endogenous variable and Z_i is a candidate instrumental variable. In the spirit of Nevo and Rosen (2012), suppose that Z_i is an imperfect instrument in the sense that it satisfies $E_{P_n}[Z_i\varepsilon_i] \geq 0$ rather than $E_{P_n}[\varepsilon_i \mid Z_i] = 0$. Furthermore, suppose that Z_i is weak in the sense that $E_{P_n}[Z_iX_i] = \eta_n \downarrow 0$. In this case, the model implies the following moment inequality:

$$E_{P_n}[m(W_i,\theta)] = E_{P_n}[Z_i(X_i\theta - Y_i)] \leq 0.$$

Then $\Theta_I(P_n) = \{\theta \in \mathbb{R} : \theta \leq E_{P_n}[Z_iY_i]/E_{P_n}[Z_iX_i]\}$. Let $\tilde{\theta} = E_{P_n}[Z_iY_i]/E_{P_n}[Z_iX_i]$. Then for any $\theta \notin \Theta_I(P_n)$ we have:

$$\max\{E_{P_n}[m(W_i,\theta)], 0\} = E_{P_n}[Z_iX_i]\theta - E_{P_n}[Z_iY_i] = E_{P_n}[Z_iX_i](\theta - \tilde{\theta}) = \eta_n|\theta - \tilde{\theta}|.$$

[24] See Kaido, Molinari, and Stoye (2022) for a review of these conditions. The authors also show the connection between these conditions and constraint qualifications from the optimization literature.

But then clearly there *does not* exist universal constants (c, δ) such that

$$\max\{E_{P_n}[m(W_i, \theta)], 0\} \geq c \min\left\{\delta, \inf_{\tilde{\theta} \in \Theta_I(P_n)} |\theta - \tilde{\theta}|\right\}.$$

More generally, this assumption can fail when the slope of the moment conditions drift to zero near the identified set. In Section S.4.2 of the Online Supplementary Material we present simulation results for a similar example, and show that the procedure of Bugni, Canay, and Shi (2015) over-rejects under the null hypothesis, regardless of their choice of tuning parameters.

While this simple example shows that the polynomial minorant condition rules out certain null DGP sequences, in more complicated examples it is not always clear which sequences are ruled out. However, with conditional moment inequalities the polynomial minorant condition will generally prevent the correlation between the instruments and the moment functions to drift to zero under the null. This rules out weak identification of the identified set, a common occurrence with a large number of weak instrument functions.

3.4.2 Uniformly Equicontinuous Derivatives

Existing procedures also require strong smoothness conditions on the moment functions which can rule out some interesting models. In particular, both Bugni, Canay, and Shi (2015) and Andrews and Kwon (2021) require a uniform equicontinuity assumption on the gradients of the moment conditions. For the sake of comparison, consider again the case of a finite number of unconditional moment inequalities.

Assumption 3.7 (Uniformly Equicontinuous Gradients)**.** *Each $E_P[m_k(W_i, \theta)]/\sigma_{P,k}(\theta)$ is differentiable in θ. Furthermore, if $D_P(\theta)$ denotes the $K \times K$ diagonal matrix with $\sigma_{P,k}(\theta)$ in the k^{th} position, and $\nabla_\theta D_P^{-1}(\theta) E_P[m(W_i, \theta)]$ denotes the $K \times d_\theta$ Jacobian matrix for the standardized moment conditions, then the following holds:*

$$\lim_{\delta \to 0} \sup_{P \in \mathcal{P}_0} \sup_{(\theta, \theta'): ||\theta - \theta'|| \leq \delta} ||\nabla_\theta D_P^{-1}(\theta) E_P[m(W_i, \theta)] - \nabla_\theta D_P^{-1}(\theta') E_P[m(W_i, \theta')]|| = 0.$$

Requiring uniformly equicontinuous gradients can be quite strong; for instance, this condition is strictly stronger than uniform continuity of the map $\theta \mapsto \nabla_\theta \sigma_{P,k}^{-1}(\theta) E_P[m_k(W_i, \theta)]$. A possible sufficient condition is that each function $\theta \mapsto \nabla_\theta \sigma_{P,k}^{-1}(\theta) E_P[m_k(W_i, \theta)]$ is Lipschitz continuous with a common (i.e. for all P and k) Lipschitz constant. This holds, for instance, if the functions $\theta \mapsto \nabla_\theta \sigma_{P,k}^{-1}(\theta) E_P[m_k(W_i, \theta)]$ are everywhere differentiable with a (P, k)-uniformly bounded derivative. With an uncountable number of conditional moment inequalities this requires $\theta \mapsto \nabla_\theta \varsigma_P^{-1}(\theta, \tau, g) E_P[m_k(W_i, \theta, \tau, g)]$ be everywhere differentiable with a (P, τ, g)-uniformly bounded derivative.

For a practical example of when Assumption 3.7 can fail, consider again the moment conditions arising from the support function estimator in Example 1:

$$E_P\left[\tau^\top g(W_i) - E_P\left[\sup_{q\in Q_\theta(W_i,\varepsilon_i)} \tau^\top q \;\middle|\; W_i\right] \;\middle|\; X_i\right] \leq 0 \text{ a.s. } \forall \tau \in \mathbb{B}^{d_\tau}.$$

Here the parameter vector $\theta \in \Theta$ enters only through the random set $Q_\theta(W_i, \varepsilon_i)$, so that the existence of a smooth derivative for each of the implied unconditional moments can be challenging to establish unless the random set is sufficiently simple.

3.5 Power Comparisons with Previous Methods

In this section we compare the power of the SSMT procedure with previous methods. To limit the scope, we focus on a comparison with an approach based on subsampling (e.g. Politis, Romano, and Wolf (1999), Romano and Shaikh (2008)), and the approach of Bugni, Canay, and Shi (2015). However, similar comments apply to by-product tests using the procedures of Andrews and Soares (2010), Andrews and Shi (2013) and Andrews and Shi (2017), and others (see Remark 3.4).

3.5.1 Comparison with Subsampling

The sample-splitting aspect of the SSMT procedure means that it shares some superficial similarities to subsampling. In this section we briefly introduce a subsampling approach to the specification testing problem, and discuss the differences with the SSMT procedure. Consider the test statistic:

$$T_n^{ss} := \inf_{\theta\in\Theta} T_n^{ss}(\theta), \text{ where } T_n^{ss}(\theta) := \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{\sqrt{n}\bar{m}_n(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)}. \tag{3.4}$$

Following Romano and Shaikh (2008), subsampling proceeds by selecting a subsample size $q_n \ll n$ satisfying $q_n/n \to 0$ and $q_n \to \infty$. Now let $N = \binom{n}{q_n}$, let $\{S_k\}_{k=1}^N$ denote all possible subsets of $\{W_i\}_{i=1}^n$ of size q_n, and define the subsampled test statistic:

$$T_{q_n,k}^{ss\sharp} = \inf_{\theta\in\Theta} T_{q_n,k}^{ss\sharp}(\theta), \text{ where } T_{q_n,k}^{ss\sharp}(\theta) := \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{\sqrt{q_n}\bar{m}_{q_n}^{ss\sharp}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)}, \tag{3.5}$$

where $\bar{m}_{q_n,k}^{ss\sharp}(\theta,\tau,g)$ is the sample average moment computed using only the observations in the subsample S_k. The subsampling critical value is the $1-\alpha$ quantile of the subsampling distribution of $T_{q_n,k}^{ss\sharp}$.[25] Thus, subsampling uses the full sample to compute the test statistic (3.4), and uses subsamples to compute the critical value. In contrast, the SSMT procedure uses the full sample to compute $\hat{\theta}_n$, uses subsamples to compute the sub-test statistics $T_{q_n}^{(r)}$ and the sub-test critical values $c_{q_n}^{(r)\sharp}(1-\alpha/\rho_n+\eta)$, and then aggregates the results from the multiple tests.

[25] Following Politis, Romano, and Wolf (1999) and Remark 3.2 in Romano and Shaikh (2008), the subsampling critical value can also be taken as the $1-\alpha$ quantile of the subsampling distribution of $\sqrt{q_n}(T_{q_n,k}^{ss\sharp}/\sqrt{q_n} - T_n^{ss}/\sqrt{n})$.

Remark 3.3. *To our knowledge, primitive conditions for the validity of the subsampling procedure just described have never been established, and it is unclear whether this subsampling procedure requires a polynomial minorant condition (or another similar condition). In particular, the validity of subsampling established in Theorem 3.4 of Romano and Shaikh (2008) relies on their high-level condition* (38)*. Bugni, Canay, and Shi (2017) (Remark 4.3) conjecture that the existence of a polynomial minorant may be required to satisfy this condition.*

Unlike in the SSMT procedure, the vector $\hat{\theta}_n$ cannot be recycled when subsampling; that is, it is not possible to use $T^{ss\sharp}_{q_n,k}(\hat{\theta}_n)$ in place of $T^{ss\sharp}_{q_n,k}$. This makes subsampling much more computationally burdensome for the applications we have in mind, since the infimum in (3.5) must be recomputed for every subsample. To see why subsampling fails when recycling $\hat{\theta}_n$, suppose for simplicity that $\mathcal{T}$ and $\mathcal{G}_n$ both contain a single element, suppose that $\hat{\varsigma}(\theta,\tau,g) = 1$, and suppose that $\hat{\theta}_n$ obtains the infimum in (3.4). Then:

$$T^{ss}_n(\hat{\theta}_n) = \underbrace{\sqrt{n}(\bar{m}_n(\hat{\theta}_n,\tau,g) - E_P[m(W_i,\hat{\theta}_n,\tau,g)])}_{\text{Empirical Process}} + \underbrace{\sqrt{n}E_P[m(W_i,\hat{\theta}_n,\tau,g)]}_{\text{Recentering}}, \tag{3.6}$$

$$T^{ss\sharp}_{q_n,k}(\hat{\theta}_n) = \underbrace{\sqrt{q_n}(\bar{m}^{ss\sharp}_{q_n,k}(\hat{\theta}_n,\tau,g) - E_P[m(W_i,\hat{\theta}_n,\tau,g)])}_{\text{Empirical Process}} + \underbrace{\sqrt{q_n}E_P[m(W_i,\hat{\theta}_n,\tau,g)]}_{\text{Recentering}}.$$

Now suppose that instead of using $T^{ss\sharp}_{q_n,k}$, the subsampling critical value is based on the distribution of $T^{ss\sharp}_{q_n,k}(\hat{\theta}_n)$. Then:

$$T^{ss\sharp}_{q_n,k}(\hat{\theta}_n) = \sqrt{q_n}\left(\bar{m}_{q_n,k}(\hat{\theta}_n,\tau,g) - E_P[m(W_i,\hat{\theta}_n,\tau,g)]\right) + o_{\mathcal{P}_0}(1), \tag{3.7}$$

where $o_{\mathcal{P}_0}(1)$ represents a term that converges in probability to zero, uniformly over $\mathcal{P}_0$. In this case we see that the distribution of (3.6) asymptotically dominates the distribution of (3.7) whenever $\sqrt{n}E_P[m(W_i,\hat{\theta}_n,\tau,g)] \to c > 0$, which is possible under the null.[26] In these cases, the critical values based on the distribution of (3.7) will be asymptotically "too small," leading to over-rejection.

However, although subsampling is more computationally demanding, it has power against some local alternatives for which the SSMT procedure has no (non-trivial) local power. This can be illustrated using a simple example. Consider a model with a single moment inequality of the form $E_{P_n}[m(W_i,\theta,\tau,g)] = E_{P_n}[W_i] \leq 0$, where $E_{P_n}[W_i] = c/\sqrt{n}$. Under some mild assumptions, straightforward calculation shows, for any $\varepsilon > 0$, there exists a value of $c > 0$ large enough such that a test based on subsampling has rejection probability exceeding $1-\varepsilon$. Furthermore, if $E_{P_n}[W_i] = c/\sqrt{m_n}$ for $m_n/n \to 0$, then a test based on subsampling has rejection probability tending to 1. This should be contrasted with Theorem 3.3, which suggests that the SSMT procedure has no (non-trivial) local power against any such sequences whenever $m_n/q_n \to \infty$.

[26] See Section S.4.2 of the Online Supplementary Material for an example.

3.5.2 Comparison with Bugni, Canay, and Shi (2015)

Here we briefly compare power with the approach of Bugni, Canay, and Shi (2015). To place our test on equal footing, we focus on the case with a finite number of unconditional moment inequalities. Consider the test statistic:

$$T_n^{bcs} := \inf_{\theta\in\Theta} \max_{1\le k\le K} \left[\frac{\sqrt{n}\bar{m}_{n,k}(\theta)}{\hat{\sigma}_{n,k}(\theta)}\right]_+, \tag{3.8}$$

where $[\cdot]_+ = \max\{0, \cdot\}$. Furthermore, consider the bootstrap test statistic:

$$T_n^{bcs\sharp} := \inf_{\theta\in\hat{\Theta}_I} \max_{1\le k\le K} \left[\frac{\sqrt{n}(\bar{m}_{n,k}^{\sharp}(\theta) - \bar{m}_{n,k}(\theta))}{\hat{\sigma}_{n,k}(\theta)} + \varphi\left(\frac{\kappa_n^{-1}\sqrt{n}\bar{m}_{n,k}(\theta)}{\hat{\sigma}_{n,k}(\theta)}\right)\right]_+, \tag{3.9}$$

where $\bar{m}_{n,k}^{\sharp}(\theta)$ is the sample average moment computed on the bootstrap sample, $\hat{\Theta}_I := \arg\min T_n^{bcs}(\theta)$, κ_n is a sequence satisfying $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$, and φ is a GMS function, satisfying the assumptions in Andrews and Soares (2010). Examples of φ include $\varphi(x) = -\infty \cdot 1\{x < -1\}$ (where $-\infty \cdot 0 = 0$), $\varphi(x) = \min\{x, 0\}$, and $\varphi(x) = x$. Test RS ("Re-Sampling") in Bugni, Canay, and Shi (2015) rejects when T_n exceeds the $1-\alpha$ quantile of the bootstrap distribution of $T_n^{bcs\sharp}$.

It is straightforward to show that, as in the SSMT procedure (but unlike subsampling), the vector $\hat{\theta}_n$ can be recycled when computing the bootstrap test statistic $T_n^{\sharp}$ in (3.9). Recycling $\hat{\theta}_n$ improves the computational tractability of the procedure of Bugni, Canay, and Shi (2015), but potentially at the cost of power. However, size control for the procedure of Bugni, Canay, and Shi (2015) relies on their high-level assumption A.6, which they show is implied by a polynomial minorant condition (e.g. see Assumption 3.6) and an equicontinuity condition on the gradients of the moments (e.g. see Assumption 3.7). In Section S.4.2 of the Online Supplementary Material we present an example where the polynomial minorant condition fails, and where the method of Bugni, Canay, and Shi (2015) over-rejects under the null with or without GMS.

However, simple examples also show that the procedure of Bugni, Canay, and Shi (2015) has power against some local alternatives for which the SSMT procedure has no (non-trivial) local power. Indeed, in the same example from the comparison with subsampling where $E_{P_n}[m(W_i, \theta, \tau, g)] = E_{P_n}[W_i] = c/\sqrt{m_n}$ for $c > 0$ and $m_n/n \to 0$, the approach of Bugni, Canay, and Shi (2015) has power tending to 1 for any GMS function. In contrast, Theorem 3.3 shows that if $m_n/q_n \to \infty$ the SSMT procedure has no (non-trivial) local power. Furthermore, the use of GMS in Bugni, Canay, and Shi (2015) means their test is insensitive to moment conditions that are very "slack" at $\hat{\theta}_n$, whereas these slack moments can affect the critical value in the SSMT procedure and reduce power in finite sample.

Remark 3.4. *By-product specification tests reject when the null hypothesis $H_0 : \theta \in \Theta_I$ is rejected for every $\theta \in \Theta$. The latter null can be tested in various models using methods proposed by Andrews*

and Soares (2010), Andrews and Shi (2013) and Andrews and Shi (2017), among many others. Following an argument analogous to the one presented above, these by-product tests also have power against some local-alternatives for which the SSMT procedure has no non-trivial local power.

3.6 Tuning Parameter Selection

The main tuning parameters for the SSMT procedure are q_n, r_n, η, and ρ_n, although q_n and ρ_n are the most important. While any choice of these tuning parameters satisfying Assumption 3.3 are asymptotically valid, their choices may make a difference in finite samples. To summarize our recommendations, we suggest setting $q_n = \lfloor n^{4/5} \rfloor$, $r_n = \lfloor n^{1/5} \rfloor$, $\eta = 10^{-6}$, and ρ_n such that $r_n/\rho_n \in \mathbb{N} \setminus \{1\}$. These recommendations are motivated by the theoretical considerations as well as both reported and unreported simulation results. In addition to the simulation exercises presented in the next section, some additional simulation exploration of various values for the tuning parameters can be found in Section S.4 of the Online Supplementary Material.

For q_n, we focus on choices that are polynomial in n. When q_n is large, the finite sample power properties tend to be improved, but the test may over-reject under the null in small samples. On the other hand, setting q_n too small can lead to under-rejection under alternatives that are close to the null. In practice, we found that setting $q_n = \lfloor n^{4/5} \rfloor$ balances these two concerns, as is illustrated in the simulations in the next section and in the simulations in Section S.4 of the Online Supplementary Material. We recommend that the parameter r_n then be determined directly from the choice of q_n to ensure that $q_n \cdot r_n \approx n$. However, other methods of choosing q_n are also possible. For instance, despite the differences between the proposed approach and an approach based on subsampling, the problem of choosing the tuning parameter q_n shares some similarities to the problem of choosing the subsample size. General methods for selecting the subsample size have been developed (see Politis, Romano, and Wolf (1999) Chapter 9) and these methods are also applicable to the choice of q_n.

The parameter η is exactly the infinitesimal uniformity factor from Andrews and Shi (2013) and Andrews and Shi (2017), which is required to avoid certain high level assumptions on the asymptotic distribution of the test statistics.[27] The parameter η is both added to the critical value—which is required near the end of the proofs of Lemmas B.3 and B.5—and added to the confidence level—which is required at the end of the proof of Lemma B.5.[28] We experiment with different values of η throughout the simulation results in the next section, and in Section S.4 of the Online Supplementary Material. However, while η is useful in the main proofs, the simulation exercises suggest that it plays a minor role in practice. This is consistent with Andrews and Shi

[27] See Andrews and Shi (2013) p. 625, or Andrews and Shi (2017) p. 281 and footnote 20.

[28] This constant is not required in Bugni, Canay, and Shi (2015), although they make additional assumptions on the limiting distribution of the test statistic (see Bugni, Canay, and Shi (2015) Assumption A.7 and Remark B.2).

(2013), who suggest it can be set to 10^{-6}, and Andrews and Shi (2017), who suggest it can be set to zero in applications. We recommend $\eta = 10^{-6}$ for concreteness, although any value $\eta \leq 10^{-3}$ gives similar results.

For ρ_n, we find that different values have strong implications for finite sample power, although simulation evidence suggests that the value of ρ_n becomes less important in large samples (see Figure 4 in the Online Supplementary Material). Regardless of the sample size, some values of ρ_n can be shown to unambiguously dominate the others. First, ρ_n should never be chosen outside the interval $[1, r_n]$. In particular, the SSMT procedure will never reject when $\rho_n < 1$. When $\rho_n > r_n$ the test only rejects if at least one sub-test rejects at level α/ρ_n. But this test is dominated by the test with $\rho_n' = r_n$, which only rejects if at least one sub-test rejects at level $\alpha/\rho_n' > \alpha/\rho_n$. Furthermore, in order to maximize power ρ_n should always be set so that r_n/ρ_n is an integer. To see why, note that when $r_n/\rho_n = m \in \mathbb{N}$ the SSMT procedure rejects only if exactly m sub-tests reject at level α/ρ_n. For slightly larger values $\rho_n' > \rho_n$ satisfying $\rho_n' \cdot (m-1) < r_n < \rho_n' \cdot m$, the test still rejects only if m sub-tests reject, but this time at level $\alpha/\rho_n' < \alpha/\rho_n$. Since the required number of sub-test rejections is the same, but the nominal level of each test has declined, power is unnecessarily reduced using ρ_n' relative to ρ_n.

Since ρ_n is required to be a decreasing sequence, and r_n is an increasing sequence, the choice of $\rho_n = r_n$ is not possible asymptotically. Furthermore, both the theoretical results (namely, the proof of Theorem 3.1) and the simulation results (both examples in Section S.4.1 of the Online Supplementary Material) warn against this choice. Thus, we never recommend setting $\rho_n = r_n$. However, there is a discontinuity in the rejection probability of the SSMT procedure at $\rho_n = r_n$, and larger values of ρ_n that may be close to (but are still strictly below) r_n may still work well in terms of both size control and power.

Combining everything, we recommend setting ρ_n so that $r_n/\rho_n \in \mathbb{N}\backslash\{1\}$. Note these restrictions already substantially limit the possible values for ρ_n; for instance, if $n = 2000$, then using the recommended values for q_n and r_n we have $q_n = 437$ and $r_n = 4$, and our recommendations for ρ_n are $\rho_n \in \{1, 1.3\bar{3}, 2\}$. Since ρ_n changes only the quantile for the critical value and the aggregation of the sub-tests, the SSMT procedure can be run for multiple values of ρ_n at no additional computational cost. Thus, researchers who prioritize transparency can easily report the outcome of the SSMT procedure for all recommended values $r_n/\rho_n \in \mathbb{N} \setminus \{1\}$.

3.7 Computational Details

To use the SSMT procedure, the researcher must obtain an approximate minimizer $\hat{\theta}_n$ in the optimization problem:

$$\inf_{\theta \in \Theta} \sup_{\tau \in \mathcal{T}} \max_{g \in \mathcal{G}_n} \frac{\sqrt{n}\bar{m}_n(\theta, \tau, g)}{\hat{\varsigma}_n(\theta, \tau, g)}.$$

Once an approximate minimizer $\hat{\theta}_n$ is obtained, the researcher must repeatedly bootstrap the quantity:

$$T_{q_n}^{(r)\sharp}(\hat{\theta}_n) = \sup_{\tau \in \mathcal{T}} \max_{g \in \mathcal{G}_n} \left(\frac{\sqrt{q_n}(\bar{m}_{q_n}^{(r)\sharp}(\hat{\theta}_n, \tau, g) - \bar{m}_n(\hat{\theta}_n, \tau, g))}{\hat{\varsigma}_{q_n}^{(r)}(\hat{\theta}_n, \tau, g)} \right). \tag{3.10}$$

Beginning with the test statistic, we can break the problem into two stages, denoted as the *inner problem* and the *outer problem*. The inner problem computes the profiled test statistic:

$$T_n(\theta) := \sup_{\tau \in \mathcal{T}} \max_{g \in \mathcal{G}_n} \frac{\sqrt{n}\bar{m}_n(\theta, \tau, g)}{\hat{\varsigma}_n(\theta, \tau, g)}. \tag{3.11}$$

The outer problem then computes an approximate infimum of the profiled test statistic over $\theta \in \Theta$. In many examples, the inner problem can be efficiently computed for a fixed value $\theta \in \Theta$. This is the case in models with a finite number of moment conditions (in which case (3.11) is simply the maximum over a finite number of elements), or in models where the moment functions are concave in the index parameter $\tau \in \mathcal{T}$. The latter occurs, for example, when using moment conditions implied by support function characterizations of the identified set (see Example 1). In some cases, concavity of the moment conditions can be directly exploited by specialized modeling software like `CVX` for Matlab or `CVXR` for R to efficiently compute (3.11). For more complicated moment functions, concavity can still be exploited by super-gradient ascent methods provided that the researcher can provide a single super-gradient of the objective function with respect to τ. These methods are applicable when the moment functions are not differentiable, do not require second-order information on the objective function, and are guaranteed to converge to a global optimum given a sufficiently large number of iterations. In the simulation section, we try a stochastic super-gradient ascent algorithm based on the algorithm of Nemirovski, Juditsky, Lan, and Shapiro (2009), and find that it is fast and reliable.

The outer problem (i.e. the minimization of (3.11) over $\theta \in \Theta$) is generally more difficult, even when the inner problem (3.11) is relatively easy to solve. In particular, the value function of the inner problem can easily fail to be differentiable in $\theta \in \Theta$, even when each moment function is infinitely differentiable in $\theta \in \Theta$. When the inner problem is inexpensive, the outer problem

can be solved by heuristic methods like simulated annealing, or differential evolution.[29] When the inner problem is expensive to solve, we find that Bayesian optimization algorithms work well. Our own implementation is based on the efficient global optimization (EGO) algorithm popularized by Jones, Schonlau, and Welch (1998) and recently used by Kaido, Molinari, and Stoye (2019).[30]

Turning to the construction of the critical values, the difficulty of computing (3.10) depends on how the index parameter τ enters the moment functions. Computation is straightforward if $\mathcal{T}$ is finite. Otherwise, some simplifications also arise if the moment functions are concave in τ, which is the case when using support function characterizations of the identified set. In this case, (3.10) requires solving an optimization problem involving the difference of two concave functions. The later can be reformulated as a minimization problem involving the difference of two convex functions, which is a well-studied problem in convex analysis (see An and Tao (2005), Le Thi and Pham Dinh (2018), Le Thi, Le, Phan, and Tran (2020)).

4 Simulation Examples

In this section we provide simulation results for three examples: (i) a static game of complete information, (ii) a random coefficient binary choice model with endogenous regressors, and (iii) a simple model with moment inequalities that are linear in parameters. For each example the design of the DGPs is similar: by adjusting a single parameter, we are able to construct DGPs that are partially identified, point-identified (or on the "boundary" between the null and alternative hypothesis), and misspecified.

The main objective of the simulation exercises is to investigate the size control and power properties of the SSMT procedure under various DGPs, to compare the procedure to existing methods, and to illustrate the theoretical properties discussed in the previous sections. We investigate the sensitivity of the procedure to the subsample size by considering $r_n = \lfloor n^{1/r} \rfloor$ and $q_n = \lfloor n/r_n \rfloor \approx \lfloor n^{(r-1)/r} \rfloor$ for $r \in \{4, 5, 6\}$. We also investigate the impact of ρ_n, and report results for $\rho_n \in \{1, 2, \ldots, r_n\}$.[31] For examples (ii) and (iii), we compare the SSMT procedure to the RS test of Bugni, Canay, and Shi (2015), the by-product test using the method from Andrews and Shi (2017), and the by-product test based on a least favorable critical values, which uses the method of Andrews and Shi (2017) without GMS. We provide no comparison with previous methods for example (i) since the method of Bugni, Canay, and Shi (2015) does not apply, and the by-product tests are computationally intractable. For all examples we use $B = 999$ bootstrap samples and we compute the rejection probabilities across 500 replications.

[29] We find the latter works well in our simulation examples, and use the `DEOptim` package in R by Mullen, Ardia, Gil, Windover, and Cline (2011).

[30] See also Jones (2001) for a review of response-surface optimization methods.

[31] Note that some of these values (intentionally) do not match the recommended values from Section 3.6.

4.1 Static Game of Complete Information Games

Here we apply the method to a static game of complete information with pure strategy Nash equilibria, using moment conditions derived from the support function characterization of the identified set in Beresteanu, Molchanov, and Molinari (2011). In particular, we consider a two player game and let $Y_{i1}, Y_{i2} \in \{0,1\}$ denote the binary decisions of the two players. Players make the choice that maximizes their payoff, and we assume the payoff function for player j in the i^{th} game is given by:

$$\pi_j(Y_{ij}, Y_{i(-j)}, X_{ij}, \varepsilon_{ij}, \theta) := Y_{ij}(Y_{i(-j)}\delta_j + X_{ij}\beta + \varepsilon_{ij}),$$

where $-j$ refers to player j's competitor, and $\theta = (\delta_1, \delta_2, \beta)$. In all DGPs, $X_{i1}, X_{i2} \sim$ Uniform$[1/3, 2/3]$ and $\varepsilon_{i1}, \varepsilon_{i2} \sim$ Uniform$[0,1]$ for $i = 1, \ldots, n$. Let $\mathcal{Y} = \{(y_1^k, y_2^k)\}_{k=1}^4$, where:

$$(y_1^1, y_2^1) = (0,0), \qquad (y_1^3, y_2^3) = (0,1),$$
$$(y_1^2, y_2^2) = (1,0), \qquad (y_1^4, y_2^4) = (1,1).$$

Following Beresteanu, Molchanov, and Molinari (2011), the identified set of structural parameters is the collection of vectors $\theta \in \Theta$ that satisfy the moment inequalities:

$$E_P[m(W_i, \theta, \tau, g)] \leq 0, \qquad \forall (\tau, g) \in \mathcal{T} \times \mathcal{G},$$

where $\mathcal{T} := \left\{\tau \in \mathbb{R}^3 : ||\tau|| \leq 1\right\}$ is the unit ball, $W_i = (Y_{i1}, Y_{i2}, X_{i1}, X_{i2})$, and:

$$m(W_i, \theta, \tau, g) := \left(\sum_{k=1}^{3} \tau_k 1\{(Y_{i1}, Y_{i2}) = (y_1^k, y_2^k)\} - E_P\left[\max_{\sigma \in S_\theta(X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})} \langle \tau, \sigma \rangle \,\middle|\, X_{i1}, X_{i2} \right] \right) g(X_{i1}, X_{i2}).$$

Here $S_\theta(X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})$ is a random set described in Section S.3.1 of the Online Supplementary Material which contains the set of pure strategy Nash equilibria for a given realization of the vector $(X_{i1}, X_{i2}, \varepsilon_{i1}, \varepsilon_{i2})$ at a fixed vector of structural parameters. In the simulation exercises, we impose the parameter space constraints $\delta_1, \delta_2 \leq 0$ when computing our test statistic, and we use $\mathcal{G}_n \subset \mathcal{G}$ with $\mathcal{G}_n = \{g_1, g_2, g_3, g_4\}$, where:

$$g_1(\boldsymbol{x}) = \mathbb{1}\{\boldsymbol{x} \in (0, 0.5]^2\}, \qquad g_3(\boldsymbol{x}) = \mathbb{1}\{\boldsymbol{x} \in (0.5, 1] \times (0, 0.5]\},$$
$$g_2(\boldsymbol{x}) = \mathbb{1}\{\boldsymbol{x} \in (0, 0.5] \times (0.5, 1]\}, \qquad g_4(\boldsymbol{x}) = \mathbb{1}\{\boldsymbol{x} \in (0.5, 1]^2\}.$$

Notice that any rescaling of the moment functions is equivalent to a rescaling of $\mathcal{T}$, so that the scale of the moment functions in this example is implicitly determined by the fact that $\mathcal{T}$ is the unit ball. Thus, for computational simplicity we set $\hat{\varsigma}_n(\theta, \tau, g) = 1$ for all (τ, g).

We then consider four DGPs. The first DGP is correctly specified and partially identified.

This DGP is denoted by GameDGP$_1$, where we set $(\delta_1, \delta_2, \beta) = (-0.2, -0.1, -0.6)$. The second DGP is correctly specified and point identified. This DGP is denoted as GameDGP$_2$, where we set $(\delta_1, \delta_2, \beta) = (-0.2, 0, -0.6)$. Finally, we consider two "alternative" DGPs where the model is misspecified. The first alternative DGP is denoted as GameDGP$_3$ with $(\delta_1, \delta_2, \beta) = (0.1, 0, -0.6)$, and the second alternative DGP is denoted as GameDGP$_4$ with $(\delta_1, \delta_2, \beta) = (0.2, 0, -0.6)$.

In Section S.3.1 of the Online Supplementary Material we prove that the identified set is a singleton under GameDGP$_2$, and is empty under both GameDGP$_3$ and GameDGP$_4$. In particular, Proposition S.3.1(i) shows that θ is point-identified whenever $\delta_2 = 0$, regardless of the sign of δ_1. Applying this result, if the (point-identified) data-generating vector θ_0 has $\delta_1 > 0$ and $\delta_2 = 0$, then there can be no vector $\theta \in \Theta$ with $\delta_1, \delta_2 \leq 0$ that satisfies all moment conditions. This ensures that, for any choice probabilities generated with $\delta_1 > 0$ and $\delta_2 = 0$ (as in GameDGP$_3$ and GameDGP$_4$), the moment conditions are violated for all $\theta \in \Theta$ with $\delta_1, \delta_2 \leq 0$. Furthermore, among the alternative DGPs, the violations of the moment conditions are more severe under GameDGP$_4$ than GameDGP$_3$, so that we expect our test to have the highest power under GameDGP$_4$. For comparison, with $n = 10^5$ observations the minimum value of $n^{-1/2}T_n(\theta)$, a (simulated) measure of the level of misspecification, was -0.014 for GameDGP$_1$, 0.000 for GameDGP$_2$, 0.003 for GameDGP$_3$ and 0.021 for GameDGP$_4$

The results are displayed in Table 1. The results for GameDGP$_1$ show that the simulated rejection probability is below the nominal level for almost all sample sizes, all subsample sizes and all test functions. In ChoiceDPG$_2$ (correctly specified and point identified) the rejection probabilities are below or around the nominal level for all sample sizes and all values of ρ_n except when $\rho_n = r_n$. However, as discussed in Section 3.6, this choice of ρ_n is not theoretically valid and is not recommended, even in finite sample. The results for GameDGP$_3$ and GameDGP$_4$ show that the SSMT procedure has substantial power against fixed alternatives, with the power increasing rapidly with the subsample size.

We do not compare our approach with alternative methods for this example, since the test of Bugni, Canay, and Shi (2015) does not apply to settings with a continuum of moment inequalities, and by-product sets based on the procedure of Andrews and Shi (2017) are computationally intractable. In particular, the average time to evaluate $T_n(\theta)$ at a single parameter vector is about 9.93 seconds, owing mainly to the difficulty of computing the supremum over $\tau \in \mathcal{T}$.[32] By-product tests then require repeated evaluation of the test statistic over a grid in the parameter space, and it takes about $20^3 \times 9.93$ sec ≈ 22 hours to evaluate the test statistic for this example across a sparse grid of 20 points in each dimension of the parameter space with three parameters. This does not include the time required to compute the critical value at each point, which would require another 999 evaluations of the bootstrap test statistic at each point, or an additional $999 \times 20^3 \times 9.93$ sec $\approx 22,045$

[32] This is the average time across 1000 evaluations.

Table 1: SSMT rejection rates for the static game of complete information

	GameDPG_1			GameDPG_2			GameDPG_3			GameDPG_4		
	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$n = 500$												
$r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 1$	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.04	0.08	0.05	0.19	0.32
$r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 2$	0.00	0.00	0.01	0.00	0.01	0.01	0.06	0.18	0.29	0.17	0.42	0.58
$r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.07
$r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 2$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.01	0.08	0.15
$r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 3$	0.00	0.01	0.02	0.00	0.00	0.02	0.02	0.09	0.16	0.08	0.24	0.38
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 2$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.03	0.09
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 3$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.05
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 4$	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.04	0.09	0.03	0.11	0.19
$n = 2000$												
$r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 1$	0.00	0.00	0.01	0.01	0.02	0.03	0.22	0.41	0.52	0.51	0.80	0.88
$r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 2$	0.02	0.04	0.06	0.03	0.07	0.10	0.37	0.59	0.68	0.73	0.94	0.97
$r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 3$	0.06	0.11	0.14	0.10	0.19	0.25	0.59	0.76	0.82	0.91	0.97	0.99
$r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 1$	0.00	0.00	0.01	0.00	0.01	0.01	0.07	0.22	0.31	0.20	0.52	0.69
$r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 2$	0.00	0.02	0.03	0.02	0.06	0.08	0.27	0.53	0.61	0.60	0.89	0.95
$r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 3$	0.00	0.02	0.02	0.01	0.04	0.07	0.22	0.45	0.56	0.53	0.81	0.91
$r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 4$	0.03	0.05	0.07	0.05	0.12	0.16	0.45	0.65	0.72	0.77	0.94	0.98
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.07	0.03	0.13	0.22
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 2$	0.00	0.02	0.03	0.01	0.03	0.05	0.06	0.24	0.39	0.18	0.51	0.72
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 3$	0.01	0.03	0.05	0.02	0.05	0.08	0.11	0.32	0.45	0.30	0.64	0.83
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 4$	0.01	0.02	0.04	0.01	0.04	0.06	0.10	0.29	0.41	0.26	0.56	0.75
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 5$	0.01	0.02	0.04	0.01	0.03	0.05	0.10	0.24	0.37	0.23	0.51	0.67
$r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 6$	0.02	0.05	0.07	0.06	0.10	0.14	0.25	0.48	0.60	0.47	0.75	0.88

Notes: The values displayed in the table are the proportion of tests rejected across 500 experiments using the SSMT procedure. Both ChoiceDPG_1 and ChoiceDPG_2 are correctly specified: ChoiceDPG_1 is partially identified and ChoiceDPG_2 is point identified. Both models ChoiceDPG_3 and ChoiceDPG_4 are misspecified.

hours for each replication.

4.2 Random Coefficient Binary Choice with Endogenous Regressors

Next we apply the SSMT procedure to a random coefficient binary choice model. This model was studied in detail in Chesher and Rosen (2014), and was also used as an example in Andrews and Shi (2017). We use a similar DGP to these papers. In particular, we set:

$$Y_i = \mathbb{1}\{\beta_{i0} + \beta_{i1} X_i > 0\}, \tag{4.1}$$

where β_{i0} and β_{i1} are random coefficients, and X_i is a scalar covariate. The researcher assumes that X_i is endogenous (although this may not be the case), and has access to an instrumental variable Z_i. We assume:

$$\begin{bmatrix} \beta_{i0} \\ \beta_{i1} \\ X_i^* \end{bmatrix} \sim N\left(\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma_1 & \delta_1 \\ \gamma_1 & \gamma_2 + \gamma_1^2 & \delta_1 \\ \delta_1 & \delta_1 & 1 \end{pmatrix} \right).$$

We then set $Z_i^* = X_i^* + \delta_2 \xi_i$, where $\xi_i \sim N(0,1)$. Letting F_x denote the distribution of X_i^*, and F_z the distribution of Z_i^*, we then construct the covariate X_i and instrument Z_i as follows:

$$\begin{aligned} X_i = (-1)\mathbb{1}\{X_i^* \leq F_x^{-1}(0.25)\} &+ (0)\mathbb{1}\{F_x^{-1}(0.25) < X_i^* \leq F_x^{-1}(0.5)\} \\ &+ (1)\mathbb{1}\{F_x^{-1}(0.5) < X_i^* \leq F_x^{-1}(0.75)\} \\ &+ (2)\mathbb{1}\{F_x^{-1}(0.75) < X_i^*\}, \end{aligned}$$

$$\begin{aligned} Z_i = (-1)\mathbb{1}\{Z_i^* \leq F_z^{-1}(0.25)\} &+ (0)\mathbb{1}\{F_z^{-1}(0.25) < Z_i^* \leq F_z^{-1}(0.5)\} \\ &+ (1)\mathbb{1}\{F_z^{-1}(0.5) < Z_i^* \leq F_z^{-1}(0.75)\} \\ &+ (2)\mathbb{1}\{F_z^{-1}(0.75) < Z_i^*\}. \end{aligned}$$

The parameters to be estimated are $\theta := (\alpha_0, \alpha_1, \gamma_1, \gamma_2)$, which control the distribution of the random coefficients. We impose the parameter space constraints $\alpha_0 \in [-2, 0]$, $\alpha_1 \in [-2, 2]$, $\gamma_0 \in [-2, 2]$ and $\gamma_1 \in [0, 2]$. The parameter δ_1 controls the correlation between the covariate and instrument with the random coefficients. When $\delta_1 = 0$, both the covariate X_i and the instrument Z_i are independent of the random coefficients. In this case, Z_i is a valid instrument for X_i. When $\delta_2 = 0$ we have $Z_i = X_i$, so that Z_i is a perfect instrument and Chesher and Rosen (2014) prove that the parameter θ is point-identified.[33] When $\delta_2 > 0$, Z_i is a valid instrument, but is not perfectly correlated with X_i, so that the model is partially identified. Finally, when $\delta_2 = 0$ and $\delta_1 > 0$ we have $Z_i = X_i$, but both X_i and the instrument Z_i are not independent of the random coefficients. In this case, the model is misspecified and the identified set is empty.[34]

Denote the distribution of the random coefficients as F_θ, and define the set-valued mapping:

$$\mathcal{T}(x, y) := \text{cl}\left\{(\beta_0, \beta_1) \in \mathbb{R}^2 : y = \mathbb{1}\{\beta_0 + x\beta_1 > 0\}\right\}. \tag{4.2}$$

That is, $\mathcal{T}(x, y)$ delivers the set of all random coefficient values (β_0, β_1) that are consistent with the observed pair (x, y) through the binary choice model in (4.1). Note that each set $\mathcal{T}(x, y)$ is a closed halfspace through the origin in $\mathbb{R}^2$. Following Chesher and Rosen (2014), the identified set is given by:

$$\Theta_I := \{\theta \in \Theta : \forall S \in \mathcal{S}, F_\theta(S) \geq P(\mathcal{T}(X, Y) \subseteq S \mid Z = z), P_Z - a.s.\}.$$

[33] See Appendix C of Chesher and Rosen (2014).
[34] See the Online Supplementary Appendix S.3.3 for a proof.

Here $\mathcal{S}$ is a collection of test sets:

$$\mathcal{S} := \{\mathcal{T}(x_1, y_1) \cup \mathcal{T}(x_2, y_2) : x_1, x_2 \in \mathcal{X}, y_1, y_2 \in \{0, 1\}\},$$

where $\mathcal{X}$ is the support of X_i. Notice that $\mathcal{S}$ contains all pairwise unions of halfspaces of the form (4.2). Indexing the sets in $\mathcal{S}$ as $S_1, \ldots, S_K$, the random coefficient binary choice model with endogenous regressors has a non-empty identified set if and only if:

$$\inf_{\theta \in \Theta} \max_{k=1,\ldots,K} \sup_{g \in \mathcal{G}} E_P[m_k(W_i, \theta, g)] \leq 0,$$

where $\mathcal{G}$ is a sufficiently rich set of instrument functions, and:

$$m_k(W_i, \theta, g) = (\mathbb{1}\{\mathcal{T}(X_i, Y_i) \subseteq S_k\} - F_\theta(S_k))\, g(Z_i).$$

Since Z_i is discrete, the instrument functions $g : \mathcal{Z} \to [0, 1]$ are taken as indicator functions $g(Z_i) = \mathbb{1}\{Z_i = z_k\}$ for $z_k \in \{-1, 0, 1, 2\}$. Details on how to compute $F_\theta(S_k)$ are provided in Chesher and Rosen (2014). In this simulation exercise the moment conditions have a natural scale, being the difference of two probabilities. Thus, we set $\hat{\varsigma}_n(\theta, \tau, g) = 1$ for all (τ, g). To generate the data, we set the true parameter vector as $\theta_0 := (0, -1, -1, 1)$. In this example, it is computationally inexpensive to determine if the moment conditions are violated at a given value of the parameter vector, so the infimum in the test statistic is computed using differential evolution.[35]

In our simulation exercises we consider four DGPs. The first DGP is "interior" to the null, where the model is partially identified. This DGP is denoted as ChoiceDGP$_1$, where we set $\delta_1 = 0$ and $\delta_2 = 0.5$. The second DGP is on the "boundary" of the null and alternative, where the model is point identified. We denote this DGP by ChoiceDGP$_2$ and set $\delta_1 = 0$ and $\delta_2 = 0$. The last two DGPs fall under the alternative. The first alternative DGP is denoted ChoiceDGP$_3$ with $\delta_1 = 0.2$ and $\delta_2 = 0$, and the second alternative DGP is denoted as ChoiceDGP$_4$ with $\delta_1 = 0.4$ and $\delta_2 = 0$. Of the two alternative DGPs, the violation of the moment conditions is most severe under ChoiceDGP$_4$, so we expect our test to reject the null most often under ChoiceDGP$_4$. For comparison, with $n = 10^5$ observations the minimum value of $n^{-1/2}T_n(\theta)$, a (simulated) measure of the magnitude of misspecification, was -0.039 for ChoiceDGP$_1$, 0.000 for ChoiceDGP$_2$, 0.026 for ChoiceDGP$_3$ and 0.084 for ChoiceDGP$_4$.

We also compare the SSMT procedure with three alternative approaches: the RS test of Bugni, Canay, and Shi (2015), a by-product test using Andrews and Shi (2017) with GMS, and a "least-favorable" by-product test using Andrews and Shi (2017) without GMS. When implementing the RS test from Bugni, Canay, and Shi (2015), we use the criterion function from (3.8), which is the

[35] Note $\gamma_2 \geq 0$ is necessary to ensure the covariance matrix for (β_0, β_1) is positive semi-definite.

criterion function that most closely matches our test statistic. We also use the GMS function:

$$\varphi_k^{bcs}(\theta, g) = -\infty \cdot \mathbb{1}\left\{\sqrt{n}\bar{m}_{n,k}(\theta, g) < -0.1\sqrt{\log(n)}\right\}, \tag{4.3}$$

which is the same one used in the simulations of Bugni, Canay, and Shi (2015).[36] In our implementation we keep $\hat{\theta}_n$ fixed during the bootstrap procedure rather than reoptimizing the test statistic in each bootstrap sample. This is done for computational tractability, although it means that the simulation results likely underestimate the rejection rates for their test.

When performing the by-product test using Andrews and Shi (2017), we use the same criterion function as the RS test, but use the GMS function:

$$\varphi_k^{as}(\theta, g) = -\sqrt{0.4\log(n)/\log(\log(n))}\mathbb{1}\left\{\sqrt{n}\bar{m}_n(\theta, \tau, g) < -\sqrt{0.3\log(n)}\right\}, \tag{4.4}$$

as recommended in Section 4 of Andrews and Shi (2017). Recall the by-product test rejects if the null hypothesis $H_0 : \theta \in \Theta_I(P)$ is rejected for every vector θ in a grid over the parameter space. For computational reasons, we use a coarse grid of 5 equally spaced points in each of the four dimensions of the parameter space (625 points in total). In particular, the average time to evaluate the test statistic $T_n(\theta)$ at a single parameter vector in this model is about 0.22 seconds, requiring about $5^4 \times 0.22$ sec ≈ 2.3 min to evaluate all test statistics and $5^4 \times 999 \times 0.22$ sec ≈ 38 hrs to evaluate all bootstrap test statistics for each replication.[37] Our use of a coarse grid means that the simulation results likely overestimate the rejection rates for the by-product test.

Finally, when implementing the least-favorable by-product test we use the same method as the previous by-product test, but with the GMS function set to 0 for all moment inequalities.

The results are displayed in Table 2 for sample sizes $n \in \{500, 2000\}$. For ChoiceDPG$_1$ and ChoiceDPG$_2$, we see that the rejection frequencies are smaller than the nominal level for all sample sizes, all subsample sizes, and all methods. In ChoiceDPG$_1$ the model is partially identified, and none of the tests ever reject. For ChoiceDPG$_2$, the model is point identified, but the rejection probabilities for the SSMT method are almost always below the rejection probabilities of the other methods, which are closer to nominal. In ChoiceDPG$_3$, the SSMT procedure and the RS test have similar power, but both have lower power than the by-product tests. Finally, all tests have substantially larger power in ChoiceDPG$_4$.

4.3 Simple moment conditions linear in parameters

Our final example implements the SSMT procedure on a simple example from Bugni, Canay, and Shi (2015). The simplicity of the final example allows us to obtain a closed-form expression for

[36] See Section 7 of Bugni, Canay, and Shi (2015).

[37] Average times are reported across 1000 evaluations. Our implementation of this example follows the instructions in Chesher and Rosen (2014).

Table 2: Rejection rates for the random coefficient discrete choice model.

	ChoiceDPG$_1$			ChoiceDPG$_2$			ChoiceDPG$_3$			ChoiceDPG$_4$		
Methods	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
						$n = 500$						
SSMT: $r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 1$	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.01	0.07	0.15
SSMT: $r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 2$	0.00	0.00	0.00	0.01	0.03	0.05	0.00	0.02	0.06	0.06	0.20	0.37
SSMT: $r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03
SSMT: $r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 2$	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.09
SSMT: $r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 3$	0.00	0.00	0.00	0.01	0.03	0.06	0.00	0.04	0.08	0.05	0.17	0.28
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 2$	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.01	0.00	0.03	0.09
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 3$	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.02	0.06
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 4$	0.00	0.00	0.00	0.01	0.03	0.05	0.01	0.04	0.07	0.03	0.16	0.28
BCS2015	0.00	0.00	0.00	0.02	0.04	0.05	0.00	0.02	0.02	0.10	0.35	0.53
AS2017	0.00	0.00	0.00	0.01	0.04	0.07	0.00	0.04	0.14	0.52	0.84	0.94
Least Favorable	0.00	0.00	0.00	0.00	0.03	0.05	0.00	0.02	0.08	0.42	0.77	0.90
						$n = 2000$						
SSMT: $r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.52	0.69
SSMT: $r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 2$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.71	0.85
SSMT: $r_n = \lfloor n^{1/6} \rfloor$, $\rho_n = 3$	0.00	0.00	0.00	0.01	0.02	0.04	0.02	0.06	0.11	0.56	0.86	0.94
SSMT: $r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.18	0.38
SSMT: $r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 2$	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.03	0.22	0.64	0.82
SSMT: $r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 3$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.15	0.53	0.71
SSMT: $r_n = \lfloor n^{1/5} \rfloor$, $\rho_n = 4$	0.00	0.00	0.00	0.00	0.02	0.04	0.01	0.07	0.14	0.47	0.78	0.89
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 2$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.27	0.54
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 3$	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.03	0.09	0.42	0.66
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 4$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.33	0.54
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 5$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.28	0.47
SSMT: $r_n = \lfloor n^{1/4} \rfloor$, $\rho_n = 6$	0.00	0.00	0.00	0.01	0.03	0.06	0.01	0.04	0.11	0.28	0.60	0.76
BCS2015	0.00	0.00	0.00	0.00	0.02	0.03	0.01	0.05	0.12	0.94	1.00	1.00
AS2017	0.00	0.00	0.00	0.01	0.04	0.09	0.20	0.53	0.68	1.00	1.00	1.00
Least Favorable	0.00	0.00	0.00	0.00	0.03	0.07	0.15	0.47	0.62	1.00	1.00	1.00

Notes: The values displayed in the table are the proportion of tests rejected across 500 experiments. Both ChoiceDPG$_1$ and ChoiceDPG$_2$ are correctly specified: ChoiceDPG$_1$ is partially identified and ChoiceDPG$_2$ is point identified. Both ChoiceDPG$_3$ and ChoiceDPG$_4$ are misspecified. BCS2015 refers to the RS test of Bugni, Canay, and Shi (2015) using the GMS function from (4.3). AS2017 refers to the by-product test using Andrews and Shi (2017) with the GMS function from (4.4), implemented over a uniform grid of 625 points from the four-dimensional parameter space. "Least Favorable" refers to a least favorable procedure similar to AS2017 but with the GMS function set to 0.

the identified set, which in turn allows us to better understand the magnitude of misspecification under each alternative DGP.

Let $W_i = (X_{i1}, X_{i2}, X_{i3}) \in \mathbb{R}^3$ be such that $W_i \sim N(\mu_n, \Sigma)$ where $\Sigma = I_{3\times 3}$ and $\mu_n = (0, -\zeta_n, 0)$,

Table 3: Rejection rates for the model with linear moment conditions.

	LinearDGP_1			LinearDGP_2			LinearDGP_3			LinearDGP_4			LinearDGP_5		
Methods	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
SSMT: $\rho_n = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.16	0.32	0.58	0.84	0.94
SSMT: $\rho_n = 2$	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.02	0.12	0.37	0.55	0.82	0.97	0.99
SSMT: $\rho_n = 3$	0.00	0.01	0.04	0.01	0.04	0.08	0.03	0.08	0.14	0.41	0.69	0.81	0.96	1.00	1.00
BCS2015	0.00	0.00	0.01	0.00	0.02	0.06	0.01	0.06	0.14	0.71	0.93	0.97	1.00	1.00	1.00
AS2017	0.00	0.03	0.04	0.01	0.11	0.18	0.07	0.26	0.37	0.81	0.97	0.99	1.00	1.00	1.00
Least Favorable	0.00	0.02	0.04	0.01	0.09	0.17	0.06	0.21	0.35	0.79	0.95	0.98	1.00	1.00	1.00

Notes: The values displayed in the table are the proportion of tests rejected across 500 experiments with $n = 1000$ observations. LinearDGP_1 is correctly specified and strictly partially identified. All the other DGPs are misspecified. The SSMT procedure uses $r_n = \lfloor n^{1/5} \rfloor = 3$. BCS2015 refers to the RS test of Bugni, Canay, and Shi (2015) using the GMS function from (4.3). AS2017 refers to the by-product test using Andrews and Shi (2017) with the GMS function from (4.4), implemented using a grid of 25^2 equally spaced points belonging to $[-1, 1]^2$. "Least Favorable" refers to a least favorable procedure similar to AS2017 but with the GMS function set to 0.

where ζ_n is defined below. Now consider the following moment conditions:

$$E_{P_n}[X_{i1} - \theta_1] \leq 0, \qquad E_{P_n}[\theta_1 - X_{i2}] \leq 0, \qquad E_{P_n}[X_{i3} - \theta_2] \leq 0.$$

In this model the identified set is given by $\Theta_I(P_n) := \{\theta \in \Theta : \theta_1 \in [0, -\zeta_n], \theta_2 \geq 0\}$. The model is therefore correctly specified and strictly partially identified if $\zeta_n = 0$, a case we consider in LinearDPG_1. Furthermore, it is misspecified if $\zeta_n > 0$, and we consider the cases $\zeta_n = 1/\sqrt{n}$ in LinearDGP_2, $\zeta_n = 1/\sqrt{q_n}$ in LinearDGP_3, $\zeta_n = 6/\sqrt{n}$ in LinearDGP_4 and $\zeta_n = 6/\sqrt{q_n}$ in LinearDGP_5. We set $\hat{\varsigma}_n(\theta, \tau, g)$ equal to the sample standard deviation of the moment functions, and consider $n = 1000$, $r_n = \lfloor n^{1/5} \rfloor = 3$ and $\rho_n \in \{1, 2, 3\}$ when implementing the SSMT procedure. Since the identified set is easy to estimate in this example, we are able to minimize each bootstrap test statistic over the estimated identified set when implementing the RS test of Bugni, Canay, and Shi (2015). Furthermore, the simplicity of the example allows us to use a finer grid of 25^2 equally spaced points over $[-1, 1]^2$ when implementing the by-product tests. For these alternative methods, we use the same criterion function and GMS functions as in the previous simulation example.

The results are reported in Table 3. As expected, all three tests deliver rejection rates below the nominal level in LinearDGP_1 where the model is correctly specified. Interestingly, while the SSMT procedure has lower power than the RS test in LinearDGP_3, LinearDGP_4 and LinearDGP_5, both tests have rejection rates less than the nominal level in LinearDGP_2. Overall, the performance of the SSMT procedure against $q_n^{-1/2}$ alternatives is consistent with the discussion in Section 3.5. The procedure has non-trivial power against some distant $q_n^{-1/2}$ alternatives, such as in LinearDGP_5, but has low power against alternatives that are close to the null, such as in LinearDGP_2, LinearDGP_3 and LinearDGP_4.

5 Conclusion

This paper proposes a simple specification test for models defined by conditional moment inequalities. Our test is valid under weak assumptions on the moment conditions, and is especially useful for models in which the identified set and associated confidence sets are difficult to compute. The procedure obtains a computational advantage by reusing the minimizer of our MinMax test statistic when computing the critical value using a bootstrap procedure. Under our weak assumptions, reusing the minimizer introduces new theoretical complications which we overcome using a sample-splitting procedure. We prove that the procedure controls size uniformly over a large class of data generating processes, and has power tending to 1 for fixed and local alternatives. This paper continues an existing line of research that seeks to provide computationally accessible methods for inference in partially identified models. We believe that developing computationally-minded methods of inference under weak assumptions requires novel solutions, and we hope our unconventional testing procedure might inspire other researchers to develop unconventional procedures of their own.

References

An, L. T. H., and P. D. Tao (2005): "The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems," *Annals of operations research*, 133(1), 23–46.

Andrews, D. W., and P. Guggenberger (2009): "Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities," *Econometric Theory*, 25(3), 669–709.

Andrews, D. W., and S. Kwon (2019): "Inference in moment inequality models that is robust to spurious precision under model misspecification," .

——— (2021): "Misspecified Moment Inequality Models: Diagnostics and Inference," .

Andrews, D. W., and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81(2), 609–666.

——— (2017): "Inference based on many conditional moment inequalities," *Journal of Econometrics*, 196(2), 275–287.

Andrews, D. W., and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78(1), 119–157.

Barseghyan, L., M. Coughlin, F. Molinari, and J. C. Teitelbaum (2021): "Heterogeneous choice sets and preferences," *Econometrica*, 89(5), 2015–2048.

Belloni, A., F. A. Bugni, and V. Chernozhukov (2019): "Subvector inference in PI models with many moment inequalities," Discussion paper, cemmap working paper.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79(6), 1785–1821.

Bugni, F. A., I. A. Canay, and X. Shi (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185(1), 259–282.

——— (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8(1), 1–38.

Chesher, A., and A. M. Rosen (2014): "An instrumental variable random-coefficients model for binary outcomes," *The econometrics journal*, 17(2), S1–S19.

——— (2017): "Generalized instrumental variable models," *Econometrica*, 85(3), 959–989.

——— (2020): "Generalized instrumental variable models, methods, and applications," in *Handbook of Econometrics*, vol. 7, pp. 1–110. Elsevier.

Dudley, R. M. (2002): *Real analysis and probability*. CRC Press.

——— (2014): *Uniform central limit theorems*, vol. 142. Cambridge university press.

Galichon, A., and M. Henry (2009): "A test of non-identifying restrictions and confidence regions for partially identified parameters," *Journal of Econometrics*, 152(2), 186–196.

Giné, E., and R. Nickl (2021): *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press.

Guggenberger, P., J. Hahn, and K. Kim (2008): "Specification testing under moment inequalities," *Economics Letters*, 99(2), 375–378.

Jones, D. R. (2001): "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, 21(4), 345–383.

Jones, D. R., M. Schonlau, and W. J. Welch (1998): "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, 13(4), 455–492.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence intervals for projections of partially identified parameters," *Econometrica*, 87(4), 1397–1432.

——— (2022): "Constraint qualifications in partial identification," *Econometric Theory*, 38(3), 596–619.

Kaido, H., and H. White (2013): "Estimating misspecified moment inequality models," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 331–361. Springer.

Kédagni, D., L. Li, and I. Mourifié (2020): "Discordant relaxations of misspecified models," *arXiv preprint arXiv:2012.11679*.

Kitagawa, T. (2015): "A test for instrument validity," *Econometrica*, 83(5), 2043–2063.

Kosorok, M. R. (2007): *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

Le Thi, H. A., H. M. Le, D. N. Phan, and B. Tran (2020): "Stochastic DCA for minimizing a large sum of DC functions with application to multi-class logistic regression," *Neural Networks*, 132, 220–231.

Le Thi, H. A., and T. Pham Dinh (2018): "DC programming and DCA: thirty years of developments," *Mathematical Programming*, 169(1), 5–68.

Magnolfi, L., and C. Roncoroni (2023): "Estimation of discrete games with weak assumptions on information," *The Review of Economic Studies*, 90(4), 2006–2041.

Masten, M. A., and A. Poirier (2021): "Salvaging falsified instrumental variable models," *Econometrica*, 89(3), 1449–1469.

Meng, X.-L. (1994): "Posterior predictive p-values," *The annals of statistics*, 22(3), 1142–1160.

Molinari, F. (2020): "Microeconometrics with partial identification," *Handbook of econometrics*, 7, 355–486.

Mourifié, I., and Y. Wan (2017): "Testing local average treatment effect assumptions," *Review of Economics and Statistics*, 99(2), 305–313.

Mullen, K., D. Ardia, D. L. Gil, D. Windover, and J. Cline (2011): "DEoptim: An R package for global optimization by differential evolution," *Journal of Statistical Software*, 40(6), 1–26.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009): "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, 19(4), 1574–1609.

Nevo, A., and A. M. Rosen (2012): "Identification with imperfect instruments," *Review of Economics and Statistics*, 94(3), 659–671.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*. Springer Science & Business Media.

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

Romano, J. P., and A. M. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.

Rüschendorf, L. (1982): "Random variables with maximum sums," *Advances in Applied Probability*, 14(3), 623–632.

Santos, A. (2012): "Inference in nonparametric instrumental variables with partial identification," *Econometrica*, 80(1), 213–275.

Sheehy, A., and J. A. Wellner (1992): "Uniform Donsker classes of functions," *The Annals of Probability*, 20(4), 1983–2030.

Słoczyński, T., S. D. Uysal, and J. M. Wooldridge (2022): "Abadie's Kappa and Weighting Estimators of the Local Average Treatment Effect," *arXiv preprint arXiv:2204.07672*.

Stoye, J. (2020): "A simple, short, but never-empty confidence interval for partially identified parameters," *arXiv preprint arXiv:2010.10484*.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak convergence*. Springer.

A Proofs of the Main Results

A.1 Preliminaries

Let $(\mathbb{W}, \mathscr{W})$ be a measureable space, let $\{P_n\}_{n\geq 0} \subset \mathcal{P}$, and let $\mathscr{P}$ denote the collection of all probability measures on $\mathbb{W}$. To accommodate drifting sequences in the proofs, we take the underlying probability space to be of the form $(\Omega, \mathfrak{F}, \mathbb{P}) := (\mathbb{W}^1, \mathscr{W}^1, P_1^1) \times (\mathbb{W}^2, \mathscr{W}^2, P_2^2) \times \ldots \times (\mathbb{W}^n, \mathscr{W}^n, P_n^n) \times \ldots \times ([0,1], \mathscr{L}, \lambda)$, where λ denotes the Lebesgue measure, and $\mathscr{L}$ denotes the Lebesgue subsets of $[0,1]$.[38] We then view $W_i : \Omega \to \mathbb{W}$ as coordinate projections on the probability space $(\mathbb{W}, \mathscr{W}, P_n)$ for each $i \in \mathbb{N}$. Note this implies $W_1, \ldots, W_n$ are independent and "row-wise" identically distributed according to P_n. We denote the product probability P_n^n on the measurable space $(\mathbb{W}^n, \mathscr{W}^n)$ as $\Pr_{P_n}$, and denote the expectation with respect to $\Pr_{P_n}$ as E_{P_n}. Since we work with coordinate projections, along a fixed sequence $\{P_n\}_{n\geq 0}$, probability statements with respect to $\Pr_{P_n}$ are identical to probability statements with respect to $\mathbb{P}$. We often use $\Pr_{P_n}$ instead of $\mathbb{P}$ in the proofs in order to emphasize the underlying sequence $\{P_n\}_{n\geq 0}$.

Define $\mathbb{T} := \Theta \times \mathcal{T} \times \mathcal{G}$, and let $t = (\theta, \tau, g)$ denote a typical element of $\mathbb{T}$. Now define the class of functions:

$$\mathcal{M}_P := \{m(\,\cdot\,, t)/\varsigma_P(t) : t \in \mathbb{T}\}\,.$$

The class $\mathcal{M}_P$ can then be equipped with the envelope $\boldsymbol{M}$ from Assumptions 3.1. Under Assumption 3.1, we have $\sup_{f_t \in \mathcal{M}_P} |f_t(w) - E_P[f_t(W_i)]| < \infty$ for all $w \in \mathbb{W}$ and each $P \in \mathcal{P}$. Furthermore, under Assumption 3.1, the class $\mathcal{M}_P$ is pointwise measurable and satisfies Dudley's entropy condition unformly in $P \in \mathcal{P}$. The map $t \mapsto f_t(w) - E_P[f_t(W_i)]$ for $f_t \in \mathcal{M}_P$ can thus be viewed as an element of $\ell^\infty(\mathbb{T})$, the Banach space of bounded real-valued functions on $\mathbb{T}$ equipped with the sup norm. Throughout, $\ell^\infty(\mathbb{T})$ is equipped with the Borel $\sigma-$algebra. We denote the (random) empirical measure as:

$$\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{W_i},$$

where δ_{W_i} is the Dirac-delta. We can then view the normalized empirical process:

$$v_{n,P_n}(t) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{m(W_i,t)}{\varsigma_P(t)} - \frac{E_{P_n}[m(W_i,t)]}{\varsigma_P(t)}\right),$$

as an element of $\ell^\infty(\mathbb{T})$. Along a given sequence $\{P_n \in \mathcal{P}\}_{n\geq 1}$, the map $(t,n) \mapsto f_n(t) := f_n(\,\cdot\,, t) = m(\,\cdot\,, t)/\varsigma_{P_n}(t)$ is onto. Thus, we will often switch between indexing the empirical process v_{n,P_n} by t (which is convenient for the main proofs) or by $f_n \in \mathcal{M}_{P_n}$ (which is convenient for the proofs in

[38] A countable product of probability spaces is a probability space: see Dudley (2002) Theorem 8.2.2.

the Online Supplementary Material). For $P \in \mathscr{P}$, $t, t' \in \mathbb{T}$, define the semi-metric:

$$\rho_P^2(t,t') := \mathrm{Var}_P\left(\frac{m(W_i,t)}{\varsigma_P(t)} - \frac{m(W_i,t')}{\varsigma_P(t')}\right).$$

For every $P \in \mathscr{P}$, we let v_P represent the mean-zero real-valued Gaussian process with covariance kernel:

$$h_{2,P}(t,t') := \mathbb{C}ov(v_P(t), v_P(t')) = E_P\left[\frac{m(W_i,t)}{\varsigma_P(t)}\frac{m(W_i,t')}{\varsigma_P(t')}\right] - E_P\left[\frac{m(W_i,t)}{\varsigma_P(t)}\right]E_P\left[\frac{m(W_i,t')}{\varsigma_P(t')}\right].$$

If there exists a version of v_P, say $\tilde{v}_P$, with ρ_P-uniformly continuous sample paths, then we set v_P equal to that version. The space of all covariance kernels on $\mathbb{T}$ is given by:

$$\mathcal{H}_2 := \{h_{2,P}(\,\cdot\,,\,\cdot\,) : P \in \mathcal{P}\},$$

equipped with the metric:

$$d(h_2^{(1)}, h_2^{(2)}) = \sup_{t,t' \in \mathbb{T}} \left| h_2^{(1)}(t,t') - h_2^{(2)}(t,t') \right|.$$

We often use "$\overset{u}{\to}$" to denote convergence in the sup norm. We also use the notation:

$$h_{1,q_n,P}(\theta,\tau,g) := \frac{\sqrt{q_n}E_P[m(W_i,\theta,\tau,g)]}{\varsigma_P(\theta,\tau,g)}. \tag{A.1}$$

Weak convergence of a sequence of (possibly non-measurable) random elements to separable limit is metrizable by the bounded Lipschitz metric (see van der Vaart and Wellner (1996) Theorem 1.12.4). In particular, consider the metric space $\ell^\infty(\mathbb{T})$ and let:

$$BL_1(\ell^\infty(\mathbb{T})) := \{h : \ell^\infty(\mathbb{T}) \to \mathbb{R} : ||h||_\infty \leq 1 \text{ and } |h(x) - h(y)| \leq |x - y| \text{ for all } x \neq y\}.$$

Then a sequence of stochastic processes $\mathbb{X}_n(t,\omega)$ on $\mathbb{T}$ with bounded sample paths converges in law to a measurable, separable process $\mathbb{X}(t,\omega)$ along the sequence $\{P_n\}_{n\geq 0} \subset \mathcal{P}$ if and only if:

$$d_{BL_1}(\mathbb{X}_n, \mathbb{X}) := \sup_{h \in BL_1(\ell^\infty(\mathbb{T}))} |E_{P_n}^* h(\mathbb{X}_n) - E_{P_n} h(\mathbb{X})| \to 0,$$

as $n \to \infty$, where $E_{P_n}^*$ denotes the outer expectation.

For the bootstrap results, we also require a precise description of the underlying probability space. Suppose $W_1^{(r)}, \ldots, W_{q_n}^{(r)}$ are independent and identically distributed with distribution P_{q_n} on $(\mathbb{W}, \mathscr{W})$, and recall that $\mathbb{P}_{q_n}$ denotes the empirical distribution for a sample of size q_n. Now let $W_1^{(r)\sharp}, \ldots, W_{q_n}^{(r)\sharp}$ denote random variables that are independent and identically distributed with

distribution $\mathbb{P}_{q_n}$ on $(\mathbb{W}, \mathscr{W})$, and define:

$$v_{q_n}^{(r)\sharp}(t) = \frac{1}{\sqrt{q_n}} \sum_{i=1}^{q_n} \left(\frac{m(W_i^{(r)\sharp}, t)}{\hat{\varsigma}_n(t)} - \frac{m(W_i^{(r)}, t)}{\hat{\varsigma}_n(t)} \right), \tag{A.2}$$

$$v_{q_n, P_n}^{(r)\sharp}(t) = \frac{1}{\sqrt{q_n}} \sum_{i=1}^{q_n} \left(\frac{m(W_i^{(r)\sharp}, t)}{\varsigma_{P_n}(t)} - \frac{m(W_i^{(r)}, t)}{\varsigma_{P_n}(t)} \right). \tag{A.3}$$

Then we can take the underlying probability space to be the product of $(\Omega, \mathfrak{F}, \mathbb{P})$ with a probability space $(\Omega^\sharp, \mathfrak{F}^\sharp, \mathbb{P}^\sharp)$ on which we can define the random variables $i_1^{(r)}, \ldots, i_{q_n}^{(r)}$ with uniform distribution on $\{1, \ldots, q_n\}$ for all r. Then take $W_j^{(r)\sharp}(\omega, \omega^\sharp) := W^{(r)}_{i_j^{(r)}(\omega^\sharp)}(\omega)$.[39] Note that $v_{q_n}^{(r)\sharp}(t)$ depends on $\omega^\sharp$, and implicitly depends on $\omega \in \Omega$ through the empirical measure $\mathbb{P}_{q_n} = \mathbb{P}_{q_n}(\omega)$. Occasionally it will be useful to emphasize this dependence, in which case we write $v_{q_n}^{(r)\sharp}(\omega^\sharp, \omega)(t)$.

Throughout the proofs, we ignore measurability issues for simplicity unless they are crucial to the argument, in which case we use a superscript $*$ to denote an outer measure, outer expectation, or measurable majorant, depending on the context. Throughout the proofs we use the fact that:

$$\left| \sup_{x \in \mathcal{X}} f(x) - \sup_{x \in \mathcal{X}} g(x) \right| \leq \sup_{x \in \mathcal{X}} |f(x) - g(x)|. \tag{A.4}$$

We also repeatedly use the fact that:

$$\left| \inf_{y \in \mathcal{Y}} \sup_{x \in \mathcal{X}} f(x, y) - \inf_{y \in \mathcal{Y}} \sup_{x \in \mathcal{X}} g(x, y) \right| \leq \sup_{y \in \mathcal{Y}} \sup_{x \in \mathcal{X}} |f(x, y) - g(x, y)|. \tag{A.5}$$

A.2 Proofs

Proof of Lemma 3.1. Condition (B.1) in Lemma B.1 is trivially satisfied for every sequence $\{P_n \in \mathcal{P}_0\}_{n \geq 1}$. The result then follows immediately from Lemma B.1 and B.2. ■

Proof of Theorem 3.1. Consider the limit in (3.1). Since $\rho_n \in [1, \infty)$ is a decreasing sequence bounded from below, the monotone convergence theorem implies that $\rho_n \downarrow \rho$ for some $\rho \in [1, \infty)$. Thus we have:

$$\begin{aligned}
&\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_0 \colon (h_{2,P}, h_{3,P}) \in \mathscr{H}_{cpt}} (\Pr_P \times \mathbb{P}^\sharp)(\phi_n(\rho_n, \alpha) = 1) \\
&= \limsup_{n \to \infty} \sup_{P \in \mathcal{P}_0 \colon (h_{2,P}, h_{3,P}) \in \mathscr{H}_{cpt}} (\Pr_P \times \mathbb{P}^\sharp) \left(\frac{1}{r_n} \sum_{r=1}^{r_n} \mathbb{1}\{T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1 - \alpha/\rho_n + \eta) + \eta\} \geq \frac{1}{\rho_n} \right) \\
&\overset{(1)}{\leq} \limsup_{n \to \infty} \sup_{P \in \mathcal{P}_0 \colon (h_{2,P}, h_{3,P}) \in \mathscr{H}_{cpt}} \int \frac{\rho_n}{r_n} \sum_{r=1}^{r_n} \mathbb{1}\{T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1 - \alpha/\rho_n + \eta) + \eta\} \, d(\Pr_P \times \mathbb{P}^\sharp) \\
&\overset{(2)}{=} \limsup_{n \to \infty} \sup_{P \in \mathcal{P}_0 \colon (h_{2,P}, h_{3,P}) \in \mathscr{H}_{cpt}} \frac{\rho_n}{r_n} \sum_{r=1}^{r_n} \int \mathbb{1}\{T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1 - \alpha/\rho_n + \eta) + \eta\} \, d(\Pr_P \times \mathbb{P}^\sharp)
\end{aligned}$$

[39] See Dudley (2014) p. 324.

$$
\begin{aligned}
&= \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} \frac{\rho_n}{r_n}\sum_{r=1}^{r_n}(\Pr_P\times\mathbb{P}^\sharp)\left(T_{q_n}^{(r)}(\hat\theta_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho_n+\eta)+\eta\right)\\
&\overset{(3)}{=} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} \rho_n(\Pr_P\times\mathbb{P}^\sharp)\left(T_{q_n}^{(1)}(\hat\theta_n) > c_{q_n}^{(1)\sharp}(1-\alpha/\rho_n+\eta)+\eta\right)\\
&\overset{(4)}{\leq} \rho\cdot\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} (\Pr_P\times\mathbb{P}^\sharp)\left(T_{q_n}^{(1)}(\hat\theta_n) > c_{q_n}^{(1)\sharp}(1-\alpha/\rho+\eta)+\eta\right),
\end{aligned}
$$

where (1) follows from Markov's inequality, (2) follows from Tonelli's theorem, and (3) follows from the fact that:

$$
\begin{aligned}
&(\Pr_P\times\mathbb{P}^\sharp)\left(T_{q_n}^{(r)}(\hat\theta_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho_n+\eta)+\eta\right)\\
&\qquad = (\Pr_P\times\mathbb{P}^\sharp)\left(T_{q_n}^{(q)}(\hat\theta_n) > c_{q_n}^{(q)\sharp}(1-\alpha/\rho_n+\eta)+\eta\right),
\end{aligned}
$$

for all $r,q\in\mathbb{N}$ by the i.i.d. assumption in Assumption 3.1, and by construction of the test statistics and critical values. Finally, (4) follows from the fact that $\limsup(a_n\cdot b_n)\leq(\limsup a_n)\cdot(\limsup b_n)$ for any sequences $\{a_n\}_{n\geq1}$ and $\{b_n\}_{n\geq1}$. The rest of the proof continues for a fixed $r\in\{1,\ldots,r_n\}$ and $\rho\in[1,\infty)$. Fix any $\varepsilon\in(0,\eta)$ and let $A_{n,P}$ denote the event:

$$
A_{n,P} := \left\{\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_n}\frac{\sqrt{q_n}E_P[m(W_i,\hat\theta_n,\tau,g)]}{\varsigma_P(\hat\theta_n,\tau,g)}\leq\varepsilon\right\}. \tag{A.6}
$$

We have:

$$
\begin{aligned}
&\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} (\Pr_P\times\mathbb{P}^\sharp)\left(T_{q_n}^{(r)}(\hat\theta_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho+\eta)+\eta\right)\\
&\leq \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} (\Pr_P\times\mathbb{P}^\sharp)\left(\left\{T_{q_n}^{(r)}(\hat\theta_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho+\eta)+\eta\right\}\cap A_{n,P}\right)\\
&\qquad + \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} (\Pr_P\times\mathbb{P}^\sharp)\left(\left\{T_{q_n}^{(r)}(\hat\theta_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho+\eta)+\eta\right\}\cap A_{n,P}^c\right)\\
&\qquad \leq \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} (\Pr_P\times\mathbb{P}^\sharp)\left(\left\{T_{q_n}^{(r)}(\hat\theta_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho+\eta)+\eta\right\}\cap A_{n,P}\right)\\
&\qquad\qquad + \limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} \Pr_P\left(A_{n,P}^c\right).
\end{aligned}
$$

By Lemma 3.1 we have:

$$
\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} \Pr_P\left(A_{n,P}^c\right) = 0.
$$

Now there exists a sequence $\{(P_n,h_{2,P_n},h_{3,P_n})\in\mathcal{P}_0\times\mathscr{H}_{cpt}:n\geq1\}$ such that:

$$
\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0:\ (h_{2,P},h_{3,P})\in\mathscr{H}_{cpt}} (\Pr_P\times\mathbb{P}^\sharp)\left(\left\{T_{q_n}^{(r)}(\hat\theta_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho+\eta)+\eta\right\}\cap A_{n,P}\right)
$$

$$= \limsup_{n\to\infty} (\Pr_{P_n} \times \mathbb{P}^\sharp) \left(\left\{ T^{(r)}_{q_n}(\hat{\theta}_n) > c^{(r)\sharp}_{q_n}(1-\alpha/\rho+\eta) + \eta \right\} \cap A_{n,P_n} \right).$$

Furthermore, by compactness of $\mathscr{H}_{cpt}$ there exists a further subsequence $\{a_n\}_{n\geq 1} \subset \{n\}_{n\geq 1}$, some $P_0 \in \mathcal{P}$ and some corresponding $(h_{2,P_0}, h_{3,P_0}) \in \mathscr{H}_{cpt}$ such that $h_{2,P_{a_n}} \xrightarrow{u} h_{2,P_0}$ and $h_{3,P_{a_n}} \xrightarrow{u} h_{3,P_0}$. We continue the proof along this subsequence. By Lemma B.5 we have for any $\delta \in (\varepsilon, \eta)$:

$$\begin{aligned}
&\limsup_{n\to\infty} (\Pr_{P_{a_n}} \times \mathbb{P}^\sharp) \left(\left\{ T^{(r)}_{q_{a_n}}(\hat{\theta}_{a_n}) > c^{(r)\sharp}_{q_{a_n}}(1-\alpha/\rho+\eta) + \eta \right\} \cap A_{a_n,P_{a_n}} \right) \\
&\qquad \leq \limsup_{n\to\infty} \Pr_{P_{a_n}} \left(\left\{ T^{(r)}_{q_{a_n}}(\hat{\theta}_{a_n}) > c^{(r)}_0(\hat{\theta}_{a_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha/\rho) + \delta \right\} \cap A_{a_n,P_{a_n}} \right),
\end{aligned}$$

where $c^{(r)}_0(\hat{\theta}_{a_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha/\rho)$ is the $1-\alpha/\rho$ quantile of the distribution of:

$$T_0(\hat{\theta}_{a_n}, 0, h_{2,P_0}, h_{3,P_0}) = \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left\{ v_{P_0}(\hat{\theta}_{a_n}, \tau, g) \right\},$$

where v_{P_0} is a mean zero Gaussian process on $\Theta \times \mathcal{T} \times \mathcal{G}$ with covariance kernel h_{2,P_0}. Furthermore, by Lemma B.6 we have:

$$\begin{aligned}
&\limsup_{n\to\infty} \Pr_{P_{a_n}} \left(\left\{ T^{(r)}_{q_{a_n}}(\hat{\theta}_{a_n}) > c^{(r)}_0(\hat{\theta}_{a_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha/\rho) + \delta \right\} \cap A_{a_n,P_{a_n}} \right) \\
&\qquad \leq \limsup_{n\to\infty} \Pr_{P_{a_n}} \left(\left\{ T^{(r)}_{q_{a_n}}(\hat{\theta}_{a_n}) > c^{(r)}_0(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha/\rho) + \delta - \varepsilon \right\} \cap A_{a_n,P_{a_n}} \right),
\end{aligned}$$

where $c^{(r)}_0(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha/\rho)$ is the $1-\alpha/\rho$ quantile of the distribution of:

$$\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left\{ v_{P_0}(\hat{\theta}_{a_n}, \tau, g) + h_{1,q_{a_n},P_{a_n}}(\hat{\theta}_{a_n}, \tau, g) \right\}.$$

Finally, from Lemma B.4 we have:

$$\limsup_{n\to\infty} \Pr_{P_{a_n}} \left(T^{(r)}_{q_{a_n}}(\hat{\theta}_{a_n}) > c^{(r)}_0(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha/\rho) + \delta - \varepsilon \right) \leq \alpha/\rho.$$

This completes the proof. ■

Proof of Theorem 3.2. Since $\rho_n \in [1,\infty)$ is decreasing, it is bounded from above by ρ_1. Note that for any $M > 0$:

$$\begin{aligned}
&(\Pr_{P_n} \times \mathbb{P}^\sharp)(\phi_n(\rho_n, \alpha) = 1) \\
&= (\Pr_{P_n} \times \mathbb{P}^\sharp) \left(\sum_{r=1}^{r_n} \mathbb{1} \left\{ T^{(r)}_{q_n}(\hat{\theta}_n) > c^{(r)\sharp}_{q_n}(1-\alpha/\rho_n+\eta) + \eta \right\} \geq \frac{r_n}{\rho_n} \right) \\
&\geq (\Pr_{P_n} \times \mathbb{P}^\sharp) \left(\sum_{r=1}^{r_n} \mathbb{1} \left\{ T^{(r)}_{q_n}(\hat{\theta}_n) > c^{(r)\sharp}_{q_n}(1-\alpha/\rho_n+\eta) + \eta \right\} \geq r_n \right) \\
&= (\Pr_{P_n} \times \mathbb{P}^\sharp) \left(T^{(1)}_{q_n}(\hat{\theta}_n) > c^{(1)\sharp}_{q_n}(1-\alpha/\rho_n+\eta) + \eta, \ldots, T^{(r_n)}_{q_n}(\hat{\theta}_n) > c^{(r_n)\sharp}_{q_n}(1-\alpha/\rho_n+\eta) + \eta \right)
\end{aligned}$$

$$
\begin{aligned}
&\geq \max\left\{\sum_{r=1}^{r_n}(\Pr_{P_n}\times\mathbb{P}^\sharp)\left(T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho_n+\eta)+\eta\right)-(r_n-1),0\right\}\\
&= \max\left\{r_n\cdot(\Pr_{P_n}\times\mathbb{P}^\sharp)\left(T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho_n+\eta)+\eta\right)-(r_n-1),0\right\}\\
&\geq \max\left\{r_n\cdot(\Pr_{P_n}\times\mathbb{P}^\sharp)\left(T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho_1+\eta)+\eta\right)-(r_n-1),0\right\},
\end{aligned}
$$

where the first inequality follows from the fact that $\rho_n \geq 1$, the second inequality follows from the Frechet-Hoeffding bounds, and the final inequality follows from the fact that $\rho_n \leq \rho_1$. Thus, it suffices to show that, if $\sqrt{q_n}\mu_n \to \infty$, then:

$$
\liminf_{n\to\infty}(\Pr_{P_n}\times\mathbb{P}^\sharp)\left(T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho_1+\eta)+\eta\right)=1.
$$

To this end, note that:

$$
\begin{aligned}
&(\Pr_{P_n}\times\mathbb{P}^\sharp)(T_{q_n}^{(r)}(\hat{\theta}_n) > c_{q_n}^{(r)\sharp}(1-\alpha/\rho_1+\eta)+\eta)\\
&\geq (\Pr_{P_n}\times\mathbb{P}^\sharp)(T_{q_n}^{(r)}(\hat{\theta}_n) \geq M, c_{q_n}^{(r)\sharp}(1-\alpha/\rho_1+\eta)+\eta < M)\\
&\geq \max\Big\{(\Pr_{P_n}\times\mathbb{P}^\sharp)(T_{q_n}^{(r)}(\hat{\theta}_n)\geq M)\\
&\qquad\qquad +(\Pr_{P_n}\times\mathbb{P}^\sharp)(c_{q_n}^{(r)\sharp}(1-\alpha/\rho_1+\eta)+\eta<M)-1,0\Big\},
\end{aligned}
$$

where the last inequality follows from the Frechet-Hoeffding bounds. The proof now consists of two steps. In the first step we show that:

$$
\lim_{M_1\to\infty}\limsup_{n\to\infty}\ (\Pr_{P_n}\times\mathbb{P}^\sharp)(c_{q_n}^{(r)\sharp}(1-\alpha/\rho_1+\eta)+\eta > M_1)=0.
$$

In the second step we show that for any $M_2 > 0$:

$$
\limsup_{n\to\infty}(\Pr_{P_n}\times\mathbb{P}^\sharp)(T_{q_n}^{(r)}(\hat{\theta}_n)\leq M_2)=0.
$$

Step 1: We begin by showing that $c_{q_n}^{(r)\sharp}(1-\alpha/\rho_1+\eta)$ is stochastically bounded from above. By definition (holding fixed $\omega\in\Omega$):

$$
\begin{aligned}
c_{q_n}^{(r)\sharp}(1-\alpha/\rho_1+\eta) &:= \inf\left\{x:\mathbb{P}^\sharp\left(\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_n}\left|v_{q_n}^{(r)\sharp}(\hat{\theta}_n,\tau,g)\right|\leq x\right)\geq 1-\alpha/\rho_1+\eta\right\}\\
&= \inf\left\{x:\mathbb{P}^\sharp\left(\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_n}\left|v_{q_n}^{(r)\sharp}(\hat{\theta}_n,\tau,g)\right|> x\right)\leq \alpha/\rho_1-\eta\right\}\\
&\leq \inf\left\{x:\mathbb{P}^\sharp\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v_{q_n}^{(r)\sharp}(\theta,\tau,g)\right|> x\right)\leq \alpha/\rho_1-\eta\right\}, \qquad \text{(A.7)}
\end{aligned}
$$

where $v_{q_n}^{(r)\sharp}(\theta,\tau,g)$ is defined in (A.2). Furthermore, by Lemma S.2.3, we have for any $M>0$:

$$
\mathbb{P}^\sharp\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v_{q_n,P_n}^{(r)\sharp}(\theta,\tau,g)\right|>M\right)\leq\frac{1}{M}\mathbb{E}^\sharp\left[\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v_{q_n,P_n}^{(r)\sharp}(\theta,\tau,g)\right|\right]\leq\frac{C^\sharp}{M},
$$

where $v^{(r)\sharp}_{q_n,P_n}(\theta,\tau,g)$ is defined in (A.3), and $C^\sharp$ is a constant (i.e. does not depend on either n or $\omega \in \Omega$). Conclude that, for any $\varepsilon > 0$ there is an M such that:

$$\mathbb{P}^\sharp\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n,P_n}(\theta,\tau,g)\right| > M\right) < \varepsilon,$$

for all $n \geq 1$ and $\omega \in \Omega$, and thus:

$$(\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n,P_n}(\theta,\tau,g)\right| > M\right) < \varepsilon.$$

Conclude that:

$$\lim_{M_1\to\infty}\limsup_{n\to\infty}(\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n,P_n}(\theta,\tau,g)\right| > M_1\right) = 0. \tag{A.8}$$

Now consider the events (for any $\delta > 0$):

$$B_{n,M_1,P_n} := \left\{\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n}(\theta,\tau,g)\right| > M_1\right\}, \quad F_{n,\delta,P_n} := \left\{\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_n}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} - 1\right| > \delta\right\}.$$

Then by Assumption 3.2 we have:

$$\lim_{M_1\to\infty}\limsup_{n\to\infty}(\Pr_{P_n} \times \mathbb{P}^\sharp)\left(B_{n,M_1,P_n}\right) \leq \lim_{M_1\to\infty}\limsup_{n\to\infty}(\Pr_{P_n} \times \mathbb{P}^\sharp)\left(B^{(1)}_{n,M_1,P_n} \cap F^c_{n,\delta,P_n}\right).$$

Furthermore, note that on the event $B^{(1)}_{n,M_1,P_n} \cap F^c_{n,\delta,P_n}$:

$$\begin{aligned}
\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n}(\theta,\tau,g)\right| &\leq \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|\frac{\varsigma_P(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)}\right| \cdot \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n,P_n}(\theta,\tau,g)\right| \\
&\leq \left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|\frac{\varsigma_P(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} - 1\right| + 1\right) \cdot \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n,P_n}(\theta,\tau,g)\right| \\
&\leq (1+\delta) \cdot \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n,P_n}(\theta,\tau,g)\right|.
\end{aligned}$$

Combining this with (A.7) and (A.8), conclude that:

$$\begin{aligned}
&\lim_{M_1\to\infty}\limsup_{n\to\infty}(\Pr_{P_n} \times \mathbb{P}^\sharp)\left(c^{(r)\sharp}_{q_n}(1-\alpha/\rho_1+\eta) + \eta > M_1\right) \\
&\leq \lim_{M_1\to\infty}\limsup_{n\to\infty}(\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n}(\theta,\tau,g)\right| > M_1\right) \\
&\leq \lim_{M_1\to\infty}\limsup_{n\to\infty}(\Pr_{P_n} \times \mathbb{P}^\sharp)\left((1+\delta) \cdot \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_n,P_n}(\theta,\tau,g)\right| > M_1\right) = 0,
\end{aligned}$$

for any $\alpha > 0$ and $\rho_1 \in [1,\infty)$.

Step 2: We will show that for every $M_2 > 0$:

$$\limsup_{n\to\infty}(\Pr_{P_n} \times \mathbb{P}^\sharp)(T_{q_n}^{(r)}(\hat{\theta}_n) \leq M_2) = 0.$$

Note that:

$$\begin{aligned} T_{q_n}^{(r)}(\hat{\theta}_n) \geq T_{q_n}^{(r)} &:= \inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_n}\frac{\sqrt{q_n}\bar{m}_{q_n}^{(r)}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} \\ &= \inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_n}\left\{\frac{\sqrt{q_n}(\bar{m}_{q_n}^{(r)}(\theta,\tau,g) - E_{P_n}[m(W_i,\theta,\tau,g)])}{\hat{\varsigma}_n(\theta,\tau,g)} + \frac{\sqrt{q_n}E_{P_n}[m(W_i,\theta,\tau,g)]}{\hat{\varsigma}_n(\theta,\tau,g)}\right\}. \end{aligned}$$

Thus it suffices to show that for any $M_2 > 0$ and any $\varepsilon > 0$ there exists an N such that $\forall n \geq N$:

$$(\Pr_{P_n} \times \mathbb{P}^\sharp)(T_{q_n}^{(r)} \leq M_2) < \varepsilon.$$

To do so, fix arbitrary $M_2 > 0$ and $\varepsilon > 0$, let $\varsigma_{P_n}(\theta,\tau,g)$ be as in Assumption 3.2, and define:

$$T_n := \inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\frac{\sqrt{q_n}E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)}.$$

By Assumption 3.4, we have:

$$\begin{aligned} T_n &\overset{(1)}{\geq} \inf_{\theta\in\Theta}\frac{\sqrt{q_n}E_{P_n}[m(W_i,\theta,\tau_{n,\theta},g_{n,\theta})]}{\varsigma_{P_n}(\theta,\tau_{n,\theta},g_{n,\theta})} \\ &\overset{(2)}{\geq} \frac{1}{\Delta_\varsigma}\inf_{\theta\in\Theta}\sqrt{q_n}E_{P_n}[m(W_i,\theta,\tau_{n,\theta},g_{n,\theta})] \\ &\overset{(3)}{\geq} \frac{\sqrt{q_n}\mu_n}{\Delta_\varsigma}\inf_{\theta\in\Theta}E_{P_n}[g_{n,\theta}(X)] \\ &\overset{(4)}{\geq} \frac{\sqrt{q_n}\mu_n}{\Delta_\varsigma}\inf_{\theta\in\Theta}\int_{x\in\mathcal{X}_n(\theta)}g_{n,\theta}(x)dP_X \\ &\overset{(5)}{\geq} \frac{\sqrt{q_n}\mu_n\varepsilon_v}{\Delta_\varsigma}\inf_{\theta\in\Theta}Pr_{P_n}(\mathcal{X}_n(\theta)) \\ &\overset{(6)}{\geq} \frac{\sqrt{q_n}\mu_n\eta_v\varepsilon_v}{\Delta_\varsigma} > 0, \end{aligned} \tag{A.9}$$

where the third inequality holds by Assumption 3.4 and the law of iterated expectations, and the remaining inequalities follow from Assumption 3.4. Fix any $\delta > 0$ (the role of δ is clarified at the end of the proof). Now set $\kappa \in (0, \eta_v\varepsilon_v/2\Delta_\varsigma)$, set $c = M_2/\kappa(1+\delta)$, and note that for all $n \geq N$ (where N is from Assumption 3.4):

$$\begin{aligned} &(\Pr_{P_n} \times \mathbb{P}^\sharp)(T_{q_n}^{(r)} \leq M_2) \\ &= (\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\frac{1}{c}T_{q_n}^{(r)} \leq \frac{M_2}{c}\right) \end{aligned}$$

$$
\begin{aligned}
&\leq (\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\frac{1}{c}T_{q_n}^{(r)} \leq \frac{M_2}{c}, \frac{1}{c}T_n - \frac{1}{c(1+\delta)}T_{q_n}^{(r)} \leq \kappa\right) + (\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\frac{1}{c(1+\delta)}T_{q_n}^{(r)} - \frac{1}{c}T_n > \kappa\right) \\
&\leq (\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\frac{1+\delta}{c}T_n \leq \frac{M_2}{c} + \kappa(1+\delta)\right) + (\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\frac{1}{c(1+\delta)}T_{q_n}^{(r)} - \frac{1}{c}T_n > \kappa\right) \\
&= (\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\frac{1+\delta}{c}T_n \leq 2\kappa(1+\delta)\right) + (\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\frac{\kappa}{M_2}T_{q_n}^{(r)} - \frac{\kappa(1+\delta)}{M_2}T_n > \kappa\right) \\
&\leq 1\left\{\frac{\sqrt{q_n}\mu_n\eta_v\varepsilon_v}{c\Delta_\varsigma} \leq 2\kappa\right\} + (\Pr_{P_n} \times \mathbb{P}^\sharp)\left(T_{q_n}^{(r)} - (1+\delta)T_n > M_2\right),
\end{aligned}
$$

where the first inequality follows from the union bound, the second equality follows by taking $c = M_2/\kappa(1+\delta)$, and the last inequality follows from (A.9). By assumption $\sqrt{q_n}\mu_n \to C > c$. Taking N larger if necessary, we can assume $\sqrt{q_n}\mu_n > c$ for all $n \geq N$. In this case we have:

$$
1\left\{\frac{\sqrt{q_n}\mu_n\eta_v\varepsilon_v}{c\Delta_\varsigma} \leq 2\kappa\right\} = 1\left\{\frac{C\eta_v\varepsilon_v}{c\Delta_\varsigma} \leq 2\kappa\right\} \leq 1\left\{\frac{\eta_v\varepsilon_v}{\Delta_\varsigma} \leq 2\kappa\right\} = 0,
$$

which follows from our choice of $\kappa < \eta_v\varepsilon_v/2\Delta_\varsigma$. Thus, it remains only to show that:

$$
(\Pr_{P_n} \times \mathbb{P}^\sharp)\left(\left(\frac{T_{q_n}^{(r)}}{1+\delta} - T_n\right) > \frac{M_2}{(1+\delta)}\right) < \varepsilon.
$$

By Assumption 3.2 we have for any $M_2, \delta > 0$ (taking N larger if necessary):

$$
\begin{aligned}
&\Pr_{P_n}\left(\left(\frac{T_{q_n}^{(r)}}{1+\delta} - T_n\right) > \frac{M_2}{(1+\delta)}\right) \\
&\leq \Pr_{P_n}\left(\left(\frac{T_{q_n}^{(r)}}{1+\delta} - T_n\right) > \frac{M_2}{(1+\delta)}, \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_n}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} - 1\right| \leq \delta\right) \\
&\qquad + \Pr_{P_n}\left(\left(\frac{T_{q_n}^{(r)}}{1+\delta} - T_n\right) > \frac{M_2}{(1+\delta)}, \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_n}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} - 1\right| > \delta\right) \\
&\leq \Pr_{P_n}\left(\left(\frac{T_{q_n}^{(r)}}{1+\delta} - T_n\right) > \frac{M_2}{(1+\delta)}, \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_n}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} - 1\right| \leq \delta\right) \\
&\qquad + \Pr_{P_n}\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_n}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} - 1\right| > \delta\right) \\
&\leq \Pr_{P_n}\left(\left(\frac{T_{q_n}^{(r)}}{1+\delta} - T_n\right) > \frac{M_2}{(1+\delta)}, \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_n}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} - 1\right| \leq \delta\right) + \varepsilon/2 \\
&= \Pr_{P_n}\left(\left(\frac{1}{1+\delta}\right)\left(\inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_n}\frac{\sqrt{q_n}\bar{m}_{q_n}^{(r)}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)}\right)\right. \\
&\qquad \left. - \left(\inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\frac{\sqrt{q_n}E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)}\right) > \frac{M_2}{(1+\delta)}, \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_n}(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} - 1\right| \leq \delta\right) + \varepsilon/2 \\
&\leq \Pr_{P_n}\left(\left(\inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_n}\frac{\sqrt{q_n}\bar{m}_{q_n}^{(r)}(\theta,\tau,g)}{\varsigma_P(\theta,\tau,g)}\right) - \left(\inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\frac{\sqrt{q_n}E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_P(\theta,\tau,g)}\right) > M_2\right) + \varepsilon/2
\end{aligned}
$$

$$\leq \Pr_{P_n}\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\sqrt{q_n}(\bar{m}^{(r)}_{q_n}(\theta,\tau,g)-E_{P_n}[m(W_i,\theta,\tau,g)])}{\varsigma_P(\theta,\tau,g)}\right|>M_2\right)+\varepsilon/2$$

$$<\varepsilon,$$

where the second last inequality follows from the fact that $\mathcal{G}_n \subset \mathcal{G}$, and the last inequality follows from Lemma B.2, taking N larger if necessary. Combining the results above, the conclusion follows. ■

Proof of Theorem 3.3. Note that Lemmas B.4, B.5, and B.6 hold for any sequence $\{P_n \in \mathcal{P} : n \geq 1\}$. Thus, repeating an identical proof to the proof of Theorem 3.1, it suffices to show that:

$$\limsup_{n\to\infty}\Pr_{P_n}\left(A^c_{n,P_n}\right)=0,$$

where:

$$A^c_{n,P}:=\left\{\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_n}\frac{\sqrt{q_n}E_P[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_P(\hat{\theta}_n,\tau,g)}>\varepsilon\right\}.$$

(Note this does not follow from Lemma 3.1, which holds only for null sequences $\{P_n \in \mathcal{P}_0 : n \geq 1\}$). However, note that:

$$\begin{aligned}
\inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_n}\frac{\sqrt{q_n}E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} &\leq \inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\frac{\sqrt{q_n}E_P[m(W_i,\theta,\tau,g)]}{\varsigma_P(\theta,\tau,g)}\\
&\leq \frac{1}{\underline{\Delta}_\varsigma}\inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\sqrt{q_n}E_P[m(W_i,\theta,\tau,g)]\\
&\leq \frac{1}{\underline{\Delta}_\varsigma}\inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\sqrt{q_n}E_X[E_P[m(W_i,\theta,\tau)\mid X_i]g(X_i)]\\
&\leq \frac{\bar{G}}{\underline{\Delta}_\varsigma}\inf_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sqrt{q_n}E_X[E_P[m(W_i,\theta,\tau)\mid X_i]]\\
&\leq \frac{\bar{G}}{\underline{\Delta}_\varsigma}\sqrt{q_n}E_X\left[\sup_{\tau\in\mathcal{T}}E_P[m(W_i,\theta_n,\tau)\mid X_i]\right]\\
&\leq \frac{\bar{G}}{\underline{\Delta}_\varsigma}\sqrt{q_n}\mu_n\\
&\to 0.
\end{aligned}$$

Thus, the sequence $\{P_n \in \mathcal{P}\setminus\mathcal{P}_0\}_{n\geq 1}$ satisfies condition (B.1) from Lemma B.1. Applying Lemma B.1 and Lemma B.2, the result follows. ■

B Additional Results

Lemma B.1. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold. Fix any $\varepsilon > 0$ and any sequence $\{a_n\}_{n=1}^{\infty} \subset \mathbb{N}$, and let $\{P_n \in \mathcal{P}\}_{n\geq 1}$ be any sequence satisfying:*

$$\limsup_{n\to\infty} \mathbb{1}\left\{\inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} > \varepsilon\right\} = 0. \tag{B.1}$$

Then for any $M > 0$:

$$\limsup_{n\to\infty} Pr_{P_n}\left(\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} > M\right) \leq \sum_{j=1}^{3} \limsup_{n\to\infty} Pr_{P_n}\left(\widehat{C}_{n,P_n,j}\right),$$

where:

$$\widehat{C}_{n,P_n,j} := \left\{\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}_n} \left|\frac{a_n(\bar{m}_n(\theta,\tau,g) - E_{P_n}[m(W_i,\theta,\tau,g)])}{\varsigma_{P_n}(\theta,\tau,g)}\right| > M\cdot C_j\right\},$$

for $j = 1,2,3$, where each constant $C_j > 0$ can be chosen independent of n.

Remark B.1. *Under the null we have:*

$$\inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} \leq 0,$$

for every sequence $\{a_n\}_{n=1}^{\infty}$ and every $\mathcal{G}_n \subset \mathcal{G}$, so that (B.1) is trivially satisfied for every sequence $\{P_n \in \mathcal{P}_0\}_{n\geq 1}$, which makes the result useful in the proof of Lemma 3.1. However, (B.1) can also be satisfied by some alternative sequences, which makes the result useful for the proof of Theorem 3.3.

Proof of Lemma B.1. Define:

$$\begin{aligned}
\widehat{B}^{(0)}_{n,M,P_n} &:= \left\{\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} > M\right\} \\
\widehat{B}^{(1)}_{n,P_n} &:= \left\{\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \geq 0, \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \geq 0\right\}, \\
\widehat{B}^{(2)}_{n,P_n} &:= \left\{\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \geq 0, \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} < 0\right\}, \\
\widehat{B}^{(3)}_{n,P_n} &:= \left\{\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} < 0\right\}.
\end{aligned}$$

Then:

$$\limsup_{n\to\infty} \Pr_{P_n}\left(\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} > M\right) = \limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n}\right)$$

$$\leq \limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(1)}_{n,P_n}\right) + \limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(2)}_{n,P_n}\right) + \limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(3)}_{n,P_n}\right),$$

which follows from the union bound, and the fact that $\widehat{B}^{(1)}_{n,P_n}$, $\widehat{B}^{(2)}_{n,P_n}$ and $\widehat{B}^{(3)}_{n,P_n}$ exhaust all possible cases. Since $\varsigma_{P_n}(\theta,\tau,g) > 0 \ \forall(\theta,\tau,g)$ by Assumption 3.1, on the event $\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(3)}_{n,P_n}$ we have:

$$\begin{aligned}
&\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \\
&\overset{(1)}{<} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \\
&\leq \left| \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \right| \\
&\overset{(2)}{\leq} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \right| \\
&\leq \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} - \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \right|,
\end{aligned}$$

where (1) follows since we are on the event $\widehat{B}^{(3)}_{n,P_n}$, and (2) follows from (A.4). Thus we have $\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(3)}_{n,P_n} \subset \widehat{C}_{n,P_n,1}$, where:

$$\widehat{C}_{n,P_n,1} := \left\{ \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n(\bar{m}_n(\theta,\tau,g) - E_{P_n}[m(W_i,\theta,\tau,g)])}{\varsigma_{P_n}(\theta,\tau,g)} \right| > M \right\}.$$

Now consider the events $\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(1)}_{n,P_n}$ and $\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(2)}_{n,P_n}$. Along the sequence $\{P_n \in \mathcal{P}\}_{n\geq 1}$ we have:

$$\limsup_{n\to\infty} \mathbb{1}\left\{ \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} > \varepsilon \right\} = 0.$$

Now define:

$$\widehat{B}^{(4)}_{n,\varepsilon,P_n} := \left\{ \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} \leq \varepsilon \right\}.$$

Then:

$$\begin{aligned}
\limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(1)}_{n,P_n}\right) &= \limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(1)}_{n,P_n} \cap \widehat{B}^{(4)}_{n,\varepsilon,P_n}\right), \\
\limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(2)}_{n,P_n}\right) &= \limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(2)}_{n,P_n} \cap \widehat{B}^{(4)}_{n,\varepsilon,P_n}\right).
\end{aligned}$$

Now fix any $\delta \in (0,1)$, and consider the event:

$$F_{n,\delta,P} := \left\{ \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \sup_{g\in\mathcal{G}} \left| \frac{\varsigma_P(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} - 1 \right| > \delta \right\}.$$

Note that, on $F^c_{n,\delta,P}$, we have:

$$1-\delta \leq \frac{\varsigma_P(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} \leq 1+\delta, \qquad \forall(\theta,\tau,g) \in \Theta\times\mathcal{T}\times\mathcal{G},$$

which implies:

$$\frac{1}{\varsigma_P(\theta,\tau,g)} \leq \frac{(1-\delta)^{-1}}{\hat{\varsigma}_n(\theta,\tau,g)}, \qquad \forall(\theta,\tau,g) \in \Theta\times\mathcal{T}\times\mathcal{G}, \tag{B.2}$$

$$\frac{1}{\hat{\varsigma}_n(\theta,\tau,g)} \leq \frac{(1+\delta)}{\varsigma_P(\theta,\tau,g)}, \qquad \forall(\theta,\tau,g) \in \Theta\times\mathcal{T}\times\mathcal{G}. \tag{B.3}$$

By Assumption 3.2 we have:

$$\limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(1)}_{n,P_n} \cap \widehat{B}^{(4)}_{n,M,P_n}\right) = \limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(1)}_{n,P_n} \cap \widehat{B}^{(4)}_{n,M,P_n} \cap F^c_{n,\delta,P_n}\right),$$

$$\limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(2)}_{n,P_n} \cap \widehat{B}^{(4)}_{n,M,P_n}\right) = \limsup_{n\to\infty} \Pr_{P_n}\left(\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(2)}_{n,P_n} \cap \widehat{B}^{(4)}_{n,M,P_n} \cap F^c_{n,\delta,P_n}\right).$$

Now on $\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(1)}_{n,P_n} \cap \widehat{B}^{(4)}_{n,M,P_n} \cap F^c_{n,\delta,P_n}$ we have:

$$\begin{aligned}
&\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \\
&\leq \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \varepsilon \\
&= \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \\
&\qquad + \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \varepsilon \\
&\overset{(1)}{\leq} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \\
&\qquad + (1-\delta)^{-1} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\hat{\varsigma}_n(\hat{\theta}_n,\tau,g)} \\
&\qquad\qquad - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \varepsilon \\
&\overset{(2)}{\leq} \sup_{\theta\in\Theta} \left| \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} - \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \right|
\end{aligned}$$

$$
\begin{aligned}
&+ (1-\delta)^{-1} \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} + (1-\delta)^{-1}\epsilon_n \\
&\qquad - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \varepsilon \\
&\overset{(3)}{\leq} \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} - \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \right| \\
&\qquad + \left(\frac{(1+\delta)}{(1-\delta)}\right) \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} + (1-\delta)^{-1}\epsilon_n \\
&\qquad - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \varepsilon \\
&\overset{(4)}{\leq} \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} - \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \right| \\
&\qquad + \left(\frac{(1+\delta)}{(1-\delta)}\right) \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} + (1-\delta)^{-1}\epsilon_n \\
&\qquad - \left(\frac{(1+\delta)}{(1-\delta)}\right) \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \left(\frac{(1+\delta)}{(1-\delta)}\right)\varepsilon \\
&\overset{(5)}{\leq} \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} - \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \right| \\
&\qquad + \left(\frac{(1+\delta)}{(1-\delta)}\right) \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} - \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} \right| \\
&\qquad + (1-\delta)^{-1}\epsilon_n + \left(\frac{(1+\delta)}{(1-\delta)}\right)\varepsilon \\
&\overset{(6)}{\leq} \frac{(2+\delta)}{(1-\delta)} \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} - \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \right| + \frac{\epsilon_n}{1-\delta} + \left(\frac{(1+\delta)}{(1-\delta)}\right)\varepsilon,
\end{aligned}
$$

where (1) follows from (B.2) and the fact that we are on event $\widehat{B}^{(1)}_{n,P_n}$, (2) follows from the definition of $\hat{\theta}_n$ from (2.6), (3) follows from (A.4) and from (B.3) coupled with the fact that we are on event $\widehat{B}^{(1)}_{n,P_n}$, (4) follows from the fact that $P \in \mathcal{P}_0$ (implying the last term is negative), (5) follows from (A.5), and (6) follows after collecting terms. Thus we have $\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(1)}_{n,P_n} \cap \widehat{B}^{(4)}_{n,M,P_n} \cap F^c_{n,\delta,P_n} \subset \widehat{C}_{n,P_n,2}$, where:

$$
\widehat{C}_{n,P_n,2} := \left\{ \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n(\bar{m}_n(\theta,\tau,g) - E_{P_n}[m(W_i,\theta,\tau,g)])}{\varsigma_{P_n}(\theta,\tau,g)} \right| + \left(\frac{1}{2+\delta}\right)\epsilon_n + \left(\frac{(1+\delta)}{(2+\delta)}\right)\varepsilon > \frac{M(1-\delta)}{(2+\delta)} \right\}.
$$

Similarly, on $\widehat{B}^{(0)}_{n,M,P_n} \cap \widehat{B}^{(2)}_{n,P_n} \cap \widehat{B}^{(4)}_{n,M,P_n} \cap F^c_{n,\delta,P_n}$:

$$
\begin{aligned}
&\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \\
&\leq \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \varepsilon \\
&= \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)}
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \varepsilon \\
&\overset{(1)}{\leq} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\hat{\theta}_n,\tau,g)]}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} - \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\varsigma_{P_n}(\hat{\theta}_n,\tau,g)} \\
&\quad + (1-\delta)^{-1} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\hat{\theta}_n,\tau,g)}{\hat{\varsigma}_n(\hat{\theta}_n,\tau,g)} \\
&\quad - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \varepsilon \\
&\overset{(2)}{\leq} \sup_{\theta\in\Theta} \left| \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} - \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \right| \\
&\quad + (1-\delta)^{-1} \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} + (1-\delta)^{-1}\epsilon_n \\
&\quad - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} + \varepsilon \\
&\overset{(3)}{\leq} \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} - \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \right| \\
&\quad + \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} + (1-\delta)^{-1}\epsilon_n \\
&\quad - \inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\hat{\varsigma}_n(\theta,\tau,g)} + \varepsilon \\
&\overset{(4)}{\leq} 2 \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n E_{P_n}[m(W_i,\theta,\tau,g)]}{\varsigma_{P_n}(\theta,\tau,g)} - \frac{a_n \bar{m}_n(\theta,\tau,g)}{\varsigma_{P_n}(\theta,\tau,g)} \right| + (1-\delta)^{-1}\epsilon_n + \varepsilon,
\end{aligned}
$$

where (1) follows from (B.2) and the fact that we are on event $\widehat{B}^{(2)}_{n,P_n}$, (2) follows from the definition of $\hat{\theta}_n$ from (2.6), (3) follows from the fact that we are on $\widehat{B}^{(2)}_{n,P_n}$, which implies:

$$
\inf_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \frac{a_n \bar{m}_n(\theta,\tau,g)}{\hat{\varsigma}_n(\theta,\tau,g)} \leq 0,
$$

and (4) follows from (A.5) and collecting terms. Thus we have $\widehat{B}^{(4)}_{n,\varepsilon,P_n} \cap \widehat{B}^{(2)}_{n,P_n} \cap F^c_{n,\delta,P_n} \subset \widehat{C}_{n,P_n,3}$, where:

$$
\widehat{C}_{n,P_n,3} := \left\{ \sup_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left| \frac{a_n(\bar{m}_n(\theta,\tau,g) - E_{P_n}[m(W_i,\theta,\tau,g)])}{\varsigma_{P_n}(\theta,\tau,g)} \right| + \frac{1}{2(1-\delta)}\epsilon_n + \frac{\varepsilon}{2} > \frac{M}{2} \right\}.
$$

This completes the proof. ■

Lemma B.2. *Suppose that Assumption 3.1 holds. Then:*

$$
\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} Pr_P\left(B_{n,M,P}\right) = 0,
$$

where:

$$B_{n,M,P} := \left\{ \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}} \left| \frac{\sqrt{n}\left(\bar{m}_n(\theta,\tau,g) - E_P[m(W_i,\theta,\tau,g)]\right)}{\varsigma_P(\theta,\tau,g)} \right| > M \right\}.$$

Proof of Lemma B.2. Define:

$$||v_{n,P}||_{\mathbb{T}} := \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}} \left| \frac{\sqrt{n}(\bar{m}_n(\theta,\tau,g) - E_P\left[m(W_i,\theta,\tau,g)\right])}{\varsigma_P(\theta,\tau,g)} \right|.$$

It suffices to show that for any $\varepsilon > 0$ there exists an $M > 0$ and an N such that:

$$\sup_{n\geq N}\sup_{P\in\mathcal{P}} \Pr_P\left(||v_{n,P}||_{\mathbb{T}} > M\right) < \varepsilon. \tag{B.4}$$

By Markov's inequality, we have:

$$\sup_{P\in\mathcal{P}} \Pr_P\left(||v_{n,P}||_{\mathbb{T}} > M\right) \leq \sup_{P\in\mathcal{P}} \frac{E_P\,||v_{n,P}||_{\mathbb{T}}}{M}. \tag{B.5}$$

The remainder of the proof will show that $E_P\,||v_{n,P}||_{\mathbb{T}}$ is bounded above by a constant that is independent of both n and P. To this end, note that under Assumptions 3.1, the class $\mathcal{M}_P$ is pointwise measurable and satisfies Dudley's entropy condition for the envelope $\boldsymbol{M}$ uniformly in $P \in \mathcal{P}$. In particular:

$$\begin{aligned}\sup_{P\in\mathcal{P}} \int_0^\infty \sup_{Q\in\mathcal{Q}} \sqrt{\log N(\epsilon\cdot||\boldsymbol{M}||_{Q,2}, \mathcal{M}_P, L_2(Q))}\, d\epsilon \\ = \sup_{P\in\mathcal{P}} \int_0^1 \sup_{Q\in\mathcal{Q}} \sqrt{\log N(\epsilon\cdot||\boldsymbol{M}||_{Q,2}, \mathcal{M}_P, L_2(Q))}\, d\epsilon < \infty,\end{aligned} \tag{B.6}$$

with the inner supremum taken over all probability measures with finite support. Now define:

$$J(\delta, \mathcal{M}_P) := \sup_{Q\in\mathcal{Q}} \int_0^\delta \sqrt{1 + \log N(\epsilon\cdot||\boldsymbol{M}||_{Q,2}, \mathcal{M}_P, ||\cdot||_{Q,2})}\, d\epsilon.$$

Then (B.6) implies that $\sup_{P\in\mathcal{P}} J(1,\mathcal{M}_P) < \infty$. By Theorem 2.14.1 in van der Vaart and Wellner (1996), we have:

$$E_P\,||v_{n,P}||_{\mathbb{T}} \leq C' J(1,\mathcal{M}_P)||\boldsymbol{M}||_{P,2}, \tag{B.7}$$

for some finite constant C'. By Assumption 3.1, there exists an $\eta > 0$ such that $\sup_{P\in\mathcal{P}} E_P[\boldsymbol{M}^{2+\eta}] \leq C''$ for some $C'' < \infty$, so that by Hölder's inequality $\sup_{P\in\mathcal{P}} ||\boldsymbol{M}||_{P,2} \leq \sup_{P\in\mathcal{P}} (E_P[\boldsymbol{M}^{2+\eta}])^{2/(2+\eta)} \leq (C'')^{2/(2+\eta)}$. Thus, conclude that $\sup_{P\in\mathcal{P}} C' J(1,\mathcal{M}_P)||\boldsymbol{M}||_{P,2} \leq \sup_{P\in\mathcal{P}} C' J(1,\mathcal{M}_P)(C'')^{2/(2+\eta)} < \infty$. Combining this with (B.4), (B.5) and (B.7) completes the proof. ■

The following Lemma is similar to Theorem 1 in Andrews and Shi (2013) and Theorem D.3 in Andrews and Shi (2017), although adapted to accommodate our infimum test statistic.

Lemma B.3. *Suppose Assumptions 3.1, 3.2 and 3.3 hold, let $\{P_n \in \mathcal{P} : n \geq 1\}$ be any sequence, and let $\{a_n\}_{n\geq 1}$ be any subsequence along which $h_{2,P_{a_n}} \overset{u}{\to} h_{2,P_0}$ and $h_{3,P_{a_n}} \overset{u}{\to} h_{3,P_0}$ for some $P_0 \in \mathcal{P}$. Then for all $x_{a_n} := x_{a_n}(\omega) \in \mathbb{R}$ and $\delta > 0$ we have:*

$$\limsup_{n\to\infty} \left[Pr_{P_{a_n}}(T^{(r)}_{q_{a_n}}(\hat{\theta}_{a_n}) > x_{a_n}) - Pr_{P_{a_n}}(T^{(r)}_0(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}) + \delta > x_{a_n}) \right] \leq 0,$$

where:

$$T^{(r)}_0(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}) := \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_{a_n}} \left\{ v_{P_0}(\hat{\theta}_{a_n}, \tau, g) + h_{1,q_{a_n},P_{a_n}}(\hat{\theta}_{a_n}, \tau, g) \right\}, \quad \text{(B.8)}$$

where v_{P_0} is a real-valued tight normalized Gaussian process on $\mathbb{T}$ with covariance kernel h_{2,P_0}, and where $h_{1,q_{a_n},P_{a_n}}(\theta, \tau, g)$ is as defined in (A.1).

Proof of Lemma B.3. There exists a further subsequence $\{b_n\}_{n\geq 1} \subset \{a_n\}_{n\geq 1}$ such that:

$$\begin{aligned}
&\limsup_{n\to\infty} \left[\Pr_{P_{a_n}}(T^{(r)}_{q_{a_n}}(\hat{\theta}_{a_n}) > x_{a_n}) - \Pr_{P_{a_n}}(T^{(r)}_0(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}) + \delta > x_{a_n}) \right] \\
&= \lim_{n\to\infty} \left[\Pr_{P_{b_n}}(T^{(r)}_{q_{b_n}}(\hat{\theta}_{b_n}) > x_{b_n}) - \Pr_{P_{b_n}}(T^{(r)}_0(\hat{\theta}_{b_n}, h_{1,q_{b_n},P_{b_n}}, h_{2,P_0}, h_{3,P_0}) + \delta > x_{b_n}) \right].
\end{aligned}$$

We continue the proof along this subsequence. Note that:

$$\begin{aligned}
&\sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_{b_n}} \left(\frac{\sqrt{q_{b_n}} \bar{m}^{(r)}_{q_{b_n}}(\theta, \tau, g)}{\hat{\varsigma}_{b_n}(\theta, \tau, g)} \right) \\
&= \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_{b_n}} \left(\frac{\sqrt{q_{b_n}} \left(\bar{m}^{(r)}_{q_{b_n}}(\theta, \tau, g) - E_{P_{b_n}}[m(W_i, \theta, \tau, g)] \right)}{\hat{\varsigma}_{b_n}(\theta, \tau, g)} + \frac{\sqrt{q_{b_n}} E_{P_{b_n}}[m(W_i, \theta, \tau, g)]}{\hat{\varsigma}_{b_n}(\theta, \tau, g)} \right) \\
&= \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_{b_n}} \left(\frac{\varsigma_{P_{b_n}}(\theta, \tau, g)}{\hat{\varsigma}_{b_n}(\theta, \tau, g)} \right) \left(\frac{\sqrt{q_{b_n}} \left(\bar{m}^{(r)}_{q_{b_n}}(\theta, \tau, g) - E_{P_{b_n}}[m(W_i, \theta, \tau, g)] \right)}{\varsigma_{P_{b_n}}(\theta, \tau, g)} \right. \\
&\qquad \left. + \frac{\sqrt{q_{b_n}} E_{P_{b_n}}[m(W_i, \theta, \tau, g)]}{\varsigma_{P_{b_n}}(\theta, \tau, g)} \right) \\
&= \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_{b_n}} \left(\frac{\varsigma_{P_{b_n}}(\theta, \tau, g)}{\hat{\varsigma}_{b_n}(\theta, \tau, g)} \right) \left(v^{(r)}_{q_{b_n},P_{b_n}}(\theta, \tau, g) + h_{1,q_{b_n},P_{b_n}}(\theta, \tau, g) \right),
\end{aligned}$$

where:

$$v^{(r)}_{q_{b_n},P_{b_n}}(\theta, \tau, g) := \frac{\sqrt{q_{b_n}} \left(\bar{m}^{(r)}_{q_{b_n}}(\theta, \tau, g) - E_{P_{b_n}}[m(W_i, \theta, \tau, g)] \right)}{\varsigma_{P_{b_n}}(\theta, \tau, g)}.$$

By Lemma S.2.1 in the Online Supplementary Material we have $v^{(r)}_{q_{b_n},P_{b_n}} \rightsquigarrow v_{P_0}$ in $\ell^\infty(\mathbb{T})$ where v_{P_0} is a tight Gaussian process with covariance kernel h_{2,P_0} with almost all sample paths bounded and in $UC(\mathbb{T}, \rho_{P_0})$, the space of uniformly continuous real-valued functions on $(\mathbb{T}, \rho_{P_0})$ equipped with the sup norm. It follows from Assumption 3.2 and Slutsky's Theorem (e.g. Kosorok (2007)

Theorem 7.15) that:

$$\overline{v}^{(r)}_{q_{b_n},P_{b_n}} := (\varsigma_{P_{b_n}}(\theta,\tau,g)/\hat{\varsigma}_{b_n}(\theta,\tau,g))v^{(r)}_{q_{b_n},P_{b_n}} \rightsquigarrow v_{P_0},$$

in $\ell^\infty(\mathbb{T})$ as $n \to \infty$. By the almost-sure representation theorem (e.g. Dudley (2014) Theorem 3.24) there exists a probability space $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{\mathbb{P}})$ and perfect measurable functions s_{b_n} from $(\tilde{\Omega}, \tilde{\mathfrak{F}})$ to $(\Omega, \mathfrak{F})$ for each $n = 0, 1, \ldots$ such that $\tilde{\mathbb{P}} \circ s_{b_n}^{-1} = \mathbb{P}$ on $\mathfrak{F}$ for each n and:

$$\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}||\tilde{v}_{b_n}(\tilde{\omega})(\theta,\tau,g) - \tilde{v}(\tilde{\omega})(\theta,\tau,g)||\right)^* \to 0 \text{ almost surely as } n \to \infty, \tag{B.9}$$

where $\tilde{v}_{b_n}(\tilde{\omega})(\,\cdot\,) := \overline{v}^{(r)}_{q_{b_n},P_{b_n}}(s_{b_n}(\tilde{\omega}))(\,\cdot\,)$ and $\tilde{v}(\tilde{\omega}) := v_{P_0}(s_{b_0}(\tilde{\omega}))(\,\cdot\,)$ for $\tilde{\omega} \in \tilde{\Omega}$, and where $(\,\cdot\,)^*$ denotes the measurable majorant. Now define $\tilde{\theta}_{b_n}(\tilde{\omega}) := \hat{\theta}_{b_n} \circ s_{b_n}(\tilde{\omega})$ and $\tilde{x}_{b_n}(\tilde{\omega}) := x_{b_n} \circ s_{b_n}(\tilde{\omega}) \in \mathbb{R}$, and set:

$$\begin{aligned}
\tilde{T}_{b_n}(\tilde{\omega}) &:= \sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}}\left(\tilde{v}_{b_n}(\tilde{\omega})(\tilde{\theta}_{b_n}(\tilde{\omega}),\tau,g) + h_{1,q_{b_n},P_{b_n}}(\tilde{\theta}_{b_n}(\tilde{\omega}),\tau,g)\right),\\
\tilde{T}_{0,b_n}(\tilde{\omega}) &:= \sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}}\left(\tilde{v}(\tilde{\omega})(\tilde{\theta}_{b_n}(\tilde{\omega}),\tau,g) + h_{1,q_{b_n},P_{b_n}}(\tilde{\theta}_{b_n}(\tilde{\omega}),\tau,g)\right).
\end{aligned}$$

By construction, $\tilde{T}_{b_n}(\tilde{\omega})$ and $T^{(r)}_{b_n}(\hat{\theta}_{b_n})$ are identically distributed and $\tilde{T}_{0,b_n}(\tilde{\omega})$ and $T^{(r)}_0(\hat{\theta}_{b_n}, h_{1,q_{b_n},P_{b_n}}, h_{2,P_0}, h_{3,P_0})$ are identically distributed. Thus it suffices to prove:

$$A := \lim_{n\to\infty}\left[\tilde{\mathbb{P}}\left(\tilde{T}_{b_n}(\tilde{\omega}) > \tilde{x}_{b_n}(\tilde{\omega})\right) - \tilde{\mathbb{P}}\left(\tilde{T}_{0,b_n}(\tilde{\omega}) + \delta > \tilde{x}_{b_n}(\tilde{\omega})\right)\right] \le 0.$$

We have:

$$\begin{aligned}
&\left|\tilde{T}_{b_n}(\tilde{\omega}) - \tilde{T}_{0,b_n}(\tilde{\omega})\right|\\
&= \Bigg|\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}}\left\{\tilde{v}_{b_n}(\tilde{\omega})(\tilde{\theta}_{b_n}(\tilde{\omega}),\tau,g) + h_{1,q_{b_n},P_{b_n}}(\tilde{\theta}_{b_n}(\tilde{\omega}),\tau,g)\right\}\\
&\qquad\qquad - \sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}}\left\{\tilde{v}(\tilde{\omega})(\tilde{\theta}_{b_n}(\tilde{\omega}),\tau,g) + h_{1,q_{b_n},P_{b_n}}(\tilde{\theta}_{b_n}(\tilde{\omega}),\tau,g)\right\}\Bigg|\\
&\le \sup_{\theta\in\Theta}\Bigg|\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}}\left\{\tilde{v}_{b_n}(\tilde{\omega})(\theta,\tau,g) + h_{1,q_{b_n},P_{b_n}}(\theta,\tau,g)\right\}\\
&\qquad\qquad - \sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}}\left\{\tilde{v}(\tilde{\omega})(\theta,\tau,g) + h_{1,q_{b_n},P_{b_n}}(\theta,\tau,g)\right\}\Bigg|\\
&\le \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}}\left|\tilde{v}_{b_n}(\tilde{\omega})(\theta,\tau,g) - \tilde{v}(\tilde{\omega})(\theta,\tau,g)\right|\\
&\le \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\tilde{v}_{b_n}(\tilde{\omega})(\theta,\tau,g) - \tilde{v}(\tilde{\omega})(\theta,\tau,g)\right|\\
&\le \left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\tilde{v}_{b_n}(\tilde{\omega})(\theta,\tau,g) - \tilde{v}(\tilde{\omega})(\theta,\tau,g)\right|\right)^*
\end{aligned}$$

$$\longrightarrow 0,$$

almost surely as $n \to \infty$ by (B.9). Conclude that there exists a measurable sequence $B_{b_n}(\tilde{\omega})$ such that $|\tilde{T}_{b_n}(\tilde{\omega}) - \tilde{T}_{0,b_n}(\tilde{\omega})| \leq B_{b_n}(\tilde{\omega}) \to 0$ almost surely as $n \to \infty$. Now define:

$$\begin{aligned}\tilde{\Delta}_{b_n}(\tilde{\omega}) &= \mathbb{1}\{\tilde{T}_{b_n}(\tilde{\omega}) > \tilde{x}_{b_n}(\tilde{\omega})\} - \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + \delta > \tilde{x}_{b_n}(\tilde{\omega})\} \\ &= \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + \tilde{T}_{b_n}(\tilde{\omega}) - \tilde{T}_{0,b_n}(\tilde{\omega}) > \tilde{x}_{b_n}(\tilde{\omega})\} - \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + \delta > \tilde{x}_{b_n}(\tilde{\omega})\}.\end{aligned}$$

Note that:

$$\tilde{\Delta}_{b_n}(\tilde{\omega}) \leq (\tilde{\Delta}_{b_n}(\tilde{\omega}))^* \leq \tilde{\Delta}^{\dagger}_{b_n}(\tilde{\omega}) := \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + B_{b_n}(\tilde{\omega}) > \tilde{x}_{b_n}(\tilde{\omega})\} - \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + \delta > \tilde{x}_{b_n}(\tilde{\omega})\}.$$

Furthermore, let $\tilde{\Delta}^{+}_{b_n}(\tilde{\omega}) := \max\{\tilde{\Delta}^{\dagger}_{b_n}(\tilde{\omega}), 0\}$, and note that $\tilde{\Delta}^{\dagger}_{b_n}(\tilde{\omega}) \leq \tilde{\Delta}^{+}_{b_n}(\tilde{\omega})$. Since $B_{b_n}(\tilde{\omega}) \to 0$ a.s., $\tilde{\Delta}^{+}_{b_n}(\tilde{\omega}) \to 0$ a.s. Let $\tilde{\mathbb{E}}$ denote the expectation with respect to $\tilde{\mathbb{P}}$. Then by dominated convergence we have:

$$A = \limsup_{n\to\infty} \tilde{\mathbb{E}}^*[\tilde{\Delta}_{b_n}] \leq \limsup_{n\to\infty} \tilde{\mathbb{E}}[\tilde{\Delta}^{\dagger}_{b_n}] \leq \limsup_{n\to\infty} \tilde{\mathbb{E}}[\tilde{\Delta}^{+}_{b_n}] = 0.$$

This completes the proof.

■

Let $c_0^{(r)}(\theta, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha)$ denote the $1-\alpha$ quantile of $T_0^{(r)}(\theta, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0})$, where the statistic $T_0^{(r)}(\theta, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0})$ is as defined in (B.8). The following result is a direct consequence of Lemma B.3 and is similar to Lemma A2 in Andrews and Shi (2013).

Lemma B.4. *Suppose Assumptions 3.1, 3.2 and 3.3 hold, let $\{P_n \in \mathcal{P} : n \geq 1\}$ be any sequence, and let $\{a_n\}_{n\geq 1}$ be any subsequence along which $h_{2,P_{a_n}} \xrightarrow{u} h_{2,P_0}$ and $h_{3,P_{a_n}} \xrightarrow{u} h_{3,P_0}$ for some $P_0 \in \mathcal{P}$. Then for every $\delta > 0$:*

$$\limsup_{n\to\infty} Pr_{P_{a_n}}(T^{(r)}_{a_n}(\hat{\theta}_{a_n}) > c_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta) \leq \alpha.$$

Proof of Lemma B.4. Note that for every $\delta > 0$ we have:

$$\begin{aligned}&\limsup_{n\to\infty} \Pr_{P_{a_n}}(T^{(r)}_{a_n}(\hat{\theta}_{a_n}) > c_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta) \\ &= \limsup_{n\to\infty} \Pr_{P_{a_n}}(T_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}) > c_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha)) \\ &\quad + \limsup_{n\to\infty} \Big(\Pr_{P_{a_n}}(T^{(r)}_{a_n}(\hat{\theta}_{a_n}) > c_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta) \\ &\qquad - \Pr_{P_{a_n}}(T_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}) > c_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha)) \Big) \\ &\leq \alpha + 0,\end{aligned}$$

where the last inequality follows from Lemma B.3, taking $x_{a_n} = c_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta$, and from the fact that $c_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0}, 1-\alpha)$ is the $1-\alpha$ quantile of $T_0^{(r)}(\hat{\theta}_{a_n}, h_{1,q_{a_n},P_{a_n}}, h_{2,P_0}, h_{3,P_0})$, by definition.

■

The following result is similar to Lemma D.4 in Andrews and Shi (2017).

Lemma B.5. *Suppose Assumptions 3.1, 3.2 and 3.3 hold, let $\{P_n \in \mathcal{P} : n \geq 1\}$ be any sequence, and let $\{a_n\}_{n\geq 1}$ be any subsequence along which $h_{2,P_{a_n}} \overset{u}{\to} h_{2,P_0}$ and $h_{3,P_{a_n}} \overset{u}{\to} h_{3,P_0}$ for some $P_0 \in \mathcal{P}$. Then for every $\delta \in (0,\eta)$:*

$$\limsup_{n\to\infty}(Pr_{P_{a_n}} \times \mathbb{P}^\sharp)(\hat{c}_{a_n}^{(r)\sharp}(1-\alpha+\eta) \leq c_0^{(r)}(\hat{\theta}_{a_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta - \eta) = 0,$$

where $\hat{c}_{a_n}^{(r)\sharp}(1-\alpha+\eta)$ is the $1-\alpha+\eta$ quantile of the distribution of $T_{q_{a_n}}^{(r)\sharp}(\hat{\theta}_{a_n})$, and where $c_0^{(r)}(\hat{\theta}_{a_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha)$ is the $1-\alpha$ quantile of the distribution of:

$$T_0^{(r)}(\hat{\theta}_{a_n}, 0, h_{2,P_0}, h_{3,P_0}) = \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_{a_n}} v_{P_0}(\hat{\theta}_{a_n}, \tau, g),$$

where v_{P_0} is a mean zero Gaussian process on $\mathbb{T}$ with covariance kernel h_{2,P_0}.

Proof of Lemma B.5. Lemma S.2.2 implies that for any $\varepsilon > 0$:

$$\limsup_{n\to\infty} \Pr_{P_{a_n}}\left(d_{BL_1}(v_{q_{a_n},P_{a_n}}^{(r)\sharp}, v_{P_0})^* > \varepsilon\right) = 0, \tag{B.10}$$

where $v_{q_{a_n},P_{a_n}}^{(r)\sharp}$ is defined in (A.3), and where $v_{P_0} = v_{P_0}(\omega^\sharp, \omega)$ is a tight Gaussian-process defined on the probability space $(\Omega, \mathfrak{F}, \mathbb{P}) \times (\Omega^\sharp, \mathfrak{F}^\sharp, \mathbb{P}^\sharp)$ that is constant in its $\omega^\sharp-$coordinate. Now note that:

$$v_{q_{a_n}}^{(r)\sharp}(\theta,\tau,g) = \varsigma_{P_{a_n}}(\theta,\tau,g)\hat{\varsigma}_{a_n}^{-1}(\theta,\tau,g)v_{q_{a_n},P_{a_n}}^{(r)\sharp}(\theta,\tau,g),$$

where $v_{q_{a_n}}^{(r)\sharp}$ is defined in (A.2). Define the shorthand notation:

$$\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right) v_{q_{a_n},P_{a_n}}^{(r)\sharp} := \varsigma_{P_{a_n}}(\theta,\tau,g)\hat{\varsigma}_{a_n}^{-1}(\theta,\tau,g)v_{q_{a_n},P_{a_n}}^{(r)\sharp}(\theta,\tau,g).$$

By Assumption 3.2 we have for any $\varepsilon' > 0$:

$$\begin{aligned}
&\limsup_{n\to\infty} \Pr_{P_{a_n}}\left(d_{BL_1}\left(v_{q_{a_n}}^{(r)\sharp}, v_{P_0}\right)^* > \varepsilon\right) \\
&= \limsup_{n\to\infty} \Pr_{P_{a_n}}\left(d_{BL_1}\left(\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right) v_{q_{a_n},P_{a_n}}^{(r)\sharp}, v_{P_0}\right)^* > \varepsilon\right) \\
&= \limsup_{n\to\infty} \Pr_{P_{a_n}}\left(d_{BL_1}\left(\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right) v_{q_{a_n},P_{a_n}}^{(r)\sharp}, v_{P_0}\right)^* > \varepsilon,\ \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_{a_n}}(\theta,\tau,g)}{\hat{\varsigma}_{a_n}(\theta,\tau,g)} - 1\right| > \varepsilon'\right)
\end{aligned}$$

$$+\limsup_{n\to\infty}\Pr_{P_{a_n}}\left(d_{BL_1}\left(\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right)v^{(r)\sharp}_{q_{a_n},P_{a_n}},v_{P_0}\right)^*>\varepsilon,\ \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_{a_n}}(\theta,\tau,g)}{\hat{\varsigma}_{a_n}(\theta,\tau,g)}-1\right|\leq\varepsilon'\right)$$

$$\leq\limsup_{n\to\infty}\Pr_{P_{a_n}}\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_{a_n}}(\theta,\tau,g)}{\hat{\varsigma}_{a_n}(\theta,\tau,g)}-1\right|>\varepsilon'\right)$$
$$+\limsup_{n\to\infty}\Pr_{P_{a_n}}\left(d_{BL_1}\left(\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right)v^{(r)\sharp}_{q_{a_n},P_{a_n}},v_{P_0}\right)^*>\varepsilon,\ \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_{a_n}}(\theta,\tau,g)}{\hat{\varsigma}_{a_n}(\theta,\tau,g)}-1\right|\leq\varepsilon'\right)$$

$$=\limsup_{n\to\infty}\Pr_{P_{a_n}}\left(d_{BL_1}\left(\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right)v^{(r)\sharp}_{q_{a_n},P_{a_n}},v_{P_0}\right)^*>\varepsilon,\ \sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_{a_n}}(\theta,\tau,g)}{\hat{\varsigma}_{a_n}(\theta,\tau,g)}-1\right|\leq\varepsilon'\right). \quad \text{(B.11)}$$

Furthermore:

$$\begin{aligned}
&d_{BL_1}\left(\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right)v^{(r)\sharp}_{q_{a_n},P_{a_n}},v_{P_0}\right)\\
&=\sup_{h\in BL_1(\ell^\infty(\mathbb{T}))}\left|\mathbb{E}^\sharp h\left(\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right)v^{(r)\sharp}_{q_{a_n},P_{a_n}}\right)-E_{P_{a_n}}h(v_{P_0})\right|\\
&\leq\sup_{h\in BL_1(\ell^\infty(\mathbb{T}))}\left|\mathbb{E}^\sharp h\left(\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right)v^{(r)\sharp}_{q_{a_n},P_{a_n}}\right)-\mathbb{E}^\sharp h(v^{(r)\sharp}_{q_{a_n},P_{a_n}})\right|\\
&\qquad+\sup_{h\in BL_1(\ell^\infty(\mathbb{T}))}\left|\mathbb{E}^\sharp h(v^{(r)\sharp}_{q_{a_n},P_{a_n}})-E_{P_{a_n}}h(v_{P_0})\right|\\
&\leq\mathbb{E}^\sharp\left[\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\left(\frac{\varsigma_{P_{a_n}}(\theta,\tau,g)}{\hat{\varsigma}_{a_n}(\theta,\tau,g)}\right)v^{(r)\sharp}_{q_{a_n},P_{a_n}}(\theta,\tau,g)-v^{(r)\sharp}_{q_{a_n},P_{a_n}}(\theta,\tau,g)\right|\right]\\
&\qquad+\sup_{h\in BL_1(\ell^\infty(\mathbb{T}))}\left|\mathbb{E}^\sharp h(v^{(r)\sharp}_{q_{a_n},P_{a_n}})-E_{P_{a_n}}h(v_{P_0})\right|\\
&\leq\mathbb{E}^\sharp\left[\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_{a_n}}(\theta,\tau,g)}{\hat{\varsigma}_{a_n}(\theta,\tau,g)}-1\right|\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_{a_n},P_{a_n}}(\theta,\tau,g)\right|\right]\\
&\qquad+\sup_{h\in BL_1(\ell^\infty(\mathbb{T}))}\left|\mathbb{E}^\sharp h(v^{(r)\sharp}_{q_{a_n},P_{a_n}})-E_{P_{a_n}}h(v_{P_0})\right|\\
&=\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_{a_n}}(\theta,\tau,g)}{\hat{\varsigma}_{a_n}(\theta,\tau,g)}-1\right|\cdot\mathbb{E}^\sharp\left[\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|v^{(r)\sharp}_{q_{a_n},P_{a_n}}(\theta,\tau,g)\right|\right]\\
&\qquad+\sup_{h\in BL_1(\ell^\infty(\mathbb{T}))}\left|\mathbb{E}^\sharp h(v^{(r)\sharp}_{q_{a_n},P_{a_n}})-E_{P_{a_n}}h(v_{P_0})\right|\\
&\leq C^\sharp\cdot\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\frac{\varsigma_{P_{a_n}}(\theta,\tau,g)}{\hat{\varsigma}_{a_n}(\theta,\tau,g)}-1\right|\\
&\qquad+\sup_{h\in BL_1(\ell^\infty(\mathbb{T}))}\left|\mathbb{E}^\sharp h(v^{(r)\sharp}_{q_{a_n},P_{a_n}})-E_{P_{a_n}}h(v_{P_0})\right|, \qquad \text{(B.12)}
\end{aligned}$$

where $C^\sharp$ is the constant from Lemma S.2.3. Combine (B.10), (B.11) and (B.12) and conclude that

for $0 < \varepsilon' < \varepsilon$:

$$\limsup_{n\to\infty} \Pr_{P_{a_n}}\left(d_{BL_1}\left(v_{q_{a_n}}^{(r)\sharp}, v_{P_0}\right)^* > \varepsilon\right) = \limsup_{n\to\infty} \Pr_{P_{a_n}}\left(d_{BL_1}\left(\left(\frac{\varsigma_{P_{a_n}}}{\hat{\varsigma}_{a_n}}\right) v_{q_{a_n},P_{a_n}}^{(r)\sharp}, v_{P_0}\right)^* > \varepsilon\right) = 0.$$

Now by Lemma 1.9.2 in van der Vaart and Wellner (1996) there exists a further subsequence $\{b_n\}_{n=1}^{\infty} \subset \{a_n\}_{n=1}^{\infty}$ along which:

$$d_{BL_1}\left(v_{q_{b_n}}^{(r)\sharp}, v_{P_0}\right)^* \to 0 \text{ almost surely on } \Omega,$$

as $n \to \infty$. We will now prove that for any $\xi > 0$:

$$\lim_{n\to\infty}\left[(\mathbb{P}\times\mathbb{P}^\sharp)(T_{b_n}^{(r)\sharp}(\hat{\theta}_{b_n}) \le x_{b_n}) - \mathbb{P}(T_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}) \le x_{b_n} + \xi)\right] \le 0, \tag{B.13}$$

where, importantly, $x_{b_n} = x_{b_n}(\omega)$ may depend on ω. To this end, let Ω_0 denote the set of all $\omega \in \Omega$ such that:

$$d_{BL_1}\left(v_{q_{b_n}}^{(r)\sharp}, v_{P_0}\right)^* \to 0,$$

as $n \to \infty$. The discussion above implies $\mathbb{P}(\Omega_0) = 1$. For a fixed $\omega \in \Omega_0$, we can apply the almost-sure representation theorem (e.g. Dudley (2014) Theorem 3.24): there exists a probability space $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{\mathbb{P}})$ and perfect measurable functions s_{b_n} from $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{\mathbb{P}})$ to $(\Omega^\sharp, \mathfrak{F}^\sharp, \mathbb{P}^\sharp)$ for each $n = 0, 1, \ldots$ such that $\tilde{\mathbb{P}} \circ s_{b_n}^{-1} = \mathbb{P}^\sharp$ on $\mathfrak{F}^\sharp$ for each n and:

$$\left(\sup_{\theta\in\Theta}\sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\left|\tilde{v}_{b_n}^{(r)}(\tilde{\omega},\omega)(\theta,\tau,g) - \tilde{v}(\tilde{\omega},\omega)(\theta,\tau,g)\right|\right|\right)^* \to 0 \text{ almost surely on } \tilde{\Omega} \text{ as } n\to\infty, \tag{B.14}$$

where $\tilde{v}_{b_n}^{(r)}(\tilde{\omega},\omega)(\,\cdot\,) := v_{q_{b_n}}^{(r)\sharp}(s_{b_n}(\tilde{\omega}),\omega)(\,\cdot\,)$ and $\tilde{v}(\tilde{\omega},\omega) := v_{P_0}(s_{b_0}(\tilde{\omega}),\omega)(\,\cdot\,)$ for each $\tilde{\omega} \in \tilde{\Omega}$. Now define:

$$\tilde{T}_{0,b_n} = \sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}} \tilde{v}(\hat{\theta}_{b_n},\tau,g), \qquad \tilde{T}_{b_n}^{(r)\sharp} = \sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}} \tilde{v}_{b_n}^{(r)}(\hat{\theta}_{b_n},\tau,g).$$

By construction $\tilde{T}_{b_n}^{(r)\sharp}$ and $T_{b_n}^{(r)\sharp}$ are identically distributed, and $\tilde{T}_{0,b_n}$ and $T_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0})$ are identically distributed. Now note (B.13) follows by dominated convergence if we can show:

$$A := \limsup_{n\to\infty}\left[\tilde{\mathbb{P}}(\tilde{T}_{b_n}^{(r)\sharp} > x_{b_n} \mid \omega) - \tilde{\mathbb{P}}(\tilde{T}_{0,b_n} + \delta > x_{b_n} \mid \omega)\right] \le 0, \tag{B.15}$$

where "$\mid \omega$" is meant to emphasize that the $\omega-$coordinate is held fixed. We have:

$$\begin{aligned}\tilde{T}_{b_n}^{(r)\sharp} - \tilde{T}_{0,b_n} &= \sup_{\tau\in\mathcal{T}}\max_{g\in\mathcal{G}_{b_n}} \tilde{v}_{b_n}^{(r)}(\hat{\theta}_{b_n},\tau,g) - \sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}_{b_n}} \tilde{v}(\hat{\theta}_{b_n},\tau,g) \\ &\le \sup_{\tau\in\mathcal{T}}\sup_{g\in\mathcal{G}}\left|\tilde{v}_{b_n}^{(r)}(\hat{\theta}_{b_n},\tau,g) - \tilde{v}(\hat{\theta}_{b_n},\tau,g)\right|\end{aligned}$$

$$\leq \left(\sup_{\theta \in \Theta} \sup_{\tau \in \mathcal{T}} \sup_{g \in \mathcal{G}} \left| \tilde{v}_{b_n}^{(r)}(\theta, \tau, g) - \tilde{v}(\theta, \tau, g) \right| \right)^*$$
$$\longrightarrow 0,$$

almost surely as $n \to \infty$ by (B.14). Conclude that there exists a measurable sequence $B_{b_n}(\tilde{\omega})$ such that $\tilde{T}_{b_n}^{(r)\sharp}(\tilde{\omega}) - \tilde{T}_{0,b_n}(\tilde{\omega}) \leq B_{b_n}(\tilde{\omega}) \to 0$ almost surely as $n \to \infty$. Now define:

$$\begin{aligned}\tilde{\Delta}_{b_n}(\tilde{\omega}) &= \mathbb{1}\{\tilde{T}_{b_n}^{(r)\sharp}(\tilde{\omega}) > x_{b_n}\} - \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + \delta > x_{b_n}\} \\ &= \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + \tilde{T}_{b_n}^{(r)\sharp}(\tilde{\omega}) - \tilde{T}_{0,b_n}(\tilde{\omega}) > x_{b_n}\} - \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + \delta > x_{b_n}\}.\end{aligned}$$

Note that:

$$\tilde{\Delta}_{b_n}(\tilde{\omega}) \leq (\tilde{\Delta}_{b_n}(\tilde{\omega}))^* \leq \tilde{\Delta}_{b_n}^{\dagger}(\tilde{\omega}) := \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + B_{b_n}(\tilde{\omega}) > x_{b_n}\} - \mathbb{1}\{\tilde{T}_{0,b_n}(\tilde{\omega}) + \delta > x_{b_n}\}.$$

Furthermore, let $\tilde{\Delta}_{b_n}^{+}(\tilde{\omega}) := \max\{\tilde{\Delta}_{b_n}^{\dagger}(\tilde{\omega}), 0\}$, and note that $\tilde{\Delta}_{b_n}^{\dagger}(\tilde{\omega}) \leq \tilde{\Delta}_{b_n}^{+}(\tilde{\omega})$. Since $B_{b_n}(\tilde{\omega}) \to 0$ almost surely, $\tilde{\Delta}_{b_n}^{+}(\tilde{\omega}) \to 0$ almost surely. Let $\tilde{\mathbb{E}}[\,\cdot \mid \omega]$ denote the expectation with respect to the marginal $\tilde{\mathbb{P}}$, holding fixed ω. Then by dominated convergence we have:

$$A = \limsup_{n \to \infty} \tilde{\mathbb{E}}^*[\tilde{\Delta}_{b_n} \mid \omega] \leq \limsup_{n \to \infty} \tilde{\mathbb{E}}[\tilde{\Delta}_{b_n}^{\dagger} \mid \omega] \leq \limsup_{n \to \infty} \tilde{\mathbb{E}}[\tilde{\Delta}_{b_n}^{+} \mid \omega] = 0.$$

This completes the proof of (B.15). The result (B.13) then follows by dominated convergence. We now proceed in a similar manner to the proof of Lemma D.4 in Andrews and Shi (2017). In particular, since $\delta \in (0, \eta)$, the interval $(0, \eta - \delta)$ is non-empty, so using (B.13) we have for all $\xi \in (0, \eta - \delta)$:

$$\begin{aligned}&\limsup_{n \to \infty} (\mathbb{P} \times \mathbb{P}^\sharp)(T_{b_n}^{(r)\sharp}(\hat{\theta}_{b_n}) \leq c_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta - \eta) \\ &= \limsup_{n \to \infty} \mathbb{P}(T_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}) \leq c_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta - \eta + \xi) \\ &\quad + \limsup_{n \to \infty} \Big[(\mathbb{P} \times \mathbb{P}^\sharp)(T_{b_n}^{(r)\sharp}(\hat{\theta}_{b_n}) \leq c_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta - \eta) \\ &\quad - \mathbb{P}(T_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}) \leq c_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta - \eta + \xi) \Big] \\ &\leq 1 - \alpha,\end{aligned}$$

where the last inequality follows from (B.13) taking $x_{b_n} = c_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta - \eta$, and from the fact that $c_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha)$ is the $1-\alpha$ quantile of the statistic $T_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0})$. But by definition of $\hat{c}_{b_n}^{(r)\sharp}(1-\alpha+\eta)$, for all n we have:

$$\hat{c}_{b_n}^{(r)\sharp}(1-\alpha+\eta) := \inf\left\{ c : (\mathbb{P} \times \mathbb{P}^\sharp)(T_{b_n}^{(r)\sharp}(\hat{\theta}_{b_n}) \leq c) \geq 1-\alpha+\eta \right\}.$$

Thus the result above implies that:

$$c_0^{(r)}(\hat{\theta}_{b_n}, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \delta - \eta < \hat{c}_{b_n}^{(r)\sharp}(1-\alpha+\eta),$$

asymptotically with $\mathbb{P} \times \mathbb{P}^\sharp-$probability 1. This completes the proof. ■

The following lemma plays the same role as Lemma A3 in Andrews and Shi (2013).

Lemma B.6. *Suppose Assumptions 3.1, 3.2 and 3.3 hold. Let the event $A_{n,P}$ be as defined in (A.6), and let $\{P_n \in \mathcal{P} : n \geq 1\}$ be any sequence. Then:*

$$\limsup_{n\to\infty} Pr_{P_n}\left(\left\{c_0^{(r)}(\hat{\theta}_n, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \varepsilon < c_0^{(r)}(\hat{\theta}_n, h_{1,q_n,P_n}, h_{2,P_0}, h_{3,P_0}, 1-\alpha)\right\} \cap A_{n,P_n}\right) = 0,$$

where $c_0^{(r)}(\hat{\theta}_n, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha)$ is the $1-\alpha$ quantile of the distribution of the test statistic $T_0^{(r)}(\hat{\theta}_n, 0, h_{2,P_0}, h_{3,P_0})$ and $c_0^{(r)}(\hat{\theta}_n, h_{1,q_n,P_n}, h_{2,P_0}, h_{3,P_0}, 1-\alpha)$ is the $1-\alpha$ quantile of the distribution of $T_0^{(r)}(\hat{\theta}_n, h_{1,q_n,P_n}, h_{2,P_0}, h_{3,P_0})$.

Proof of Lemma B.6. Note that on the event A_{n,P_n} we have $h_{1,q_n,P_n}(\hat{\theta}_n, \tau, g) \leq \varepsilon$ for all $(\tau, g) \in \mathcal{T} \times \mathcal{G}_n$, which implies:

$$\begin{aligned}
T_0^{(r)}(\hat{\theta}_n, 0, h_{2,P_0}, h_{3,P_0}) &= \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} v_{P_0}(\hat{\theta}_n, \tau, g) \\
&= \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left(v_{P_0}(\hat{\theta}_n, \tau, g) - h_{1,q_n,P_n}(\hat{\theta}_n, \tau, g) + h_{1,q_n,P_n}(\hat{\theta}_n, \tau, g)\right) \\
&\geq \sup_{\tau\in\mathcal{T}} \max_{g\in\mathcal{G}_n} \left(v_{P_0}(\hat{\theta}_n, \tau, g) + h_{1,q_n,P_n}(\hat{\theta}_n, \tau, g)\right) - \varepsilon \\
&= T_0^{(r)}(\hat{\theta}_n, h_{1,q_n,P_n}, h_{2,P_0}, h_{3,P_0}) - \varepsilon.
\end{aligned}$$

This in turn implies that $c_0^{(r)}(\hat{\theta}_n, 0, h_{2,P_0}, h_{3,P_0}, 1-\alpha) + \varepsilon \geq c_0^{(r)}(\hat{\theta}_n, h_{1,q_n,P_n}, h_{2,P_0}, h_{3,P_0}, 1-\alpha)$. ■