# Asymptotically Exact Inference in Conditional Moment Inequality Models

Timothy B. Armstrong*
Stanford University

November 10, 2011

JOB MARKET PAPER

**Abstract**

This paper derives the rate of convergence and asymptotic distribution for a class of Kolmogorov-Smirnov style test statistics for conditional moment inequality models for parameters on the boundary of the identified set under general conditions. In contrast to other moment inequality settings, the rate of convergence is faster than root-$n$, and the asymptotic distribution depends entirely on nonbinding moments. The results require the development of new techniques that draw a connection between moment selection, irregular identification, bandwidth selection and nonstandard M-estimation. Using these results, I propose tests that are more powerful than existing approaches for choosing critical values for this test statistic. I quantify the power improvement by showing that the new tests can detect alternatives that converge to points on the identified set at a faster rate than those detected by existing approaches. A monte carlo study confirms that the tests and the asymptotic approximations they use perform well in finite samples. In an application to a regression of prescription drug expenditures on income with interval data from the Health and Retirement Study, confidence regions based on the new tests are substantially tighter than those based on existing methods.

*email: timothya@stanford.edu. Thanks to Han Hong and Joe Romano for guidance and many useful discussions, and to Liran Einav, Azeem Shaikh, Tim Bresnahan, Guido Imbens, Raj Chetty, Whitney Newey, Victor Chernozhukov, Jerry Hausman, Andres Santos, Elie Tamer, Vicky Zinde-Walsh, Alberto Abadie, Karim Chalak, Xu Cheng, Stefan Hoderlein, Don Andrews, Peter Phillips, Taisuke Otsu, Ed Vytlacil, Xiaohong Chen, Yuichi Kitamura and participants at seminars at Stanford and MIT for helpful comments and criticism. All remaining errors are my own. This paper was written with generous support from a fellowship from the endowment in memory of B.F. Haley and E.S. Shaw through the Stanford Institute for Economic Policy Research.

# 1 Introduction

Theoretical restrictions used for estimation of economic models often take the form of moment inequalities. Examples include models of consumer demand and strategic interactions between firms, bounds on treatment effects using instrumental variables restrictions, and various forms of censored and missing data (see, among many others, Manski, 1990; Manski and Tamer, 2002; Pakes, Porter, Ho, and Ishii, 2006; Ciliberto and Tamer, 2009; Chetty, 2010, and papers cited therein). For these models, the restriction often takes the form of moment inequalities conditional on some observed variable. That is, given a sample $(X_1, W_1), \ldots (X_n, W_n)$, we are interested in testing a null hypothesis of the form $E(m(W_i, \theta)|X_i) \geq 0$ with probability one, where the inequality is taken elementwise if $m(W_i, \theta)$ is a vector. Here, $m(W_i, \theta)$ is a known function of an observed random variable $W_i$, which may include $X_i$, and a parameter $\theta \in \mathbb{R}^{d_\theta}$, and the moment inequality defines the identified set $\Theta_0 \equiv \{\theta | E(m(W_i, \theta)|X_i) \geq 0 \text{ a.s.}\}$ of parameter values that cannot be ruled out by the data and the restrictions of the model.

In this paper, I consider inference in models defined by conditional moment inequalities. I focus on test statistics that exploit the equivalence between the null hypothesis $E(m(W_i, \theta)|X_i) \geq 0$ almost surely and $Em(W_i, \theta)I(s < X_i < s + t) \geq 0$ for all $(s, t)$. Thus, we can use $\inf_{s,t} \frac{1}{n} \sum_{i=1}^{n} m(W_i, \theta)I(s < X_i < s + t)$, or the infimum of some weighted version of the unconditional moments indexed by $(s, t)$. Following the terminology commonly used in the literature, I refer to these as Kolmogorov-Smirnov (KS) style test statistics. The main contribution of this paper is to derive the rate of convergence and asymptotic distribution of this test statistic for parameters on the boundary of the identified set under a general set of conditions. The asymptotic distributions derived in this paper and the methods used to derive them fall into a different category than other asymptotic distributions derived in the conditional moment inequalities and goodness-of-fit testing literatures. Rather, the asymptotic distributions and rates of convergence derived here resemble more closely those of maximized objective functions for nonstandard M-estimators (see, for example, Kim and Pollard, 1990), but require new methods to derive. The results draw a connection between moment selection, bandwidth selection, irregular identification and nonstandard M-estimation.

While asymptotic distribution results are available for this statistic in some cases (Andrews and Shi, 2009; Kim, 2008), the existing results give only a conservative upper bound of $\sqrt{n}$ on the rate of convergence of this test statistic in a large class of important cases. For example, in the interval regression model, the asymptotic distribution of this test statistic for parameters on the boundary of the identified set and the proper scaling needed to achieve it have so far been unknown in the generic case (see Section 2 for the definition of this model). In these cases, results available in the literature do not give an asymptotic distribution result, but state only that the test statistic converges in probability to zero when scaled up by $\sqrt{n}$. This paper derives the scaling that leads to a nondegenerate asymptotic distribution and characterizes this distribution. Existing results can be used for conservative inference in these cases (along with tuning parameters to prevent the critical value from going to zero), but lose power relative to procedures that use the results derived in this paper to choose critical values based on the asymptotic distribution of the test statistic on the boundary of the identified set.

To quantify this power improvement, I show that using the asymptotic distributions derived in this paper gives power against sequences of parameter values that approach points on the boundary of the identified set at a faster rate than those detected using root-$n$ convergence to a degenerate distribution. Since local power results have not been available for the conservative approach based on root-$n$ approximations in this setting, making this comparison involves deriving new local power results for the existing tests in addition to the new tests. The increase in power is substantial. In the leading case considered in Section 3, I find that the methods developed in this paper give power against local alternatives that approach the identified set at a $n^{-2/(d_X+4)}$ rate (where $d_X$ is the dimension of the conditioning variable), while using conservative $\sqrt{n}$ approximations only gives power against $n^{-1/(d_X+2)}$ alternatives. The power improvements are not completely free, however, as the new tests require smoothness conditions not needed for existing approaches. In another paper (Armstrong, 2011), I propose a modification of this test statistic that achieves a similar power improvement (up to a $\log n$ term) without sacrificing the robustness of the conservative approach. See Section 10 for more on these tradeoffs.

To examine how well these asymptotic approximations describe sample sizes of practical importance, I perform a monte carlo study. Confidence regions based on the tests proposed in this paper have close to the nominal coverage in the monte carlos, and shrink to the identified set at a faster rate than those based on existing tests. In addition, I provide an empirical illustration examining the relationship between out of pocket prescription spending and income in a data set in which out of pocket prescription spending is sometimes missing or reported as an interval. Confidence regions for this application constructed using the methods in this paper are substantially tighter than those that use existing methods (these confidence regions are reported in Figures 8 and 9 and Table 5; see Section 9 for the details of the empirical illustration).

While the asymptotic distribution results in this paper are technical in nature, the key insights can be described at an intuitive level. I provide a nontechnical exposition of these ideas in Section 2. Together with the statements of the asymptotic distribution results in Section 3 and the local power results in Section 7, this provides a general picture of the results of the paper. The rest of this section discusses the relation of these results to the rest of the literature, and introduces notation and definitions. Section 5 generalizes the asymptotic distribution results of Section 3, and Sections 4 and 6 deal with estimation of the asymptotic distribution for feasible inference. Section 8 presents monte carlo results. Section 9 presents the empirical illustration. In Section 10, I discuss some implications of these results beyond the immediate application to constructing asymptotically exact tests. Section 11 concludes. Proofs are in the appendix.

## 1.1 Related Literature

The results in this paper relate to recent work on testing conditional moment inequalities, including papers by Andrews and Shi (2009), Kim (2008), Khan and Tamer (2009), Chernozhukov, Lee, and Rosen (2009), Lee, Song, and Whang (2011), Ponomareva (2010), Menzel (2008) and Armstrong (2011). The results on the local power of asymptotically exact and conservative KS statistic based procedures derived in this paper are useful for comparing confidence regions based on KS statistics to other methods of inference on the identified set proposed in these papers. Armstrong (2011) derives local power results for some common alternatives to the KS statistics based on integrated moments considered in this paper (the confidence regions considered in that paper satisfy the stronger criterion of containing the entire identified set, rather than individual points, with a prespecified probability). I compare the local power calculations in this paper with those results in Section 10.

Out of these existing approaches to inference on conditional moment inequalities, the papers that are most closely related to this one are those by Andrews and Shi (2009) and Kim (2008), both of which consider statistics based on integrating the conditional inequality. As discussed above, the main contributions of the present paper relative to these papers are (1) deriving the rate of convergence and nondegenerate asymptotic distribution of this statistic for parameters on the boundary of the identified set in the common case where the results in these papers reduce to a statement that the statistic converges to zero at a root-$n$ scaling and (2) deriving local power results that show how much power is gained by using critical values based on these new results. Armstrong (2011) uses a statistic similar to the one considered here, but proposes an increasing sequence of weightings ruled out by the assumptions of the rest of the literature (including the present paper). This leads to almost the same power improvement as the methods in this paper even when conservative critical values are used. Khan and Tamer (2009) propose a statistic similar to one considered here for a model defined by conditional moment inequalities, but consider point estimates and confidence intervals based on these estimates under conditions that lead to point identification. Galichon and Henry (2009) propose a similar statistic for a class of partially identified models under a different setup. Statistics based on integrating conditional moments have been used widely in other contexts as well, and go back at least to Bierens (1982).

The literature on models defined by finitely many unconditional moment inequalities is more developed, but still recent. Papers in this literature include Andrews, Berry, and Jia (2004), Andrews and Jia (2008), Andrews and Guggenberger (2009), Andrews and Soares (2010), Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2010), Romano and Shaikh (2008), Bugni (2010), Beresteanu and Molinari (2008), Moon and Schorfheide (2009), Imbens and Manski (2004) and Stoye (2009). While most of this literature does not apply directly to the problems considered in this paper when the conditioning variable is continuous, ideas from these papers have been used in the literature on conditional moment inequality models and other problems involving inference on sets. Indeed, some of these results are stated in a broad enough way to apply to the general problem of inference on partially identified models.

## 1.2 Notation and Definitions

Throughout this paper, I use the terms asymptotically exact and asymptotically conservative to refer to the behavior of tests for a fixed parameter value under a fixed probability distribution. I refer to a test as asymptotically exact for testing a parameter $\theta$ under a data generating process $P$ such that the null hypothesis holds if the probability of rejecting $\theta$ converges to the nominal level as the number of observations increases to infinity under $P$. I refer to a test as asymptotically conservative for testing a parameter $\theta$ under a data generating process $P$ if the probability of falsely rejecting $\theta$ is asymptotically strictly less than the nominal level under $P$. While this contrasts with a definition where a test is conservative only if the size of the test is less than the nominal size taken as the supremum of the probability of rejection over a composite null of all possible values of $\theta$ and $P$ such that $\theta$ is in the identified set under $P$, it facilitates discussion of results like the ones in this paper (and other papers that deal with issues related to moment selection) that characterize the behavior of tests for different values of $\theta$ in the identified set.

I use the following notation in the rest of the paper. For observations $(X_1, W_1), \ldots, (X_n, W_n)$ and a measurable function $h$ on the sample space, $E_n h(X_i, W_i) \equiv \frac{1}{n}\sum_{i=1}^{n} h(X_i, W_i)$ denotes the sample mean. I use double subscripts to denote elements of vector observations so that $X_{i,j}$ denotes the $j$th component of the $i$th observation $X_i$. Inequalities on Euclidean space refer to the partial ordering of elementwise inequality. For a vector valued function $h : \mathbb{R}^{\ell} \to \mathbb{R}^{m}$, the infimum of $h$ over a set $T$ is defined to be the vector consisting of the infimum of each element: $\inf_{t\in T} h(t) \equiv (\inf_{t\in T} h_1(t), \ldots, \inf_{t\in T} h_m(t))$. I use $a \wedge b$ to denote the elementwise minimum and $a \vee b$ to denote the elementwise maximum of $a$ and $b$. The notation $\lceil x \rceil$ denotes the least integer greater than or equal to $x$.

# 2 Overview of Results

The asymptotic distributions derived in this paper arise when the conditional moment inequality binds only on a probability zero set. In contrast to inference with finitely many unconditional moment inequalities, in which at least one moment inequality will bind on the boundary of the identified set and limiting distributions of test statistics are degenerate only on the interior of the identified set, this lack of nondegenerate binding moments holds even on the boundary of the identified set in typical applications. This leads to a faster than root-$n$ rate of convergence to an asymptotic distribution that depends entirely on moments that are close to, but not quite binding.

To see why this case is typical in applications, consider an application of moment inequalities to regression with interval data. In the interval regression model, $E(W_i^*|X_i) = X_i'\beta$, and $W_i^*$ is unobserved, but known to be between observed variables $W_i^H$ and $W_i^L$, so that $\beta$ satisfies the moment inequalities

$$E(W_i^L|X_i) \leq X_i'\beta \leq E(W_i^H|X_i).$$

Suppose that the distribution of $X_i$ is absolutely continuous with respect to the Lebesgue measure. Then, to have one of these inequalities bind on a positive probability set, $E(W_i^L|X_i)$ or $E(W_i^H|X_i)$ will have to be linear on this set. Even if this is the case, this only means that the moment inequality will bind on this set for one value of $\beta$, and the moment inequality will typically not bind when applied to nearby values of $\beta$ on the boundary of the identified set. Figures 1 and 2 illustrate this for the case where the conditioning variable is one dimensional. Here, the horizontal axis is the nonconstant part of $x$, and the vertical axis plots the conditional mean of the $W_i^H$ along with regression functions corresponding to points in the identified set. Figure 1 shows a case where the KS statistic converges at a faster than root-$n$ rate. In Figure 2, the parameter $\beta_1$ leads to convergence at exactly a root-$n$ rate, but this is a knife edge case, since the KS statistic for testing $\beta_2$ will converge at a faster rate.

This paper derives asymptotic distributions under conditions that generalize these cases to arbitrary moment functions $m(W_i, \theta)$. In this broader setting, KS statistics converge at a faster than root-$n$ rate on the boundary of the identified set under general conditions when the model is set identified and at least one conditioning variable is continuously distributed. In interval quantile regression, contact sets for the conditional median translate to contact sets for the conditional mean of the moment function, leading to faster than root-$n$ rates of convergence in similar settings. Bounds in selection models, such as those proposed by Manski (1990), lead to a similar setup to the interval regression model, as do some of the structural models considered by Pakes, Porter, Ho, and Ishii (2006), with the intervals depending on a first stage parameter estimate. See Armstrong (2011) for primitive conditions for a set of high-level conditions similar to the ones used in this paper for some of these models.

While the results hold more generally, the rest of this section describes the results in the context of the interval regression example in a particular case. Consider deriving the rate of convergence and nondegenerate asymptotic distribution of the KS statistic for a parameter $\beta$ like the one shown in Figure 1, but with $X_i$ possibly containing more than one covariate. Since the lower bound never binds, it is intuitively clear that the KS statistic for the lower bound will converge to zero at a faster rate than the KS statistic for the upper bound, so consider the KS statistic for the upper bound given by $\inf_{s,t} E_n Y_i I(s < X_i < s + t)$ where $Y_i = W_i^H - X_i'\beta$. If $E(W_i^H|X_i = x)$ is tangent to $x'\beta$ at a single point $x_0$, and $E(W_i^H|X_i = x)$ has a positive second derivative matrix $V$ at this point, we will have $E(Y_i|X_i = x) \approx (x - x_0)'V(x - x_0)$ near $x_0$, so that, for $s$ near $x_0$ and $t$ close to zero, $EY_i I(s < X_i < s + t) \approx f_X(x_0) \int_{s_1}^{s_1+t_1} \cdots \int_{s_{d_X}}^{s_{d_X}+t_{d_X}} (x - x_0)'V(x - x_0)\, dx_{d_X} \cdots dx_1$ (here, if the regression contains a constant, the conditioning variable $X_i$ is redefined to be the nonconstant part of the regressor, so that $d_X$ refers to the dimension of the nonconstant part of $X_i$).

Since $EY_i I(s < X_i < s + t) = 0$ only when $Y_i I(s < X_i < s + t)$ is degenerate, the asymptotic behavior of the KS statistic should depend on indices $(s, t)$ where the moment inequality is not quite binding, but close enough to binding that sampling error makes $E_n Y_i I(s < X_i < s + t)$ negative some of the time. To determine on which indices $(s, t)$ we should expect this to happen, split up the process in the KS statistic into a mean zero process and a drift term: $(E_n - E)Y_i I(s < X_i < s+t) + EY_i I(s < X_i < s+t)$. In order for this to be strictly negative some of the time, there must be non-negligible probability that the mean zero process is greater in absolute value than the drift term. That is, we must have $sd((E_n - E)Y_i I(s < X_i < s+t))$ of at least the same order of magnitude as $EY_i I(s < X_i < s+t)$. The idea is similar to rate of convergence arguments for M-estimators with possibly nonstandard rates of convergence, such as those considered by Kim and Pollard (1990). We have $sd((E_n - E)Y_i I(s < X_i < s+t)) = \mathcal{O}(\sqrt{\prod_i t_i}/\sqrt{n})$ for small $t$, and some calculations show that, for $s$ close to $x_0$, $EY_i I(s < X_i < s+t) \approx f_X(x_0) \int_{s_1}^{s_1+t_1} \cdots \int_{s_{d_X}}^{s_{d_X}+t_{d_X}} (x - x_0)'V(x - x_0)\, dx_{d_X} \cdots dx_1 \geq C\|(s - x_0, t)\|^2 \prod_i t_i$ for some $C > 0$. Thus, we expect the asymptotic distribution to depend on $(s,t)$ such that $\sqrt{\prod_i t_i}/\sqrt{n}$ is of the same or greater order of magnitude than $\|(s - x_0, t)\|^2 \prod_i t_i$, which corresponds to $\|(s - x_0, t)\|^2 \sqrt{\prod_i t_i}$ less than or equal to $\mathcal{O}(1/\sqrt{n})$.

To get the main intuition for the rate of convergence, let us first suppose that $s - x_0$ is of the same order of magnitude as $t$, and the components $t_i$ of $t$ are of the same order of magnitude, and show separately that cases where components of $(s,t)$ converge at different rates do not matter for the asymptotic distribution. If $s - x_0$ and all components $t_i$ are to converge to zero at the same rate $h_n$, we must have $\|(s - x_0, t)\| = \mathcal{O}(h_n)$ and $\prod_i t_i = \mathcal{O}(h_n^{d_X})$, so that, if $\|(s - x_0, t)\|^2 \sqrt{\prod_i t_i} \leq \mathcal{O}(1/\sqrt{n})$, we will have $\mathcal{O}(1/\sqrt{n}) \geq h_n^2 \sqrt{h_n^{d_X}} = h_n^{2+d_X/2}$ so that $h_n \leq \mathcal{O}(1/n^{1/(2(2+d_X/2))}) = \mathcal{O}(n^{-1/(4+d_X)})$. Then, for $(s,t)$ with $t$ in an $h_n$-neighborhood of zero, we will have $(E_n - E)Y_i I(s < X_i < s+t) = \mathcal{O}_P(\sqrt{\prod_i t_i}/\sqrt{n}) = \mathcal{O}_P(n^{-(d_X+2)/(d_X+4)})$.

Next suppose that $s$ or converges to $x_0$ more slowly than $h_n = n^{-1/(d_X+4)}$ or that one of the components of $t$ converges to zero more slowly than $h_n$. In this case, we will have $\|(s - x_0, t)\|$ greater than some sequence $k_n$ with $k_n/h_n \to \infty$, so that, to have $\|(s - x_0, t)\|^2 \sqrt{\prod_i t_i} \leq \mathcal{O}(1/\sqrt{n})$, we would have to have $\sqrt{\prod_i t_i} \leq \mathcal{O}(1/(k_n^2 \sqrt{n}))$ so that $(E_n - E)Y_i(s < X_i < s+t)$ will be of order less than $1/(k_n^2 n)$, which goes to zero at a faster rate than the $n^{-(d_X+2)/(d_X+4)}$ rate that we get when the components of $(s,t)$ converge at the same rate.

Thus, we should expect that the values of $(s,t)$ that matter for the asymptotic distribution of the KS statistic are those with $(s - x_0, t)$ of order $n^{-1/(d_X+4)}$, and that the KS statistic will converge in distribution when scaled up by $n^{-(d_X+2)/(d_X+4)}$ to the infimum of the limit of a sequence of local objective functions indexed by $(s,t)$ with $(s - x_0, t)$ in a sequence of $n^{-1/(d_X+4)}$ neighborhoods of zero. Formalizing this argument requires showing that this intuition holds uniformly in $(s,t)$. The formal proof uses a "peeling" argument along the lines of Kim and Pollard (1990), but a different type of argument is needed for regions where, even though $\|(s - x_0, t)\|$ is far from zero, some components of $t$ are small enough that $E_n Y_i I(s < X_i < s+t)$ may be slightly negative because the region $\{s < X_i < s+t\}$ is small and happens to catch a few observations with $Y_i < 0$. The proof formalizes the intuition that these regions cannot matter for the asymptotic distribution, since $\prod_i t_i$ must be much smaller than when $s$ is close to $x_0$ and the components of $t$ are of the same order of magnitude as each other.

These results can be used for inference once the asymptotic distribution is estimated. In Section 4, I describe two procedures for estimating this asymptotic distribution. The first is a generic subsampling procedure that uses only the fact that the statistic converges to a nondegenerate distribution at a known rate. The second is based on estimating a finite dimensional set of objects that allows this distribution to be simulated.

Both procedures rely on the conditional mean having a positive definite second derivative matrix near its minimum. To form tests that are asymptotically valid under more general conditions, I propose pre-tests for these conditions, and embed these tests in a procedure that uses the asymptotic approximation to the null distribution for which the pre-test finds evidence. I describe these pre-tests in Section 6, but, before doing this, I extend the results of Section 3 to a broader class of shapes of the conditional mean in Section 5. These results are useful for the pre-tests in Section 6.1, which adapt methods from Politis, Romano, and Wolf (1999) for estimating rates of convergence to this setting. Section 6.2 describes another pre-test for the conditions of Section 3, this one based on estimating the second derivative and testing for positive definiteness. The pre-tests are valid under regularity conditions governing the smoothness of the conditional mean.

One of the appealing features of using asymptotically exact critical values over conservative ones is the potential for more power against parameters outside of the identified set. In Section 7, I consider power against local alternatives. I describe the intuition for the results in more detail in that section, but the main idea is that, for a sequence of alternatives $\theta_n$ converging to a point $\theta$ on the identified set that under which the argument described above goes through, the drift process has an additional term $E(m(W_i,\theta_n) - m(W_i,\theta))I(s < X < s+t)$, where $s - x_0$ and $t$ are of order $h_n$. The exact asymptotics will detect $\theta_n$ when this term is of order $n^{-(d_X+2)/(d_X+4)}$, while conservative asymptotics will have power only when $\theta_n$ is large enough so that this term is of order $n^{-1/2}$. This leads to power against local alternatives of order $n^{-2/(d_X+4)}$ for the asymptotically exact critical values, and $n^{-1/(d_X+2)}$ when the conservative $\sqrt{n}$ approximation is used.

# 3 Asymptotic Distribution of the KS Statistic

Given iid observations $(X_1, W_1), \ldots, (X_n, W_n)$, of random variables $X_i \in \mathbb{R}^{d_X}$, $W_i \in \mathbb{R}^{d_W}$, we wish to test the null hypothesis that $E(m(W_i, \theta)|X_i) \geq 0$ almost surely, where $m : \mathbb{R}^{d_W} \times \Theta \to \mathbb{R}^{d_Y}$ is a known measurable function and $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ is a fixed parameter value. I use the notation $\bar{m}(\theta, x)$ to denote a version of $E(m(W_i, \theta)|X = x)$ (it will be clear from context which version is meant when this matters). In some cases when it is clear which parameter value is being tested, I will define $Y_i = m(W_i, \theta)$ for notational convenience. Defining $\Theta_0$ to be the identified set of values of $\theta$ in $\Theta$ that satisfy $E(m(W_i, \theta)|X_i) \geq 0$ almost surely, these tests can then be inverted to obtain a confidence region that, for every $\theta_0 \in \Theta_0$, contains $\theta_0$ with a prespecified probability (Imbens and Manski, 2004). The tests considered here will be based on asymptotic approximations, so that these statements will only hold asymptotically.

The results in this paper allow for asymptotically exact inference using KS style statistics in cases where the $\sqrt{n}$ approximations for these statistics are degenerate. This includes the case described in the introduction in which one component of $E(m(W_i, \theta)|X_i)$ is tangent to zero at a single point and the rest are bounded away from zero. While this case captures the essential intuition for the results in this paper, I state the results in a slightly more general way in order to make them more broadly applicable. I allow each component of $E(m(W_i, \theta)|X)$ to be tangent to zero at finitely many points, which may be different for each component. This is relevant in the interval regression example for parameters for which the regression line is tangent to $E(W_i^H|X)$ and $E(W_i^L|X)$ at different points. In the case of an interval regression on a scalar and a constant, the points in the identified set corresponding to the largest and smallest values of the slope parameter will typically have this property.

I consider KS style statistics that are a function of $\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t) = (\inf_{s,t} E_n m_1(W_i, \theta) I(s < X_i < s + t), \ldots, \inf_{s,t} E_n m_{d_Y}(W_i, \theta) I(s < X_i < s + t))$. Fixing some function $S : \mathbb{R}^{d_Y} \to \mathbb{R}_+$, we can then reject for large values of $S(\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s+t))$ (which correspond to more negative values of the components of $\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t)$ for typical choices of $S$). Note that this is different in general than taking $\sup_{s,t} S(E_n m(W_i, \theta) I(s < X_i < s + t))$, although similar ideas will apply here. Also, the moments $E_n m(W_i, \theta) I(s < X_i < s + t)$ are not weighted, but the results could be extended to allow for a weighting function $\omega(s, t)$, so that the infimum is over $\omega(s, t) E_n m(W_i, \theta) I(s < X_i < s + t)$ as long as $\omega(s, t)$ is smooth and bounded away from zero and infinity. The condition that the weight function be bounded uniformly in the sample size, which is also imposed by Andrews and Shi (2009) and Kim (2008), turns out to be important (see Armstrong, 2011).

I formalize the notion that $\theta$ is at a point in the identified set such that one or more of the components of $E(m(W_i,\theta)|X_i)$ is tangent to zero at a finite number of of points in the following assumption.

**Assumption 1.** *For some version of $E(m(W_i,\theta)|X_i)$, the conditional mean of each element of $m(W_i,\theta)$ takes its minimum only on a finite set $\{x|E(m_j(W_i,\theta)|X=x)=0 \text{ some } j\} = \mathcal{X}_0 = \{x_1,\ldots,x_\ell\}$. For each $k$ from $1$ to $\ell$, let $J(k)$ be the set of indices $j$ for which $E(m_j(W_i,\theta)|X=x_k)=0$. Assume that there exist neighborhoods $B(x_k)$ of each $x_k \in \mathcal{X}_0$ such that, for each $k$ from $1$ to $\ell$, the following assumptions hold.*

*i.) $E(m_j(W_i,\theta)|X_i)$ is bounded away from zero outside of $\cup_{k=1}^{\ell} B(x_k)$ for all $j$ and, for $j \notin J(k)$, $E(m_j(W_i,\theta)|X_i)$ is bounded away from zero on $B(x_k)$.*

*ii.) For $j \in J(k)$, $x \mapsto E(m_j(W_i,\theta)|X=x)$ has continuous second derivatives inside of the closure of $B(x_k)$ and a positive definite second derivative matrix $V_j(x_k)$ at each $x_k$.*

*iii.) $X$ has a continuous density $f_X$ on $B(x_k)$.*

*iv.) Defining $m_{J(k)}(W_i,\theta)$ to have $j$th component $m_j(W_i,\theta)$ if $j \in J(k)$ and $0$ otherwise, $x \mapsto E(m_{J(k)}(W_i,\theta)m_{J(k)}(W_i,\theta)'|X_i=x)$ is finite and continuous on $B(x_k)$ for some version of this conditional second moment matrix.*

Assumption 1 is the main substantive assumption distinguishing the case considered here from the case where the KS statistic converges at a $\sqrt{n}$ rate. In the $\sqrt{n}$ case, some component of $E(m(W_i,\theta)|X_i)$ is equal to zero on a positive probability set. Assumption 1 states that any component of $E(m(W_i,\theta)|X_i)$ is equal to zero only on a finite set, and that $X_i$ has a density in a neighborhood of this set, so that this finite set has probability zero. Note that the assumption that $X_i$ has a density at certain points means that the moment inequalities must be defined so that $X_i$ does not contain a constant. Thus, the results stated below hold in the interval regression example with $d_X$ equal to the number of nonconstant regressors.

Unless otherwise stated, I assume that the contact set $\mathcal{X}_0$ in Assumption 1 is nonempty. If Assumption 1 holds with $\mathcal{X}_0$ empty so that the conditional mean $\bar{m}(\theta,x)$ is bounded from below away from zero, $\theta$ will typically be on the interior of the identified set (as long as the conditional mean stays bounded away from zero when $\theta$ is moved a small amount). For such values of $\theta$, KS statistics will converge at a faster rate (see Lemma 6 in the appendix), leading to conservative inference even if the rates of convergence derived under Assumption 1, which are faster than $\sqrt{n}$, are used.

In addition to imposing that the minimum of the components of the conditional mean $\bar{m}(\theta, x)$ over $x$ are taken on a probability zero set, Assumption 1 requires that this set be finite, and that $\bar{m}(\theta, x)$ behave quadratically in $x$ near this set. I state results under this condition first, since it is easy to interpret as arising from a positive definite second derivative matrix at the minimum, and is likely to provide a good description of many situations encountered in practice. In Section 5, I generalize these results to other shapes of the conditional mean. This is useful for the tests for rates of convergence in Section 6, since the rates of convergence turn out to be well behaved enough to be estimated using adaptations of existing methods.

The next assumption is a regularity condition that bounds $m_j(W_i, \theta)$ by a nonrandom constant. This assumption will hold naturally in models based on quantile restrictions. In the interval regression example, it requires that the data have finite support. This assumption could be replaced with an assumption that $m(W_i, \theta)$ has exponentially decreasing tails, or even a finite $p$th moment for some potentially large $p$ that would depend on $d_X$ without much modification of the proof, but the finite support condition is simpler to state.

**Assumption 2.** *For some nonrandom $\overline{Y} < \infty$, $|m_j(W_i, \theta)| \leq \overline{Y}$ with probability one for each $j$.*

Finally, I make the following assumption on the function $S$. Part of this assumption could be replaced by weaker smoothness conditions, but the assumption covers $x \mapsto \|x\|_- \equiv \|x \wedge 0\|$ for any norm $\|\cdot\|$ as stated, which should suffice for practical purposes.

**Assumption 3.** *$S : \mathbb{R}^{d_Y} \to \mathbb{R}_+$ is continuous and satisfies $S(ax) = aS(x)$ for any nonnegative scalar $a$.*

The following theorem gives the asymptotic distribution and rate of convergence for $\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s+t)$ under these conditions. The distribution of $S(\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s+t))$ under mild conditions on $S$ then follows as an easy corollary.

**Theorem 1.** *Under Assumptions 1 and 2,*

$$n^{(d_X+2)/(d_X+4)} \inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s+t) \xrightarrow{d} Z$$

*where $Z$ is a random vector on $\mathbb{R}^{d_Y}$ defined as follows. Let $\mathbb{G}_{P,x_k}(s,t)$, $k = 1, \ldots, \ell$ be independent mean zero Gaussian processes with sample paths in the space $C(\mathbb{R}^{2d_X}, \mathbb{R}^{d_Y})$ of* *continuous functions from* $\mathbb{R}^{2d_X}$ *to* $\mathbb{R}^{d_Y}$ *and covariance kernel*

$$cov(\mathbb{G}_{P,x_k}(s,t), \mathbb{G}_{P,x_k}(s',t')) = E(m_{J(k)}(W_i,\theta)m_{J(k)}(W_i,\theta)'|X_i = x_k)f_X(x_k) \int_{s\vee s'<x<(s+t)\wedge(s'+t')} dx$$

*where* $m_{J(k)}(W_i,\theta)$ *is defined to have* $j$*th element equal to* $m_j(W_i,\theta)$ *for* $j \in J(k)$ *and equal to zero for* $j \notin J(k)$*. For* $k = 1,\ldots,\ell$*, let* $g_{P,x_k} : \mathbb{R}^{2d_X} \to \mathbb{R}^{d_Y}$ *be defined by*

$$g_{P,x_k,j}(s,t) = \frac{1}{2}f_X(x_k)\int_{s_1}^{s_1+t_1}\cdots\int_{s_{d_X}}^{s_{d_X}+t_{d_X}} x'V_j(x_k)x\,dx_{d_X}\cdots dx_1$$

*for* $j \in J(k)$ *and* $g_{x_k,j}(s,t) = 0$ *for* $j \notin J(k)$*. Define* $Z$ *to have* $j$*th element*

$$Z_j = \min_{k\ s.t.\ j\in J(k)} \inf_{(s,t)\in\mathbb{R}^{2d_X}} \mathbb{G}_{P,x_k,j}(s,t) + g_{P,x_k,j}(s,t).$$

The asymptotic distribution of $S(\inf_{s,t} E_n m(W_i,\theta)I(s < X_i < s+t))$ follows immediately from this theorem.

**Corollary 1.** *Under Assumptions 1, 2, and 3,*

$$n^{(d_X+2)/(d_X+4)}S(\inf_{s,t} E_n m(W_i,\theta)I(s < X_i < s+t)) \xrightarrow{d} S(Z)$$

*for a random variable* $Z$ *with the distribution given in Theorem 1.*

These results will be useful for constructing asymptotically exact level $\alpha$ tests if the asymptotic distribution does not have an atom at the $1-\alpha$ quantile, and if the quantiles of the asymptotic distribution can be estimated. In the next section, I show that the asymptotic distribution is atomless under mild conditions and propose two methods for estimating the asymptotic distribution. The first is a generic subsampling procedure. The second is a procedure based on estimating a finite dimensional set of objects that determine the asymptotic distribution. This provides feasible methods for constructing asymptotically exact confidence intervals under Assumption 1. However, while, in many cases, this assumption characterizes the distribution of $(X_i, m(W_i,\theta))$ for most or all values of $\theta$ on the boundary of the identified set, it is not an assumption that one would want to impose a priori. Thus, these tests should be embedded in a procedure that tests between this case and cases where $E(m(W_i,\theta)|X) = 0$ on a positive probability set, or where $E(m(W_i,\theta)|X)$ is still equal to 0 only at finitely many points, but behaves like $x^4$ or the absolute value function or something else near these points rather than a quadratic function. In Section 5, I generalize Theorem 1 to handle a wider set of shapes of the conditional mean, with different rates of convergence for different cases. In Section 6, I propose procedures for testing for Assumption 1 under mild smoothness conditions. Combining one of these preliminary tests with inference that is valid in the corresponding case gives a procedure that is asymptotically valid under more general conditions. These include tests based on estimating the rate of convergence directly, which use the results of Section 5.

# 4 Inference

To ensure that the asymptotic distribution is continuous, we need to impose additional assumptions to rule out cases where components of $m(W_j, \theta)$ are degenerate. The next assumption rules out these cases.

**Assumption 4.** *For each $k$ from 1 to $\ell$, letting $j_{k,1}, \ldots, j_{k,|J(k)|}$ be the elements in $J(k)$, the matrix with $q, r$th element given by $E(m_{j_{k,q}}(W_i, \theta) m_{j_{k,r}}(W_i, \theta)|X_i = x_k)$ is invertible.*

This assumption simply says that the binding components of $m(W_i, \theta)$ have a nonsingular conditional covariance matrix at the point where they bind. A sufficient condition for this is for the conditional covariance matrix of $m(W_i, \theta)$ given $X_i$ to be nonsingular at these points.

I also make the following assumption on the function $S$, which translates continuity of the distribution of $Z$ to continuity of the distribution of $S(Z)$.

**Assumption 5.** *For any Lebesgue measure zero set $A$, $S^{-1}(A)$ has Lebesgue measure zero.*

Under these conditions, the asymptotic distribution in Theorem 1 is continuous. In addition to showing that the rate derived in that theorem is the exact rate of convergence (since the distribution is not a point mass at zero or some other value), this shows that inference based on this asymptotic approximation will be asymptotically exact.

**Theorem 2.** *Under Assumptions 1, 2, and 4, the asymptotic distribution in Theorem 1 is continuous. If Assumptions 3 and 5 hold as well, the asymptotic distribution in Corollary 1 is continuous.*

Thus, an asymptotically exact test of $E(m(W_i, \theta)|X_i) \geq 0$ can be obtained by comparing the quantiles of $S(\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t))$ to the quantiles of any consistent estimate of the distribution of $S(Z)$. I propose two methods for estimating this distribution. The first is a generic subsampling procedure. The second method uses the fact that the distribution of $Z$ in Theorem 1 depends on the data generating process only through finite dimensional parameters to simulate an estimate of the asymptotic distribution.

Subsampling is a generic procedure for estimating the distribution of a statistic using versions of the statistic formed with a smaller sample size (Politis, Romano, and Wolf, 1999). Since many independent smaller samples are available, these can be used to estimate the distribution of the original statistic as long as the distribution of the scaled statistic is stable as a function of the sample size. To describe the subsampling procedure, let $T_n(\theta) = \inf_{s,t} E_n m(W_i,\theta)I(s < X_i < s+t)$. For any set of indices $\mathcal{S} \subseteq \{1,\ldots,n\}$, define $T_{\mathcal{S}}(\theta) = \inf_{s,t} \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} m(W_i,\theta)I(s < X_i < s+t)$. The subsampling estimate of $P(S(Z) \leq t)$ is, for some subsample size $b$,

$$\frac{1}{\binom{n}{b}} \sum_{|\mathcal{S}|=b} I\left(b^{(d_X+2)/(d_X+4)} S(T_{\mathcal{S}}(\theta)) \leq t\right).$$

One can also estimate the null distribution using the centered subsampling estimate

$$\frac{1}{\binom{n}{b}} \sum_{|\mathcal{S}|=b} I\left(b^{(d_X+2)/(d_X+4)} [S(T_{\mathcal{S}}(\theta)) - S(T_n(\theta))] \leq t\right).$$

For some nominal level $\alpha$, let $\hat{q}_{b,1-\alpha}$ be the $1-\alpha$ quantile of either of these subsampling distributions. We reject the null hypothesis that $\theta$ is in the identified set at level $\alpha$ if $n^{(d_X+2)/(d_X+4)}S(T_n(\theta)) > \hat{q}_{b,1-\alpha}$ and fail to reject otherwise. The following theorem states that this procedure is asymptotically exact. The result follows immediately from general results for subsampling in Politis, Romano, and Wolf (1999).

**Theorem 3.** *Under Assumptions 1, 2, 3, 4 and 5, the probability of rejecting using the subsampling procedure described above with nominal level $\alpha$ converges to $\alpha$ as long as $b \to \infty$ and $b/n \to 0$.*

While subsampling is valid under general conditions, subsampling estimates may be less precise than estimates based on knowledge of how the asymptotic distribution relates to the data generating process. One possibility is to note that the asymptotic distribution in Theorem 1 depends on the underlying distribution only through the set $\mathcal{X}_0$ and, for points $x_k$ in $\mathcal{X}_0$, the density $f_X(x_k)$, the conditional second moment matrix $E(m_{J(k)}(W_i,\theta)m_{J(k)}(W_i,\theta)'|X = x_k)$, and the second derivative matrix $V(x_k)$ of the conditional mean. Thus, with consistent estimates of these objects, we can estimate the distribution in Theorem 1 by replacing these objects with their consistent estimates and simulating from the corresponding distribution.

In order to accommodate different methods of estimating $f_X(x_k)$, $E(m_{J(k)}(W_i,\theta)m_{J(k)}(W_i,\theta)'|X = x_k)$, and $V(x_k)$, I state the consistency of these estimators as a high level condition, and show that the procedure works as long as these estimators are consistent. Since these objects only appear as $E(m_{J(k)}(W_i,\theta)m_{J(k)}(W_i,\theta)'|X = x_k)f_X(x_0)$ and $f_X(x_k)V(x_k)$ in the asymptotic distribution, we actually only need consistent estimates of these objects.

**Assumption 6.** *The estimates $\hat{M}_k(x_k)$, $\hat{f}_X(x_k)$, and $\hat{V}(x_k)$ satisfy $\hat{f}_X(x_k)\hat{V}(x_k) \xrightarrow{p} f_X(x_k)V(x_k)$ and $\hat{M}_k(x_k)\hat{f}_X(x_k) \xrightarrow{p} E(m_{J(k)}(W_i,\theta)m_{J(k)}(W_i,\theta)'|X = x_k)f_X(x_k)$.*

For $k$ from 1 to $\ell$, let $\hat{\mathbb{G}}_{P,x_k}(s,t)$ and $\hat{g}_{P,x_k}(s,t)$ be the random process and mean function defined in the same way as $\mathbb{G}_{P,x_k}(s,t)$ and $g_{P,x_k}(s,t)$, but with the estimated quantities replacing the true quantities. We estimate the distribution of $Z$ defined to have $j$th element

$$Z_j = \min_{m \text{ s.t. } j\in J(k)} \inf_{(s,t)\in\mathbb{R}^{2d}} \mathbb{G}_{P,x_k,j}(s,t) + g_{P,x_k,j}(s,t)$$

using the distribution of $\hat{Z}$ defined to have $j$th element

$$\hat{Z}_j = \min_{k \text{ s.t. } j\in J(k)} \inf_{\|(s,t)\|\leq B_n} \hat{\mathbb{G}}_{P,x_k,j}(s,t) + \hat{g}_{P,x_k,j}(s,t)$$

for some sequence $B_n$ going to infinity. The convergence of the distribution $\hat{Z}$ to the distribution of $Z$ is in the sense of conditional weak convergence in probability often used in proofs of the validity of the bootstrap (see, for example, Lehmann and Romano, 2005). From this, it follows that tests that replace the quantiles of $S(Z)$ with the quantiles of $S(\hat{Z})$ are asymptotically exact under the conditions that guarantee the continuity of the limiting distribution.

**Theorem 4.** *Under Assumption 6, $\rho(\hat{Z}, Z) \xrightarrow{p} 0$ where $\rho$ is any metric on probability distributions that metrizes weak convergence.*

**Corollary 2.** *Let $\hat{q}_{1-\alpha}$ be the $1-\alpha$ quantile of $S(\hat{Z})$. Then, under Assumptions 1, 2, 3, 4, 5, and 6, the test that rejects when $n^{(d_X+2)/(d_X+4)}S(T_n(\theta)) > \hat{q}_{1-\alpha}$ and fails to reject otherwise is an asymptotically exact level $\alpha$ test.*

If the set $\mathcal{X}_0$ is known, the quantities needed to compute $\hat{Z}$ can be estimated consistently using standard methods for nonparametric estimation of densities, conditional moments, and their derivatives. However, typically $\mathcal{X}_0$ is not known, and the researcher will not even want to impose that this set is finite. In Section 6, I propose methods for testing Assumption 1 and estimating the set $\mathcal{X}_0$ under weaker conditions on the smoothness of the conditional mean. These conditions allow for both the $n^{(d_X+2)/(d_X+4)}$ asymptotics that arise from Assumption 1 and the $\sqrt{n}$ asymptotics that arise from a positive probability contact set.

Before describing these results, I extend the results of Section 3 to other shapes of the conditional mean. These results are needed for the tests in Section 6.1, which rely on the rate of convergence being sufficiently well behaved if it is in a certain range.

# 5 Other Shapes of the Conditional Mean

Assumption 1 states that the components of the conditional mean $\bar{m}(\theta, x)$ are minimized on a finite set and have strictly positive second derivative matrices at the minimum. More generally, if the conditional mean is less smooth, or does not take an interior minimum, $\bar{m}(\theta, x)$ could be minimized on a finite set, but behave differently near the minimum. Another possibility is that the minimizing set could have zero probability, while containing infinitely many elements (for example, an infinite countable set, or a lower dimensional set when $d_X > 1$).

In this section, I derive the asymptotic distribution and rate of convergence of KS statistics under a broader class of shapes of the conditional mean $\bar{m}(\theta, x)$. I replace part (ii) of Assumption 1 with the following assumption.

**Assumption 7.** *For $j \in J(k)$, $\bar{m}_j(\theta, x) = E(m_j(W_i, \theta)|X = x)$ is continuous on $B(x_k)$ and satisfies*

$$\sup_{\|x - x_k\| \le \delta} \left\| \frac{\bar{m}_j(\theta, x) - \bar{m}_j(\theta, x_k)}{\|x - x_k\|^{\gamma(j,k)}} - \psi_{j,k}\left(\frac{x - x_k}{\|x - x_k\|}\right) \right\| \overset{\delta \to 0}{\to} 0$$

*for some $\gamma(j, k) > 0$ and some function $\psi_{j,k} : \{t \in \mathbb{R}^{d_X} | \|t\| = 1\} \to \mathbb{R}$ with $\overline{\psi} \ge \psi_{j,k}(t) \ge \underline{\psi}$ for some $\overline{\psi} < \infty$ and $\underline{\psi} > 0$. For future reference, define $\gamma = \max_{j,k} \gamma(j, k)$ and $\tilde{J}(k) = \{j \in J(k)|\gamma(j, k) = \gamma\}$.*

When Assumption 7 holds, the rate of convergence will be determined by $\gamma$, and the asymptotic distribution will depend on the local behavior of the objective function for $j$ and $k$ with $j \in \tilde{J}(k)$.

Under Assumption 1, Assumption 7 will hold with $\gamma = 2$ and $\psi_{j,k}(t) = \frac{1}{2}tV_j(x_k)t$ (this holds by a second order Taylor expansion, as described in the appendix). For $\gamma = 1$, Assumption 7 states that $\bar{m}_j(\theta, x)$ has a directional derivative for every direction, with the approximation error going to zero uniformly in the direction of the derivative. More generally, Assumption 7 states that $\bar{m}_j(\theta, x)$ increases like $\|x - x_k\|^\gamma$ near elements $x_k$ in the minimizing set $\mathcal{X}_0$. For $d_X = 1$, this follows from simple conditions on the higher derivatives of the conditional mean with respect to $x$. With enough derivatives, the first derivative that is nonzero uniformly on the support of $X_i$ determines $\gamma$. I state this formally in the next theorem. For higher dimensions, Assumption 7 requires additional conditions to rule out contact sets of dimension less than $d_X$, but greater than 1.

**Theorem 5.** *Suppose $\bar{m}(\theta, x)$ has $p$ bounded derivatives, $d_X = 1$ and supp$(X_i) = [\underline{x}, \overline{x}]$. Then, if $\min_j \inf_x \bar{m}_j(\theta, x) = 0$, either Assumption 7 holds, with the contact set $\mathcal{X}_0$ possibly containing the boundary points $\underline{x}$ and $\overline{x}$, for $\gamma = r$ for some integer $r < p$, or, for some $x_0$ on the support of $X_i$ and some finite $B$, $\bar{m}_j(\theta, x) \le B|x - x_0|^p$ for some $j$.*

Theorem 5 states that, with $d_X = 1$ and $p$ bounded derivatives, either Assumption 7 holds for $\gamma$ some integer less than $p$, or, for some $j$, $\bar{m}_j(\theta, x)$ is less than or equal to the function $B|x - x_0|^p$, which would make Assumption 7 hold for $\gamma = p$. In the latter case, the rate of convergence for the KS statistic must be at least as slow as the rate of convergence when Assumption 1 holds with $\gamma = p$. While an interior minimum with a strictly positive second derivative or a minimum at $\underline{x}$ or $\overline{x}$ with a nonzero first derivative seem most likely, Theorem 5 shows that Assumption 7 holds under broader conditions on the smoothness of the conditional mean. This, along with the rates of convergence in Theorem 6 below, will be useful for the methods described later in Section 6 for testing between rates of convergence. With enough smoothness assumptions on the conditional mean, the rate of convergence will either be $n^\beta$ for $\beta$ in some known range, or strictly slower than $n^\beta$ for some known $\beta$. With this prior knowledge of the possible types of asymptotic behavior of $T_n(\theta)$ in hand, one can use a modified version of the estimators of the rate of convergence proposed by Politis, Romano, and Wolf (1999) to estimate $\gamma$ in Assumption 7, and to test whether this assumption holds.

Under Assumption 1 with part (ii) replaced by Assumption 7, the following modified version of Theorem 1, with a different rate of convergence and limiting distribution, will hold.

**Theorem 6.** *Under Assumption 1, with part (ii) replaced by Assumption 7, and Assumption 2,*

$$n^{(d_X+\gamma)/(d_X+2\gamma)} \inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t) \xrightarrow{d} Z$$

*where $Z$ is the random vector on $\mathbb{R}^{d_Y}$ defined as in Theorem 1, but with $J(k)$ replaced by $\tilde{J}(k)$ and $g_{P,x_k,j}(s,t)$ defined as*

$$g_{P,x_k,j}(s,t) = f_X(x_k) \int_{s_1}^{s_1+t_1} \cdots \int_{s_{d_X}}^{s_{d_X}+t_{d_X}} \psi_{j,k}\left(\frac{x}{\|x\|}\right) \|x\|^{\gamma} \, dx_{d_X} \cdots dx_1$$

*for $j \in \tilde{J}(k)$. If Assumption 3 holds as well, then*

$$n^{(d_X+\gamma)/(d_X+2\gamma)} S(\inf_{s,t} E_n m(W_i,\theta) I(s < X_i < s+t)) \xrightarrow{d} S(Z).$$

*If Assumption 4 holds as well, $Z$ has a continuous distribution. If Assumptions 3, 4 and 5 hold, $S(Z)$ has a continuous distribution.*

Theorem 6 can be used once Assumption 7 is known to hold for some $\gamma$, as long as $\gamma$ can be estimated. I treat this topic in the next section. Theorem 5 gives primitive conditions for this to hold for the case where $d_X = 1$ that rely only on the smoothness of the conditional mean. The only additional condition needed to use this theorem is to verify that the set $\mathcal{X}_0$ does not contain the boundary points $\underline{x}$ and $\overline{x}$. In fact, the requirement in Theorems 1 and 6 that $\mathcal{X}_0$ not contain boundary points could be relaxed, as long as the boundary is sufficiently smooth. The results will be similar as long as the density of $X_i$ is bounded away from zero on its support, and cases where the density of $X_i$ converges to zero smoothly near its support could be handled using a transormation of the data (see Armstrong, 2011, for an example of this approach in a slightly different setting). Alternatively, a pre-test can be done to see if the conditional mean is bounded away from zero near the boundary of the support of $X_i$ so that these results can be used as stated.

# 6 Testing Rate of Convergence Conditions

The $n^{(d_X+2)/(d_X+4)}$ convergence derived in Section 3 holds when the minimum of $\bar{m}_j(\theta,x) = E(m_j(W_i,\theta)|X_i = x)$ is taken at a finite number of points, each with a strictly positive definite second derivative matrix. The results in Section 5 extend these results to other shapes of the conditional mean near the contact set, which result in different rates of convergence. In contrast, if the minimum is taken on a positive probability set, convergence will be at the slower $\sqrt{n}$ rate. Under additional conditions on the smoothness of $\bar{m}_j(\theta,x)$ as a function of $x$, it is possible to test for the conditions that lead to the faster convergence rates. In this section, I describe two methods for testing between these conditions. In Section 6.1, I describe tests that use a generic test for rates of convergence based on subsampling proposed by Politis, Romano, and Wolf (1999). These tests are valid as long as the KS statistic converges to a nondegenerate distribution at some polynomial rate, or converges more slowly than some imposed rate, and the results in Section 5 give primitive conditions for this. In Section 6.2, I propose tests of Assumption 1 based on estimating the second derivative matrix of the conditional mean.

## 6.1 Tests Based on Estimating the Rate of Converence Directly

The pre-tests proposed in this section mostly follow Chapter 8 of Politis, Romano, and Wolf (1999), using the results in Section 5 to give primitive conditions under which the rate of convergence will be well behaved so that these results can be applied, with some modifications to accomodate the possibility that the statistic may not converge at a polynomial rate if the rate is slow enough. Following the notation of Politis, Romano, and Wolf (1999), define

$$L_{n,b}(x|\tau) \equiv \frac{1}{\binom{n}{b}} \sum_{|\mathcal{S}|=b} I(\tau_b[S(T_{\mathcal{S}}(\theta)) - S(T_n(\theta))] \leq x)$$

for any sequence $\tau_n$, and define

$$L_{n,b}(x|1) \equiv \frac{1}{\binom{n}{b}} \sum_{|\mathcal{S}|=b} I(S(T_{\mathcal{S}}(\theta)) - S(T_n(\theta)) \leq x).$$

Let

$$L^{-1}_{n,b}(t|1) = \inf\{x|L_{n,b}(x|1) \geq t\}$$

be the $t$th quantile of $L_{n,b}(x|1)$, and define $L^{-1}_{n,b}(t|\tau)$ similarly. Note that $\tau_b L^{-1}_{n,b}(t|1) = L^{-1}_{n,b}(t|\tau)$. If $\tau_n$ is the true rate of convergence, $L^{-1}_{n,b_1}(t|\tau)$ and $L^{-1}_{n,b_2}(t|\tau)$ both approximate the $t$th quantile of the asymptotic distribution. Thus, if $\tau_n = n^{\beta}$ for some $\beta$, $b_1^{\beta} L^{-1}_{n,b_1}(t|1)$ and $b_1^{\beta} L^{-1}_{n,b_1}(t|1)$ should be approximately equal, so that an estimator for $\beta$ can be formed by choosing $\hat{\beta}$ to set these quantities equal. Some calculation gives

$$\hat{\beta} = (\log L^{-1}_{n,b_2}(t|1)) - \log L^{-1}_{n,b_1}(t|1))/(\log b_1 - \log b_2). \tag{1}$$

This is a special case of the class of estimators described in Politis, Romano, and Wolf (1999) which allow averaging of more than two block sizes and more than one quantile (these estimators could be used here as well).

Note that the estimate $L_{n,b}(x|\tau)$ centers the subsampling draws around the KS statistic $S(T_n(\theta))$ rather than its limiting value, 0. This is necessary for the rate of convergence estimate not to diverge under fixed alternatives. Once the rate of convergence is known or estimated, either $L_{n,b}(x|\tau)$ or an uncentered version, defined as

$$\tilde{L}_{n,b}(x|\tau) \equiv \frac{1}{\binom{n}{b}} \sum_{|\mathcal{S}|=b} I(\tau_b S(T_{\mathcal{S}}(\theta)) \leq x),$$

can be used to estimate the null distribution of the scaled statistic.

The results in Politis, Romano, and Wolf (1999) show that subsampling with the estimated rate of convergence $n^{\hat{\beta}}$ is valid as long as the true rate of convergence is $n^{\beta}$ for some $\beta > 0$. However, this will not always be the case for the estimators considered in this paper. For example, under the conditions of Theorem 5, the rate of convergence will either be $n^{(1+\gamma)/(1+2\gamma)}$ for some $\gamma < p$ (here, $d_X = 1$), or the rate of convergence will be at least as slow as $n^{(1+p)/(1+2p)}$, but may converge at a slower rate, or oscillate between slower rates of convergence. Even if Assumption 5 holds for some $\gamma$ for $\theta$ on the boundary of the identified set, the rate of convergence will be faster for $\theta$ on the interior of the identified set, where trying not to be conservative typically has little payoff in terms of power against parameters outside of the identified set.

To remedy these issues, I propose truncated versions defined as follows. For some $1/2 \leq \underline{\beta} < \overline{\beta} < 1$, let $\hat{\beta}$ be the estimate given by (1) for $b_1 = n^{\chi_1}$ and $b_2 = n^{\chi_2}$ for some $1 > \chi_1 > \chi_2 > 0$, and let $\hat{\beta}_a$ be the estimate given by (1) for $b_2 = n^{\chi_a}$ for some $1 > \chi_a > 0$ and $b_1$ some fixed constant that does not change with the sample size (if $L_{n,b_1}^{-1}(t|1)) = 0$, replace this with an arbitrary positive constant in the formula for $\hat{\beta}_a$ so that $\hat{\beta}_a$ is well defined). The test described in the theorem below uses $\hat{\beta}_a$ to test whether the rate of convergence is slow enough that the conservative rate $n^{1/2}$ should be used, and uses $\hat{\beta}$ to estimate the rate of convergence otherwise, as long as it is not implausibly large. If the rate of convergence is estimated to be larger than $\overline{\beta}$ (which, for large enough $\overline{\beta}$, will typically only occur on the interior of the identified set), the estimate is truncated to $\overline{\beta}$. When the rate of convergence is only known to be either $n^{\beta}$ for some $\beta \in [\underline{\beta}, \overline{\beta}]$, or either slower than $n^{\underline{\beta}}$ or faster than $n^{\overline{\beta}}$, this procedure provides a conservative approach that is still asymptotically exact when the exponent of the rate of convergence is in $(\underline{\beta}, \overline{\beta})$.

**Theorem 7.** *Suppose that Assumptions 2, 3 and 5 hold, and that $S$ is convex and $E(m(W_i, \theta)m(W_i, \theta)'|X$ $x)$ is continuous and strictly positive definite. Suppose that, for some $\overline{\gamma}$, Assumptions 1 and*

*4 hold with part (ii) of Assumption 1 replaced by Assumption 7 for some $\gamma \leq \overline{\gamma}$, where the set $\mathcal{X}_0 = \{x|\bar{m}_j(\theta, x) = 0$ some $j\}$ may be empty, or, for some $x_0 \in \mathcal{X}_0$ such that $X_i$ has a continuous density in a neighborhood of $x_0$ and $B < \infty$, $\bar{m}_j(\theta, x) \leq B\|x - x_0\|^\gamma$ for some $\gamma > \overline{\gamma}$ and some $j$.*

*Let $\overline{\beta} = (d_X + \underline{\gamma})/(d_X + 2\underline{\gamma})$ for some $\underline{\gamma} < \overline{\gamma}$ and let $\underline{\beta} = (d_X + \overline{\gamma})/(d_X + 2\overline{\gamma})$. Let $\hat{\beta}$, $\hat{\beta}_a$ and be defined as above for some $0 < \chi_1 < \chi_2 < 1$ and $0 < \chi_a < 1$. Consider the following test. If $\hat{\beta}_a \geq \underline{\beta}$, reject if $n^{\hat{\beta}\wedge\overline{\beta}} S(T_n(\theta)) > L_{n,b}(1 - \alpha|b^{\hat{\beta}\wedge\overline{\beta}})$ (or if $n^{\hat{\beta}\wedge\overline{\beta}} S(T_n(\theta)) > \tilde{L}_{n,b}(1 - \alpha|b^{\hat{\beta}\wedge\overline{\beta}})$) where $b = n^{\chi_3}$ for some $0 < \chi_3 < 1$. If $\hat{\beta}_a < \underline{\beta}$, perform any (possibly conservative) asymptotically level $\alpha$ test that compares $n^{1/2}S(T_n(\theta))$ to a critical value that is bounded away from zero.*

*Under these conditions, this test is asymptotically level $\alpha$. If Assumption 1 holds with part (ii) of Assumption 1 replaced by Assumption 7 for some $\underline{\gamma} < \gamma < \overline{\gamma}$ and $\mathcal{X}_0$ nonempty, this test will be asymptotically exact level $\alpha$.*

In the one dimensional case, the conditions of Theorem 7 follow immediately from smoothness assumptions on the conditional mean by Theorem 5. As discussed above, the condition that the minimum not be taken on the boundary of the support of $X_i$ could be removed, or the result can be used as stated with a pre-test for this condition.

**Theorem 8.** *Suppose that $d_X = 1$, Assumptions 2, 3 and 5 hold, and that $S$ is convex and $E(m(W_i, \theta)m(W_i, \theta)'|X_i = x)$ is continuous and strictly positive definite. Suppose that $supp(X_i) = [\underline{x}, \overline{x}]$ and that $\bar{m}(\theta, x)$ is bounded away from zero near $\underline{x}$ and $\overline{x}$ and has $p$ bounded derivatives. Then the conditions of Theorem 7 hold for any $\overline{\gamma} < p$.*

## 6.2 Tests Based on Estimating the Second Derivative

I make the following assumptions on the conditional mean and the distribution of $X_i$. These conditions are used to estimate the second derivatives of $\bar{m}(\theta, x) = E(m_j(W_i, \theta)|X_i = x)$, and the results are stated for local polynomial estimates. The conditions and results here are from Ichimura and Todd (2007). Other nonparametric estimators of conditional means and their derivatives and conditions for uniform convergence of such estimators could be used instead. The results in this section related to testing Assumption 1 are stated for $m_j(W_i, \theta)$ for a fixed index $j$. The consistency of a procedure that combines these tests for each $j$ then follows from the consistency of the test for each $j$.

**Assumption 8.** *The third derivatives of $\bar{m}_j(\theta, x)$ with respect to $x$ are Lipschitz continuous and uniformly bounded.*

**Assumption 9.** *$X_i$ has a uniformly continuous density $f_X$ such that, for some compact set $D \in \mathbb{R}^d$, $\inf_{x \in D} f_X(x) > 0$, and $E(m_j(W_i, \theta)|X_i)$ is bounded away from zero outside of $D$.*

**Assumption 10.** *The conditional density of $X_i$ given $m_j(W_i, \theta)$ exists and is uniformly bounded.*

Note that Assumption 10 is on the density of $X_i$ given $m_j(W_i, \theta)$, and not the other way around, so that, for example, count data for the dependent variable in an interval regression is okay.

Let $\mathcal{X}_0^j$ be the set of minimizers of $\bar{m}_j(\theta, x)$ if this function is less than or equal to 0 for some $x$ and the empty set otherwise. In order to test Assumption 1, I first note that, if the conditional mean is smooth, the positive definiteness of the second derivative matrix on the contact set will imply that the contact set is finite. This reduces the problem to determining whether the second derivative matrix is positive definite on the set of minimizers of $\bar{m}_j(\theta, x)$, a problem similar to testing local identification conditions in nonlinear models (see Wright, 2003). I record this observation in the following lemma.

**Lemma 1.** *Under Assumptions 8 and 9, if the second derivative matrix of $E(m_j(W_i, \theta)|X_i = x)$ is strictly positive definite on $\mathcal{X}_0^j$, then $\mathcal{X}_0^j$ must be finite.*

According to Lemma 1, once we know that the second derivative matrix of $E(m_j(W_i, \theta)|X_i)$ is positive definite on the set of minimizers $E(m_j(W_i, \theta)|X_i)$, the conditions of Theorem 1 will hold. This reduces the problem to testing the conditions of the lemma. One simple way of doing this is to take a preliminary estimate of $\mathcal{X}_0^j$ that contains this set with probability approaching one, and then test whether the second derivative matrix of $E(m_j(W_i, \theta)|X_i)$ is positive definite on this set. In what follows, I describe an approach based on local polynomial regression estimates of the conditional mean and its second derivatives, but other methods of estimating the conditional mean would work under appropriate conditions. The methods require knowledge of a set $D$ satisfying Assumption 9. This set could be chosen with another preliminary test, an extension which I do not pursue.

Under the conditions above, we can estimate $\bar{m}_j(\theta, x)$ and its derivatives at a given point $x$ with a local second order polynomial regression estimator defined as follows. For a kernel function $K$ and a bandwidth parameter $h$, run a regression of $m_j(W_i, \theta)$ on a second order polynomial of $X_i$, weighted by the distance of $X_i$ from $x$ by $K((X - x)/h)$. That is, for each

$j$ and any $x$, define $\hat{\bar{m}}_j(\theta, x)$, $\hat{\beta}_j(x)$, and $\hat{V}_j(x)$ to be the values of $m$, $\beta$, and $V$ that minimize

$$E_n \left\{ \left[ m_j(W_i, \theta) - \left( m + (X_i - x)'\beta + \frac{1}{2}(X_i - x)'V(X_i - x) \right) \right]^2 \times K((X_i - x)/h) \right\}.$$

The pre-test uses $\hat{\bar{m}}_j(\theta, x)$ as an estimate of $\bar{m}_j(\theta, x)$ and $\hat{V}_j(x)$ as an estimate of $V_j(x)$.

The following theorem, taken from Ichimura and Todd (2007, Theorem 4.1), gives rates of convergence for these estimates of the conditional mean and its second derivatives that will be used to estimate $\mathcal{X}_0^j$ and $V_j(x)$ as described above. The theorem uses an additional assumption on the kernel $K$.

**Assumption 11.** *The kernel function $K$ is bounded, has compact support, and satisfies, for some $C$ and for any $0 \leq j_1 + \cdots + j_r \leq 5$, $|u_1^{j_1} \cdots u_r^{j_r} K(u) - v_1^{j_1} \cdots v_r^{j_r} K(v)| \leq C\|u - v\|$.*

**Theorem 9.** *Under iid data and Assumptions 2, 8, 9, 10, and 11,*

$$\sup_{x \in D} \left| \hat{V}_{j,rs}(x) - V_{j,rs}(x) \right| = \mathcal{O}_p((\log n/(nh^{d_X+4}))^{1/2}) + \mathcal{O}_p(h)$$

*for all $r$ and $s$, where $V_{j,rs}$ is the $r, s$ element of $V_j$, and*

$$\sup_{x \in D} \left| \hat{\bar{m}}_j(\theta, x) - \bar{m}_j(\theta, x) \right| = \mathcal{O}_p((\log n/(nh^{d_X}))^{1/2}) + \mathcal{O}_p(h^3).$$

For both the conditional mean and the derivative, the first term in the asymptotic order of convergence is the variance term and the second is the bias term. The optimal choice of $h$ sets both of these to be the same order, and is $h_n = (\log n/n)^{1/(d_X+6)}$ in both cases. This gives a $(\log n/n)^{1/(d_X+6)}$ rate of convergence for the second derivative, and a $(\log n/n)^{3/(d_X+6)}$ rate of convergence for the conditional mean. However, any choice of $h$ such that both terms go to zero can be used.

In order to test the conditions of Lemma 1, we can use the following procedure. For some sequence $a_n$ growing to infinity such that $a_n[(\log n/(nh^{d_X}))^{1/2} \vee h^3]$ converges to zero, let $\hat{\mathcal{X}}_0^j = \{x \in D | \hat{\bar{m}}_j(\theta, x) - (\inf_{x' \in D} \hat{\bar{m}}_j(\theta, x') \wedge 0)| \leq [a_n(\log n/(nh^{d_X}))^{1/2} \vee h^3]\}$. By Theorem 9, $\hat{\mathcal{X}}_0^j$ will contain $\mathcal{X}_0^j$ with probability approaching one. Thus, if we can determine that $V_j(x)$ is positive definite on $\hat{\mathcal{X}}_0^j$, then, asymptotically, we will know that $V_j(x)$ is positive definite on $\mathcal{X}_0^j$. Note that $\hat{\mathcal{X}}_0^j$ is an estimate of the set of minimizers of $\overline{m}_j(x, \theta)$ over $x$ if the moment inequality binds or fails to hold, and is eventually equal to the empty set if the moment inequality is slack.

Since the determinant is a differentiable map from $\mathbb{R}^{d_X^2}$ to $\mathbb{R}$, the $\mathcal{O}_p((\log n/(nh^{d_X+4}))^{1/2})+\mathcal{O}_p(h)$ rate of uniform convergence for $\hat{V}_j(x)$ translates to the same (or faster) rate of convergence for $\det \hat{V}_j(x)$. If, for some $x_0 \in \mathcal{X}_0^j$, $V_j(x_0)$ is not positive definite, then $V_j(x_0)$ will be singular (the second derivative matrix at an interior minimum must be positive semidefinite if the second derivatives are continuous in a neighborhood of $x_0$), and $\det V_j(x_0)$ will be zero. Thus, $\inf_{x\in\hat{\mathcal{X}}_0^j} \det \hat{V}_j(x) \leq \det \hat{V}_j(x_0) = \mathcal{O}_p((\log n/(nh^{d_X+4}))^{1/2}) + \mathcal{O}_p(h)$ where the inequality holds with probability approaching one. Thus, letting $b_n$ be any sequence going to infinity such that $b_n[(\log n/(nh^{d_X+4}))^{1/2} \vee h]$ converges to zero, if $V_j(x_0)$ is not positive definite for some $x_0 \in \mathcal{X}_0^j$, we will have $\inf_{x\in\hat{\mathcal{X}}_0^j} \det \hat{V}_j(x) \leq b_n[(\log n/(nh^{d_X+4}))^{1/2} \vee h]$ with probability approaching one (actually, since we are only dealing with the point $x_0$, we can use results for pointwise convergence of the second derivative of the conditional mean, so the $\log n$ term can be replaced by a constant, but I use the uniform convergence results for simplicity).

Now, suppose $V_j(x)$ is positive definite for all $x \in \mathcal{X}_0^j$. By Lemma 1, we will have, for some $B > 0$, $\det V_j(x) \geq B$ for all $x \in \mathcal{X}_0^j$. By continuity of $V_j(x)$, we will also have, for some $\varepsilon > 0$, $\det V_j(x) \geq B/2$ for all $x \in \mathcal{X}_0^{j\varepsilon}$ where $\mathcal{X}_0^{j\varepsilon} = \{x | \inf_{x'\in\mathcal{X}_0^j} \|x - x'\| \leq \varepsilon\}$ is the $\varepsilon$-expansion of $\mathcal{X}_0^j$. Since $\hat{\mathcal{X}}_0^j \subseteq \mathcal{X}_0^{j\varepsilon}$ with probability approaching one, we will also have $\inf_{x\in\hat{\mathcal{X}}_0^j} \det V_j(x) \geq B/2$ with probability approaching one. Since $\det \hat{V}_j(x) \to \det V_j(x)$ uniformly over $D$, we will then have $\inf_{x\in\hat{\mathcal{X}}_0^j} \det \hat{V}_j(x) \geq b_n[(\log n/(nh^{d_X+4}))^{1/2} \vee h]$ with probability approaching one.

This gives the following theorem.

**Theorem 10.** *Let $\hat{V}_j(x)$ and $\hat{\bar{m}}_j(\theta, x)$ be the local second order polynomial estimates defined with some kernel $K$ with $h$ such that the rate of convergence terms in Theorem 9 go to zero. Let $\hat{\mathcal{X}}_0^j$ be defined as above with $a_n[(\log n/(nh^{d_X}))^{1/2} \vee h^3]$ going to zero and $a_n$ going to infinity, and let $b_n$ be any sequence going to infinity such that $b_n[(\log n/(nh^{d_X+4}))^{1/2} \vee h]$ goes to zero. Suppose that Assumptions 2, 8, 9, 10, and 11, hold, and the null hypothesis holds with $E(m(W_i,\theta)m(W_i,\theta)'|X_i = x)$ continuous and the data are iid. Then, if Assumption 1 holds, we will have $\inf_{x\in\hat{\mathcal{X}}_0^j} \det \hat{V}_j(x) > b_n[(\log n/(nh^{d_X+4}))^{1/2} \vee h]$ for each $j$ with probability approaching one. If Assumption 1 does not hold, we will have $\inf_{x\in\hat{\mathcal{X}}_0^j} \det \hat{V}_j(x) \leq b_n[(\log n/(nh^{d_X+4}))^{1/2} \vee h]$ for some $j$ with probability approaching one.*

The purpose of this test of Assumption 1 is as a preliminary consistent test in a procedure that uses the asymptotic approximation in Theorem 1 if the test finds evidence in favor of Assumption 1, and uses the methods that are robust to different types of contact sets, but possibly conservative, such as those described in Andrews and Shi (2009), otherwise. It follows from Theorem 10 that such a procedure will have the correct size asymptotically. In the statement of the following theorem, it is understood that Assumptions 4 and 6, which refer to objects in Assumption 1, do not need to hold if the data generating process is such that Assumption 1 does not hold.

**Theorem 11.** *Consider the following test. For some $b_n \to \infty$ and $h \to 0$ satisfying the conditions of Theorem 10, perform a pre-test that finds evidence in favor of Assumption 1 iff.* $\inf_{x\in\hat{\mathcal{X}}_0} \det \hat{V}_j(x) \geq b_n[(\log n/(nh^{d_X+4}))^{1/2} \vee h]$ *for each $j$. If $\hat{\mathcal{X}}_0 = \emptyset$, do not reject the null hypothesis that $\theta \in \Theta_0$. If $\inf_{x\in\hat{\mathcal{X}}_0} \det \hat{V}_j(x) > b_n[(\log n/(nh^{d_X+4}))^{1/2} \vee h]$ for each $j$, reject the null hypothesis that $\theta \in \Theta_0$ if $n^{(d_X+2)/(d_X+4)}S(T_n(\theta)) > \hat{q}_{1-\alpha}$ where $\hat{q}_{1-\alpha}$ is an estimate of the $1-\alpha$ quantile of the distribution of $S(Z)$ formed using one of the methods in Section 4. If $\inf_{x\in\hat{\mathcal{X}}_0} \det \hat{V}_j(x) \leq b_n[(\log n/(nh^{d_X+4}))^{1/2} \vee h]$ for some $j$, perform any (possibly conservative) asymptotically level $\alpha$ test. Suppose that Assumptions 2, 3, 4, 5, 8, 9, 10, and 11 hold, $E(m(W_i,\theta)m(W_i,\theta)'|X_i = x)$ is continuous, and the data are iid. Then this provides an asymptotically level $\alpha$ test of $\theta \in \Theta_0$ if the subsampling procedure is used or if Assumption 6 holds and the procedure based on estimating the asymptotic distribution directly is used. If Assumption 1 holds, this test is asymptotically exact.*

The estimates used for this pre-test can also be used to construct estimates of the quantities in Assumption 6 that satisfy the consistency requirements of this assumption. Suppose that we have estimates $\hat{M}(x)$, $\hat{f}_X(x)$, and $\hat{V}(x)$ of $E(m(W_i,\theta)m(W_i,\theta)'|X = x)$, $f_X(x)$, and $V(x)$ that are consistent uniformly over $x$ in a neighborhood of $\mathcal{X}_0$. Then, if we have estimates of $\mathcal{X}_0$ and $J(k)$, we can estimate the quantities in Assumption 6 using $\hat{M}_k(x_k)$, $\hat{f}_X(x_k)$, and $\hat{V}(x_k)$ for each $x_k$ in the estimate of $\mathcal{X}_0$, where $\hat{M}_k(x_k)$ is a sparse version of $\hat{M}(x_k)$ with elements with indices not in the estimate of $J(k)$ set to zero.

The estimate $\hat{\mathcal{X}}_0$ contains infinitely many points, so it will not work for this purpose. Instead, define the estimate $\tilde{\mathcal{X}}_0$ of $\mathcal{X}_0$ and the estimate $\hat{J}(k)$ of $J(k)$ as follows. Let $a_n$ be as in Theorem 10, and let $\varepsilon_n^2 \to 0$ more slowly than $a_n[(\log n/(nh^{d_X}))^{1/2} \vee h^3]$. Let $\hat{\ell}_j$ be the smallest number such that $\hat{\mathcal{X}}_0^j \subseteq \cup_{k=1}^{\hat{\ell}_j} B_{\varepsilon_n}(\hat{x}_{j,k})$ for some $\hat{x}_{j,1}, \ldots, \hat{x}_{j,\hat{\ell}_j}$. Define an equivalence relation $\sim$ on the set $\{(j,k)|1 \leq j \leq d_Y, 1 \leq k \leq \hat{\ell}_j\}$ by $(j,k) \sim (j',k')$ iff. there is a sequence $(j,k) = (j_1,k_1), (j_2,k_2), \ldots, (j_r,k_r) = (j',k')$ such that $B_{\varepsilon_n}(\hat{x}_{j_s,k_s}) \cap B_{\varepsilon_n}(\hat{x}_{j_{s+1},k_{s+1}}) \neq \emptyset$ for $s$ from 1 to $r-1$. Let $\hat{\ell}$ be the number of equivalence classes, and, for each equivalence class, pick exactly one $(j,k)$ in the equivalence class and let $\tilde{x}_r = \hat{x}_{j,k}$ for some $r$ between 1 and $\hat{\ell}$. Define the estimate of the set $\mathcal{X}_0$ to be $\tilde{\mathcal{X}}_0 \equiv \{\tilde{x}_1, \ldots, \tilde{x}_{\hat{\ell}}\}$, and define the estimate $\hat{J}(r)$ for $r$ from 1 to $\hat{\ell}$ to be the set of indices $j$ for which some $(j,k)$ is in the same equivalence class as $\tilde{x}_r$.

Although these estimates of $\mathcal{X}_0$, $\ell$, and $J(1), \ldots, J(\ell)$ require some cumbersome notation to define, the intuition behind them is simple. Starting with the initial estimates $\hat{\mathcal{X}}_j$, turn these sets into discrete sets of points by taking the centers of balls that contain the sets $\hat{\mathcal{X}}_j$ and converge at a slower rate. This gives estimates of the points at which the conditional moment inequality indexed by $j$ binds for each $j$, but to estimate the asymptotic distribution in Theorem 1, we also need to determine which components, if any, of $\bar{m}(\theta, x)$ bind at the same value of $x$. The procedure described above does this by testing whether the balls used to form the estimated contact points for each index of $\bar{m}(\theta, x)$ intersect across indices.

The following theorem shows that this is a consistent estimate of the set $\mathcal{X}_0$ and the indices of the binding moments.

**Theorem 12.** *Suppose that Assumptions 1, 8, 9, 10, and 11 hold. For the estimates $\tilde{\mathcal{X}}_0$, $\hat{\ell}$ and $\hat{J}(r)$, $\hat{\ell} = \ell$ with probability approaching one and, for some labeling of the indices of $\tilde{x}_1, \ldots, \tilde{x}_{\hat{\ell}}$ we have, for $k$ from 1 to $\ell$, $\tilde{x}_k \xrightarrow{p} x_k$ and, with probability approaching one, $\hat{J}(k) = J(k)$.*

An immediate consequence of this is that this estimate of $\mathcal{X}_0$ can be used in combination with consistent estimates of $E(m(W_i, \theta)m(W_i, \theta)'|X = x)$, $f_X(x)$, and $V(x)$ to form estimates of these functions evaluated at points in $\mathcal{X}_0$ that satisfy the assumptions needed for the procedure for estimating the asymptotic distribution described in Section 4.

**Corollary 3.** *If the estimates $\hat{M}_k(x)$, $\hat{f}_X(x)$, and $\hat{V}(x)$ are consistent uniformly over $x$ in a neighborhood of $\mathcal{X}_0$, then, under Assumptions 1, 8, 9, 10, and 11, the estimates $\hat{M}_k(\tilde{x}_k)$, $\hat{f}_X(\tilde{x}_k)$, and $\hat{V}_j(\tilde{x}_k)$ satisfy Assumption 6.*

# 7 Local Alternatives

Consider local alternatives of the form $\theta_n = \theta_0 + a_n$ for some fixed $\theta_0$ such that $m(W_i, \theta_0)$ satisfies Assumption 1 and $a_n \to 0$. Here, I keep the data generating process fixed and vary the parameter being tested. Similar ideas will apply when the parameter is fixed and the data generating process is changed so that the parameter approaches the identified set. Throughout this section, I restrict attention to the conditions in Section 3, which corresponds to the more general setup in Section 5 with $\gamma = 2$. To translate the $a_n$ rate of convergence to $\theta_0$ to a rate of convergence for the sequence of conditional means, I make the following assumptions. As before, define $\bar{m}(\theta, x) = E(m(W_i, \theta)|X_i = x)$.

**Assumption 12.** *For each $x_k \in \mathcal{X}_0$, $\bar{m}(\theta, x)$ has a derivative as a function of $\theta$ in a neighborhood of $(\theta_0, x_k)$, denoted $\bar{m}_\theta(\theta, x)$, that is continuous as a function of $(\theta, x)$ at $(\theta_0, x_k)$ and, for any neighborhood of $x_k$, there is a neighborhood of $\theta_0$ such that $\bar{m}_j(\theta, x)$ is bounded away from zero for $\theta$ in the given neighborhood of $\theta_0$ and $x$ outside of the given neighborhood of $x_k$ for $j \in J(k)$ and for all $x$ for $j \notin J(k)$.*

**Assumption 13.** *For each $x_k \in \mathcal{X}_0$ and $j \in J(k)$, $E\{[m_j(W_i, \theta) - m_j(W_i, \theta_0)]^2 | X_i = x\}$ converges to zero uniformly in $x$ in some neighborhood of $x_k$ as $\theta \to \theta_0$.*

I also make the following assumption, which extends Assumption 2 to a neighborhood of $\theta_0$.

**Assumption 14.** *For some fixed $\overline{Y} < \infty$ and $\theta$ in a some neighborhood of $\theta_0$, $|m(W_i, \theta)| \leq \overline{Y}$ with probability one.*

In the interval regression example, these conditions are satisfied as long as Assumption 1 holds at $\theta_0$ and the data have finite support. These conditions are also likely to hold in a variety of models once Assumption 1 holds at $\theta_0$. Note that smoothness conditions are in terms of the conditional mean $\bar{m}(\theta, x)$, rather than $m(W_i, \theta)$, so that the conditions can still hold when the sample moments are nonsmooth functions of $\theta$.

Set $a_n = b_n a$ for some sequence of scalars $b_n \to 0$ and a constant vector $a$. Going through the argument for Theorem 1, the variance term in the local process is now

$$\begin{aligned}
&\frac{\sqrt{n}}{\sqrt{h_n^{d_X}}}(E_n - E)m(W_i, \theta_0 + b_n a)I(h_n s < X - x_k < h_n(s+t)) \\
&= \frac{\sqrt{n}}{\sqrt{h_n^{d_X}}}(E_n - E)m(W_i, \theta_0)I(h_n s < X - x_k < h_n(s+t)) \\
&+ \frac{\sqrt{n}}{\sqrt{h_n^{d_X}}}(E_n - E)[m(W_i, \theta_0 + b_n a) - m(W_i, \theta_0)]I(h_n s < X - x_k < h_n(s+t)).
\end{aligned}$$

The first term is the variance term under the null, and the second term should be small under Assumption 13.

As for the drift term,

$$\begin{aligned}
&\frac{1}{h_n^{d_X+2}} Em(W_i,\theta+b_n a)I(h_n s < X_i - x_k < h_n(s+t)) \\
&= \frac{1}{h_n^{d_X+2}} Em(W_i,\theta)I(h_n s < X_i - x_k < h_n(s+t)) \\
&+ \frac{1}{h_n^{d_X+2}} E[m(W_i,\theta+b_n a) - m(W_i,\theta)]I(h_n s < X_i - x_k < h_n(s+t)).
\end{aligned}$$

The first term is the drift term under the null. The second term is

$$\begin{aligned}
&\frac{1}{h_n^{d_X+2}} E[\bar{m}(\theta+b_n a, X_i) - \bar{m}(\theta, X_i)]I(h_n s < X_i - x_k < h_n(s+t)) \\
&\approx \frac{1}{h_n^{d_X+2}} E b_n \bar{m}_\theta(\theta, X_i) a I(h_n s < X_i - x_k < h_n(s+t)) \\
&\approx \frac{b_n}{h_n^{d_X+2}} f_X(x_k)\bar{m}_\theta(\theta,x_k)a \int_{h_n s < x - x_k < h_n(s+t)} dx = \frac{b_n}{h_n^2} f_X(x_k)\bar{m}_\theta(\theta,x_k)a \prod_i t_i.
\end{aligned}$$

Setting $b_n = h_n^2 = n^{-2/(d_X+4)}$ gives a constant that does not change with $n$, so we should expect to have power against $n^{-2/(d_X+4)}$ alternatives. The following theorem formalizes these ideas.

**Theorem 13.** *Let $\theta_0$ be such that $E(m(W_i,\theta_0)|X_i) \geq 0$ almost surely and Assumptions 1, 12, 13, and 14 are satisfied for $\theta_0$. Let $a \in \mathbb{R}^{d_\theta}$ and let $a_n = an^{-2/(d_X+4)}$. Let $Z(a)$ be a random variable defined the same way as $Z$ in Theorem 1, but with the functions $g_{P,x_k,j}(s,t)$ replaced by the functions*

$$g_{P,x_k,j,a}(s,t) = \frac{1}{2} f_X(x_k) \int_{s<x<s+t} x' V_j(x_k) x \, dx + \bar{m}_{\theta,j}(\theta_0,x_k) a f_X(x_k) \prod_i t_i$$

*for $j \in J(k)$ for each $k$ where $\bar{m}_{\theta,j}$ is the $j$th row of the derivative matrix $\bar{m}_\theta$. Then*

$$n^{(d_X+2)/(d_X+4)} \inf_{s,t} E_n m(W_i,\theta+a_n)I(s < X_i < s+t) \xrightarrow{d} Z(a).$$

Thus, an exact test gives power against $n^{-2/(d_X+4)}$ alternatives (as long as $\bar{m}_{\theta,j}(\theta_0,x_k)a$ is negative for each $j$ or negative enough for at least one $j$), but not against alternatives that converge strictly faster. The dependence on the dimension of $X_i$ is a result of the curse of dimensionality. With a fixed amount of “smoothness,” the speed at which local alternatives can converge to the null space and still be detected is decreasing in the dimension of $X_i$.

Now consider power against local alternatives of this form, with a possibly different sequence $a_n$, using the conservative estimate that $\sqrt{n}\inf_{s,t} E_n m(W_i,\theta)I(s < X_i < s+t) \xrightarrow{p} 0$ for $\theta \in \Theta_0$. That is, we fix some $\eta > 0$ and reject if $\sqrt{n}S(\inf_{s,t} E_n m(W_i, \theta_0 + a_n)I(s < X_i < s+t)) > \eta$. For the drift term $Em_j(W_i, \theta_0+a_n)I(s < X_i - x_k < s+t)$ of the local alternative, we have, for $t$ near zero and $s$ near any $x_k \in \mathcal{X}_0$,

$$\begin{aligned}
&\sqrt{n}Em_j(W_i,\theta_0 + a_n)I(s < X_i - x_k < s + t)\\
&\approx \sqrt{n}E[\overline{m}_j(\theta_0, X_i) + \overline{m}_{\theta,j}(\theta_0, X_i)a_n]I(s < X_i - x_k < s + t)\\
&\approx \sqrt{n}f_X(x_k)\int_{s<x<s+t}\left(\frac{1}{2}x'V_j(x_k)x + \overline{m}_{\theta,j}(\theta_0, x_k)a_n\right) dx.
\end{aligned}$$

For any $a$ and $b$,

$$\begin{aligned}
&f_X(x_k)\int_{s<x<s+t}\left(\frac{1}{2}x'Vx + a\right) dx = (a/b)f_X(x_k)\int_{s<x<s+t}\left\{\frac{1}{2}[(b/a)^{1/2}x]'V[(b/a)^{1/2}x] + b\right\} dx\\
&= (a/b)f_X(x_k)\int_{(b/a)^{1/2}s<x<(b/a)^{1/2}(s+t)}\left(\frac{1}{2}u'Vu + b\right)(b/a)^{-d_X/2}du.
\end{aligned}$$

For any $(s,t)$, the last line in the display is equal to $(a/b)^{(d_X+2)/2}$ times the first expression in the display evaluated at a different value of $(s,t)$ with $a$ replaced with $b$. It follows that the minimized expression for $b$ is $(a/b)^{(d_X+2)/2}$ times the minimized expression for $a$. Thus, if $a_n = ab_n$, the drift term is of order $\sqrt{n}b_n^{(d_X+2)/2}$, so we should expect to have power against local alternatives with $\sqrt{n}b_n^{(d_X+2)/2} = \mathcal{O}(1)$ or $b_n = n^{-1/(d_X+2)}$ (note that setting $n^{(d_X+2)/(d_X+4)}b_n^{(d_X+2)/2} = \mathcal{O}(1)$ so that the drift term is of the same order of magnitude as the exact rate of convergence gives the $n^{-2/(d_X+4)}$ rate derived in the previous theorem for the exact test). Since the infimum of the drift term is taken at a point where $t$ is small, we should expect the mean zero term to converge at a faster than $\sqrt{n}$ rate, so that the limiting distribution will be degenerate. This is formalized in the following theorem.

**Theorem 14.** *Let $\theta_0$ be such that $E(m(W_i,\theta_0)|X_i) \geq 0$ almost surely and Assumptions 1, 12, 13, and 14 are satisfied for $\theta_0$. Let $a \in \mathbb{R}^{d_\theta}$ and let $a_n = an^{-1/(d_X+2)}$. Then, for each $j$,*

$$\begin{aligned}
&\sqrt{n}\inf_{s,t} E_n m_j(W_i, \theta_0 + a_n)I(s < X < s + t)\\
&\xrightarrow{p} \min_{k\ s.t.\ j\in J(k)}\inf_{s,t} f_X(x_k)\int_{s<x<s+t}\left(\frac{1}{2}x'Vx + \overline{m}_{\theta,j}(\theta_0, x_k)a\right) dx.
\end{aligned}$$

The $n^{-1/(d_X+2)}$ rate is slower than the $n^{-2/(d_X+4)}$ rate for detecting local alternatives with the asymptotically exact test. As with the asymptotically exact tests, the conservative tests do worse against this form of local alternative as the dimension of the conditioning variable $X_i$ increases.

# 8 Monte Carlo

I perform a monte carlo study to examine the finite sample behavior of the tests I propose, and to see how well the asymptotic results in this paper describe the finite sample behavior of KS statistics. First, I simulate the distribution of KS statistics for various sample sizes under parameter values and data generating processes that satisfy Assumption 1, and for data generating processes that lead to a $\sqrt{n}$ rate of convergence. As predicted by Theorem 1, for the data generating process that satisfies Assumption 1, the distribution of the KS statistic is roughly stable across sample sizes when scaled up by $n^{(d_X+2)/(d_X+4)}$. For the data generating process that leads to $\sqrt{n}$ convergence, scaling by $\sqrt{n}$ gives a distribution that is stable across sample sizes. Next, I examine the size and power of KS statistic based tests using the asymptotic distributions derived in this paper. I include procedures that test between the conditions leading to $\sqrt{n}$ convergence and the faster rates derived in this paper using the subsampling estimates of the rate of convergence described in Section 6.1, as well as infeasible procedures that use prior knowledge of the correct rate of convergence to estimate the asymptotic distribution.

## 8.1 Monte Carlo Designs

Throughout this section, I consider two monte carlo designs for a mean regression model with missing data. In this model, the latent variable $W_i^*$ satisfies $E(W_i^*|X_i) = \theta_1 + \theta_2 X_i$, but $W_i^*$ is unobserved, and can only be bounded by the observed variables $W_i^H = \overline{w} I(W_i^* \text{ missing}) + W_i^* I(W_i^* \text{ observed})$ and $W_i^L = \underline{w} I(W_i^* \text{ missing}) + W_i^* I(W_i^* \text{ observed})$ are observed, where $[\underline{w}, \overline{w}]$ is an interval known to contain $W_i^*$. The identified set $\Theta_0$ is the set of values of $(\theta_1, \theta_2)$ such that the moment inequalities $E(W_i^H - \theta_1 - \theta_2 X_i|X_i) \geq 0$ and $E(\theta_1 + \theta_2 X_i - W_i^L|X_i) \geq 0$ hold with probability one. For both designs, I draw $X_i$ from a uniform distribution on $(-1, 1)$ (here, $d_X = 1$). Conditional on $X_i$, I draw $U_i$ from an independent uniform $(-1, 1)$ distribution, and set $W_i^* = \theta_{1,*} + \theta_{2,*} X_i + U_i$, where $\theta_{1,*} = 0$ and $\theta_{2,*} = .1$. I then set $W_i^*$ to be missing with probability $p^*(X_i)$ for some function $p^*$ that differs across designs. I set $[\underline{w}, \overline{w}] = [-.1 - 1, .1 + 1] = [-1.1, 1.1]$, the unconditional support of $W_i^*$. Note that, while the data are generated using a particular value of $\theta$ in the identified set and a censoring process that satisfies the missing at random assumption (that the probability of data missing conditional on $(X_i, W_i^*)$ does not depend on $W_i^*$), the data generating process is consistent with forms of endogenous censoring that do not satisfy this assumption. The identified set contains all values of $\theta$ for which the data generating process is consistent with the latent variable model for $\theta$ and some, possibly endogenous, censoring mechanism.

The shape of the conditional moment inequalities as a function of $X_i$ depends on $p^*$. For Design 1, I set $p^*(x) = (0.9481x^4 + 1.0667x^3 - 0.6222x^2 - 0.6519x + 0.3889) \wedge 1$. The coefficients of this quartic polynomial were chosen to make $p^*(x)$ smooth, but somewhat wiggly, so that the quadratic approximation to the resulting conditional moments used in Theorem 1 will not be good over the entire support of $X_i$. The resulting conditional means of the bounds on $W_i^*$ are $E(W_i^L|X_i = x) = (1-p^*(x))(\theta_{1,*}+\theta_{2,*}x)+p^*(x)\underline{w}$ and $E(W_i^H|X_i = x) = (1 - p^*(x))(\theta_{1,*} + \theta_{2,*}x) + p^*(x)\overline{w}$. In the monte carlo study, I examine the distribution of the KS statistic for the upper inequality at $(\theta_{1,D1}, \theta_{2,D1}) \equiv (1.05, .1)$, a parameter value on the boundary of the identified set for which Assumption 1 holds, along with confidence intervals for the intercept parameter $\theta_1$ with the slope parameter $\theta_2$ fixed at .1. For the confidence regions, I also restrict attention to the moment inequality corresponding to $W_i^H$, so that the confidence regions are for the one sided model with only this conditional moment inequality. Figure 3 plots the conditional means of $W_i^H$ and $W_i^L$, along with the regression line corresponding to $\theta = (1.05, .1)$. The confidence intervals for the slope parameter invert a family of tests corresponding to values of $\theta$ that move this regression line vertically.

For Design 2, I set $p^*(x) = [(|x - .5| \vee .25) - .15] \wedge .7)$. Figure 4 plots the resulting conditional means. For this design, I examine the distribution of the KS statistic for the upper inequality at $(\theta_{1,D2}, \theta_{2,D2}) = (1.1, .9)$, which leads to a positive probability contact set for the upper moment inequality and a $n^{1/2}$ rate of convergence to a nondegenerate distribution. The regression line corresponding to this parameter is plotted in Figure 4 as well. For this design, I form confidence intervals for the slope parameter $\theta_1$ with $\theta_2$ fixed at .9, using the KS statistic for the moment inequality for $W_i^H$.

The confidence intervals reported in this section are computed by inverting KS statistic based tests on a grid of parameter values. I use a grid with meshwidth .01 that covers the area of the parameter space with distance to the boundary of the identified set no more than 1. In practice, monotonicity of the KS statistic in certain parameters (in this case, the KS statistic for each moment inequality is monotonic in the intercept parameter) can often be used to get a rough estimate of the boundary of the identified set before mapping out the confidence region exactly. In this case, a rough estimate of the boundary of the identified set for the intercept parameter could be formed by finding the point where the KS statistic for the moment inequality for $W_i^H$ crosses a fixed critical value before performing the test with critical values estimated for each value of $\theta$. All of the results in this section use 1000 monte carlo draws for each sample size and monte carlo design.

## 8.2 Distribution of the KS Statistic

To examine how well Theorem 1 describes the finite sample distribution of KS statistics under Assumption 1, I simulate from Design 1 for a range of sample sizes and form the KS statistic for testing $(\theta_{1,D1}, \theta_{2,D1})$. Since Assumption 1 holds for testing this value of $\theta$ under this data generating process, Theorem 1 predicts that the distribution of the KS statistic scaled up by $n^{(d_X+2)/(d_X+4)} = n^{3/5}$ should be similar across the sample sizes. The performance of this asymptotic prediction in finite samples is examined in Figure 5, which plots histograms of the scaled KS statistic $n^{3/5}S(T_n(\theta))$ for the sample sizes $n \in \{100, 500, 1000, 2000, 5000\}$. The scaled distributions appear roughly stable across sample sizes, as predicted.

In contrast, under Design 2, the KS statistic for testing $(\theta_{1,D2}, \theta_{2,D2})$ will converge at a $n^{1/2}$ rate to a nondegenerate distribution. Thus, asymptotic approximation suggests that, in this case, scaling by $n^{1/2}$ will give a distribution that is roughly stable across sample sizes. Figure 6 plots histograms of the scaled statistic $n^{1/2}S(T_n(\theta))$ for this case. The scaling suggested by asymptotic approximations appears to give a distribution that is stable across sample sizes here as well.

## 8.3 Finite Sample Performance of the Tests

I now turn to the finite sample performance of confidence regions for the identified set based on critical values formed using the asymptotic approximations derived in this paper, along with possibly conservative confidence regions that use the $n^{1/2}$ approximation. The critical values use subsampling with different assumed rates of convergence. I report results for the tests based on subsampling estimates of the rate of convergence described in Section 6.1, tests that use the conservative rate $n^{1/2}$, and infeasible tests that use a $n^{3/5}$ rate under Design 1, and a $n^{1/2}$ rate under Design 2. The implementation details are as follows. For the critical values using the conservative rate of convergence, I estimate the .9 and .95 quantiles of the distribution of the KS statistic at each value of $\theta$ using subsampling, and add the correction factor .001 to prevent the critical value from going to zero. The critical values using estimated rates of convergence are computed as described in Section 6.1. I use the subsample sizes

$b_1 = \lceil n^{1/2} \rceil$ and $b_2 = \lceil n^{1/3} \rceil$ to estimate the rate of convergence $\hat{\beta}$ for subsampling, and $b_2 = 5$ for the rate estimate $\hat{\beta}_a$ that is used to test whether the conservative rate should be used. For both rate estimates, I average the estimates computed using the quantiles .5, .9, and .95. For the upper and lower truncation points for the rate of convergence, I use $\underline{\beta} = .55$ and $\overline{\beta} = 2/3$. These truncation points allow for exact inference for values of $\theta$ such that Assumption 7 holds with $\gamma = 2$ (twice differentiable conditional mean) or $\gamma = 1$ (directional derivatives from both sides). The upper truncation point $\overline{\beta}$ corresponds to $\gamma = 1$, and the lower truncation point $\underline{\beta}$ is halfway between the rate of convergence exponent $3/5$ for $\gamma = 2$, and the conservative rate exponent $1/2$. In addition, I truncate $\hat{\beta}$ from below at $1/2$ in cases where $\hat{\beta} < 1/2$. For both the conservative and estimated rates of convergence, I use the uncentered subsampling estimate with subsample size $\lceil n^{1/2} \rceil$. All subsampling estimates use 1000 subsample draws. For values of $\theta$ such that the pre-test finds that the conservative approximation should be used ($\hat{\beta}_a < \underline{\beta}$), I use the same method of estimating the critical values as in the tests that always use the conservative rate of convergence.

Table 1 reports the coverage probabilities for $(\theta_{1,D1}, \theta_{2,D1})$ under Design 1. As discussed above, under Design 1, $(\theta_{1,D1}, \theta_{2,D1})$ is on the boundary of the identified set and satisfies Assumption 1. As predicted, the tests that subsample with the $n^{1/2}$ rate are conservative. The nominal 95% confidence regions that use the $n^{1/2}$ rate cover $(\theta_{1,D1}, \theta_{2,D1})$ with probability at least .99 for all of the sample sizes. Subsampling with the exact $n^{3/5}$ rate of convergence, an infeasible procedure that uses prior knowledge that Assumption 1 holds under $(\theta_{1,D1}, \theta_{2,D1})$ for this data generating process, gives confidence regions that cover $(\theta_{1,D1}, \theta_{2,D1})$ with probability much closer to the nominal coverage. The subsampling tests with the estimated rate of convergence also perform well, attaining close to the nominal coverage.

Table 2 reports coverage probabilities for testing $(\theta_{1,D2}, \theta_{2,D2})$ under Design 2. In this case, subsampling with a $n^{1/2}$ rate gives an asymptotically exact test of $(\theta_{1,D2}, \theta_{2,D2})$, so we should expect the coverage probabilities for the tests that use the $n^{1/2}$ rate of convergence to be close to the nominal coverage probabilities, rather than being conservative. The coverage probabilities for the $n^{1/2}$ rate are generally less conservative here than for Design 1, as the asymptotic approximations predict, although the coverage is considerably greater than the nominal coverage, even with 5000 observations. In this case, the infeasible procedure is identical to the conservative test, since the exact rate of convergence is $n^{1/2}$. The confidence regions that use subsampling with the estimated rate contain $(\theta_{1,D2}, \theta_{2,D2})$ with probability close to the nominal coverage, but are generally more liberal than their nominal level.

Given that subsampling with the estimated rate increases type I error by having coverage probability close to the nominal coverage probability rather than being conservative, we should expect a decrease in type II error. The results in Section 7 show that critical values based on the exact $n^{3/5}$ rate of convergence lead to tests that detect local alternatives that approach the identified set at a $n^{2/(d_X+4)} = n^{2/5}$ rate, while the conservative tests detect local alternatives that approach the identified set at a slower $n^{1/(d_X+2)} = n^{1/3}$ rate. For confidence regions that invert these tests, this is reflected in the portion of the parameter space the confidence region covers outside of the true identified set.

Tables 3 and 4 summarize the portion of the parameter space outside of the identified set covered by confidence intervals for the intercept parameter $\theta_1$ with $\theta_2$ fixed at $\theta_{2,D1}$ for Design 1 and $\theta_{2,D2}$ for Design 2. The entries in each table report the upper endpoint of one of the confidence regions minus the upper endpoint of the identified set for the slope parameter, averaged over the monte carlo draws. As discussed above, the true upper endpoint of the identified set for $\theta_1$ under Design 1 with $\theta_2$ fixed at $\theta_{2,D1}$ is $\theta_{1,D1}$, and the true upper endpoint of the identified set for $\theta_1$ under Design 2 with $\theta_2$ fixed at $\theta_{2,D2}$ is $\theta_{1,D2}$, so, letting $\hat{u}_{1-\alpha}$ be the greatest value of $\theta_1$ such that $(\theta_1, \theta_{2,D1})$ is not rejected, Table 3 reports averages of $\hat{u}_{1-\alpha} - \theta_{2,D1}$, and similarly for Table 4 and Design 2.

The results of Section 7 suggest that, for the results for Design 1 reported in Table 3, the difference between the upper endpoint of the confidence region and the upper endpoint of the identified set should decrease at a $n^{2/5}$ rate for the critical values that use or estimate the exact rate of convergence (the first and third rows), and a $n^{1/3}$ rate for subsampling with the conservative rate and adding .001 to the critical value (the second row). This appears roughly consistent with the values reported in these tables. The conservative confidence regions start out slightly larger, and then converge more slowly. For Design 2, the KS statistic converges at a $n^{1/2}$ rate on the boundary of the identified set for $\theta_1$ for $\theta_2$ fixed at $\theta_{2,D2}$, and arguments in Andrews and Shi (2009) show that $n^{1/2}$ approximation to the KS statistic give power against sequences of alternatives that approach the identified set at a $n^{1/2}$ rate. The confidence regions do appear to shrink to the identified set at approximately this rate over most sample sizes, although the decrease in the width of the confidence region is larger than predicted for smaller sample sizes, perhaps reflecting time taken by the subsampling procedures to find the binding moments.

# 9 Illustrative Empirical Application

As an illustrative empirical application, I apply the methods in this paper to regressions of out of pocket prescription drug spending on income using data from the Health and Retirement Study (HRS). In this survey, respondents who did not report point values for these and other variables were asked whether the variables were within a series of brackets, giving point values for some observations and intervals of different sizes for others. The income variable used here is taken from the RAND contribution to the HRS, which adds up reported income from different sources elicited in the original survey. For illustrative purposes, I focus on the subset of respondents who report point values for income, so that only prescription drug spending, the dependent variable, is interval valued. The resulting confidence regions are valid under any potentially endogenous process governing the size of the reported interval for prescription expenditures, but require that income be missing or interval reported at random. Methods similar to those proposed in this paper could also be used along with the results of Manski and Tamer (2002) for interval reported covariates to use these additional observations to potentially gain identifying power (but still using an assumption of exogenous interval reporting for income). I use the 1996 wave of the survey and restrict attention to women with no more than \$15,000 of yearly income who report using prescription medications. This results in a data set with 636 observations. Of these observations, 54 have prescription expenditures reported as an interval of nonzero width with finite endpoints, and an additional 7 have no information on prescription expenditures.

To describe the setup formally, let $X_i$ and $W_i^*$ be income and prescription drug expenditures for the $i$th observation. We observe $(X_i, W_i^L, W_i^H)$, where $[W_i^L, W_i^H]$ is an interval that contains $W_i^*$. For observations where no interval is reported for prescription drug spending, I set $W_i^L = 0$ and $W_i^H = \infty$. I estimate an interval median regression model where the median $q_{1/2}(W_i^*|X_i)$ of $W_i^*$ given $X_i$ is assumed to follow a linear regression model $q_{1/2}(W_i^*|X_i) = \theta_1 + \theta_2 X_i$. This leads to the conditional moment inequalities $E(m(W_i, \theta)|X_i) \geq 0$ almost surely, where $m(W_i, \theta) = (I(\theta_1 + \theta_2 X_i \leq W_i^H) - 1/2, 1/2 - I(\theta_1 + \theta_2 X_i \leq W_i^L))$ and $W_i = (X_i, W_i^L, W_i^H)$.

Figure 7 shows the data graphically. The horizontal axis measures income, while the vertical axis measures out of pocket prescription drug expenditures. Observations for which prescription expenditures are reported as a point value are plotted as points. For observations where a nontrivial interval is reported, a plus symbol marks the upper endpoint, and an x marks the lower endpoint. For observations where no information on prescription expenditures is obtained in the survey, a circle is placed on the $x$ axis at the value of income reported for that observation. In order to show in detail the ranges of spending that contain most of the observations, the vertical axis is truncated at \$15,000, leading to 5 observations not being shown (although these observations are used in forming the confidence regions reported below).

I form 95% confidence intervals by inverting level .05 tests using the KS statistics described in this paper with critical values calculated using the conservative rate of convergence $n^{1/2}$, and rates of convergence estimated using the methods described in Section 6.1. For the function $S$, I set $S(t) = \max_k |t_k \wedge 0|$. The rest of the implementation details are the same as for the monte carlos in Section 8.

For comparison, I also compute point estimates and confidence regions using the least absolute deviations (LAD) estimator (Koenker and Bassett, 1978) for the median regression model with only the observations for which a point value for spending was reported. These are valid under the additional assumption that the decision to report an interval or missing value is independent of spending conditional on income. The confidence regions use Wald tests based on the asymptotic variance estimates computed by Stata. These asymptotic variance estimates are based on formulas in Koenker and Bassett (1982) and require additional assumptions on the data generating process, but I use these rather than more robust standard errors in order to provide a comparison to an alternative procedure using default options in a standard statistical package.

Figure 8 plots the outline of the 95% confidence region for $\theta$ using the pre-tests and rate of convergence estimates described above, while Figure 9 plots the outline of the 95% confidence region using the conservative approximation. Figure 10 plots the outline of the 95% confidence region from estimating a median regression model on the subset of the data with point values reported for spending. Table 5 reports the corresponding confidence intervals for the components of $\theta$. For the confidence regions based on KS tests, I use the projections of the confidence region for $\theta$ onto each component. For the confidence regions based on median regression with point observations, the 95% confidence regions use the limiting normal approximation for each component of $\theta$ separately.

The results show a sizeable increase in statistical power from using the estimated rates of convergence. With the conservative tests, the 95% confidence region estimates that a \$1,000 increase in income is associated with at least a \$3 increase in out of pocket prescription spending at the median. With the tests that use the estimated rates of convergence, the 95% confidence region bounds the increase in out of pocket prescription spending associated with a \$1,000 increase in income from below by \$11.30.

The 95% confidence region based on median regression using observations reported as points overlaps with both moment inequality based confidence regions, but gives a different picture of which parameter values can be ruled out by the data. The upper bound for the increase in spending associated with a $1,000 increase in income is $24.40 using LAD, compared to $37.20 and $34.70 using KS statistics with all observations and the conservative and estimated rates respectively. The corresponding lower bound is $10 using LAD with point observations, substantially larger than the lower bound of $3 using the conservative procedure, but actually smaller than the $11.30 lower bound under the estimated rate. Thus, while the interval reporting at random assumption for the dependent variable allows one to tighten the upper bound for the slope parameter, a lower bound close to the lower bound of the LAD confidence interval can be obtained using the new asymptotic approximations developed in this paper.

Note also that these tests could, but do not, provide evidence against the assumptions required for LAD on the point reported values. If the LAD 95% confidence region did not overlap with one of the moment inequality 95% confidence regions, there would be no parameter value consistent with this assumption at the .1 level (for any parameter value, we can reject the joint null of both models holding using Bonferroni's inequality and the results of the .05 level tests). This type of test will not necessarily have power if the interval reporting at random assumption for the dependent variable does not hold, so it should not be taken as evidence that the more robust interval regression assumptions can be replaced with LAD methods.

# 10 Discussion

Under some smoothness conditions, the asymptotic approximations derived in Sections 3 and 5 can be combined with the methods in Sections 4 and 6 to form tests that are asymptotically exact on portions of the boundary of the identified set where the $\sqrt{n}$ approximation only allows for conservative inference. Since these methods require assumptions on the conditional mean that are not needed for conservative inference using the $\sqrt{n}$ approximation, the decision of which method to use involves a tradeoff between power and robustness. The results in Section 7 quantify these tradeoffs. While approximations to the distribution of a KS statistic based on the asymptotic distribution in Section 3 and the tests in Sections 4 and 6 may not be robust to certain types of nonsmooth conditional means, when they are valid, they can detect parameters in a $n^{-2/(d_X+4)}$ region of the identified set, while the $\sqrt{n}$ approximation can only detect parameters in a $n^{-1/(d_X+2)}$ region of the identified set. It should be noted that, even if the pre-tests in Section 6 find a rate of convergence that is too fast, Lemma 6 in the Appendix shows that the rate of convergence will typically be within $\log n$ of $1/n$ for testing $\theta$ on the interior of the identified set, so the resulting confidence region, while failing to contain values of $\theta$ near the boundary of the identified set with high probability, will not be too much smaller than the true identified set.

The results in this paper also shed light on the tradeoff between the KS statistics based on integrating conditional moments to get unconditional moments considered in this paper and other methods for inference with conditional moment inequalities, such as those based on kernel or series estimation (Chernozhukov, Lee, and Rosen, 2009; Ponomareva, 2010) or increasing numbers of unconditional moments (Menzel, 2008). With the bandwidth chosen to decrease at the correct rate, kernel methods based on a supremum statistic will give close to the same $n^{-2/(d_X+4)}$ rate (up to a power of $\log n$) for detecting the local alternatives considered in this paper. With enough derivatives imposed on the conditional mean, higher order kernels or series methods could be used to get even more power. However, kernel based methods will perform worse with suboptimal bandwidth choices, or against local alternatives in which the conditional moment inequality fails to hold on a larger set. The $n^{-2/(d_X+4)}$ rate for detecting local alternatives can also be achieved within a $\log n$ term using the increasing truncation point variance weighting proposed in Armstrong (2011). Unlike the tests proposed in this paper, those methods are robust to nonsmooth conditional means. These tests also have the advantage of adapting to different shapes of the conditional mean without estimating the optimal bandwidth, as would be necessary with kernel estimates, or estimating the rate of convergence of a test statistic, as required by the tests in this paper. However, they have less power by a $\log n$ term when applied to this setting, and require choosing a conservative critical value, which decreases the power further (but not the rate at which local alternatives can converge to the identified set and still be detected).

While the results in this paper and Armstrong (2011) characterize how moment selection and weighting functions affect relative efficiency in this setting, the choice of test statistic (supremum norm, as considered here, or $L_p$ norm, as with Cramer-von Mises statistics) and instrument functions are also of interest. While the results in this paper and in Armstrong (2011) give some insight into these problems (for example, it is clear from the arguments in these papers that Cramer-von Mises style statistics will have less power in this setting unless new asymptotic distribution results or moment selection procedures are used) more complete answers to these questions are topics of ongoing research.

It is also interesting to compare the nonsimilarity problem with the statistics in this paper to nonsimilarity problems encountered with kernel based methods. The rate of convergence of supremum statistics based on kernel estimates of the conditional moments also depends on the contact set, but to a lesser extent. Ponomareva (2010) shows that the rate of convergence of these statistics differs by a factor of $\log n$ depending on the contact set. Arguing as in Section 6 of Armstrong (2011), this would lead to an increase in the rate at which local alternatives can approach the identified set and still be detected by a factor of $\log n$. In contrast, the polynomial difference in the rates of convergence of the KS statistics based on integrated moments considered in the present paper leads to increases in local power by factors of $n$ rather than $\log n$. Thus, the gains in terms of power from using exact approximations are much larger in this context.

In addition to these immediate practical applications, the results in this paper are also of independent interest in their relation to broader questions in the literatures on moment inequalities and nonparametric estimation. In testing multiple moment inequalities, the asymptotic distribution of test statistics typically only depends on inequalities that bind as equalities. Since the non binding moments do affect the finite sample distribution of the test statistic, this means that asymptotic distributions may provide poor approximations to finite sample distributions. The existing literature on moment inequalities has taken several approaches to this issue. One is to use conservative approximations using "least favorable" asymptotic distributions in which all moment inequalities bind. Another approach is to design tests that are robust to sequences where the data generating process or test statistic changes as the number of observations increases. Menzel (2008) considers asymptotic approximations in which the number of moment inequalities used for a test statistic increases with the number of observations. Andrews and Shi (2009) show that the tests they consider using test statistics similar to the ones in this paper, but using a (possibly degenerate) $\sqrt{n}$ asymptotic distribution, have the correct size asymptotically when data generating processes change with the sample size within certain classes of data generating processes. Since these classes of data generating processes include sequences where some moment inequalities are slack, but close to binding, this suggests that the methods they propose will not suffer from problems with non binding inequalities affecting the finite sample distribution.

In contrast, the asymptotic distributions presented in Sections 3 and 5 of the present paper are, to my knowledge, the first known case of the asymptotic distribution of a test statistic that uses a fixed (although, in this case, infinite) set of moment inequalities depending on moment inequalities that do not bind. These results show that, under the conditions in this paper, the "moment selection" problem takes the form of a balancing of expected value and variance of moments that are close to binding. This leads to ideas typically associated with kernel smoothing and nonstandard M-estimation applying to test statistics for moment inequalities. As with the objective functions for nonstandard M-estimation considered by Kim and Pollard (1990), the asymptotic distribution of the KS statistic is the limit of local processes under a scaling that balances a drift term and a variance term. This balancing of drift and variance terms mirrors the equating of bias and variance terms in choosing the optimal bandwidth for nonparametric kernel estimation (see, for example, Pagan and Ullah, 1999). This is especially interesting since one of the appealing features of KS style statistics in this setting is that they get rid of the need for bandwidth parameters. In the settings I consider, the choice of "bandwidth" is made automatically by the balancing of the drift and variance terms, which determines the scale of the moments that matter asymptotically. However, this shows up in the rate of convergence, so that tests to determine which "bandwidth" was chosen are still needed for exact inference. Thus, in a sense, the bandwidth selection problem shows up in the moment selection problem through the rate of convergence.

In another paper (Armstrong, 2011), I show that KS statistics similar to the ones in the present paper can be made to choose the moments that correspond to the optimal bandwidth by using a variance weighting with an increasing sequence of truncation points. This helps alleviate the problem with different rates of convergence of the KS statistic along the boundary of the identified set, but loses a $\log n$ term relative to the tests based on unweighted KS statistics (or KS statistics with bounded weights) and asymptotic approximations based on the exact rate of convergence. Thus, moment selection (in the form of testing for rates of convergence) and variance weighting play similar roles in this framework. Even without the variance weighting of Armstrong (2011), the statistics in this paper find the moments that lead to the most local power. Estimating the rate of convergence of the test statistic is only needed to find the order of magnitude (under the null) of the moments that were found.

The results in this paper are pointwise in the underlying distribution $P$. Since the procedures proposed in this paper involve pre tests, it is natural to ask for which classes of underlying distributions these tests are uniformly valid. Since uniformity in the underlying distribution is implicit in the bounds used in many of the arguments used to derive these asymptotic distributions, it seems likely that these tests could be shown to enjoy uniformity in classes of distributions with uniform bounds on the constants governing the smoothness conditions needed for the pointwise results. While this would be an interesting extension of the results in the paper, uniformity in the underlying distribution is perhaps less interesting than in other settings because many of the tradeoffs between the approach in the present paper and more conservative approaches are already clear from the pointwise results. Smoothness conditions not needed for the conservative approach to control the size uniformly in the underlying distribution are needed even for the pointwise results derived here. Thus, it is clear from the pointwise results that the power improvement achieved by the tests in this paper comes at a cost of robustness to smoothness conditions.

Many of the results in this paper assume that the conditional mean $\bar{m}(\theta, x)$ is minimized only on a finite set. For the case where $d_X = 1$, this is implied by smoothness conditions on the conditional mean (or, when it does not hold, the results in this paper bound the rate of convergence so that the tests based on estimated rates are still valid). In higher dimensions, the case where the contact set has infinitely many points but is of a dimension less than $d_X$ is likely to be more difficult, but similar ideas will apply. The results in this paper could also be extended to the case where the $\bar{m}(\theta, x)$ only approaches 0 near the (possibly infinite) boundary of the support of $x$. These cases are often relevant in performing inference on bounds on treatment effects such as those considered by Manski (1990). In the one dimensional case, $X_i$ can be transformed into a uniform random variable so that the conditions on the density of the conditioning variable used in this paper will apply (once the density is positive and well behaved on its support, the assumption that the contact point is on the interior of the support is easy to relax). If the density and conditional mean approach zero at polynomial rates, the transformed model will fit into a slight extension of Theorem 6 for some $\gamma$ that depends on these rates. These transformations are used in a slightly different setting in Armstrong (2011).

# 11 Conclusion

This paper derives the asymptotic distribution of a class of Kolmogorov-Smirnov style test statistics for conditional moment inequality models under a general set of conditions. I show how to use these results to form valid tests that are more powerful than existing approaches based on this statistic. Local power results for the new tests and existing tests are derived, which quantify this power improvement. While the increase in power comes at a cost of robustness to smoothness conditions, a complementary paper (Armstrong, 2011) proposes methods for inference that achieve almost the same power improvement while still being robust to failure of smoothness conditions.

In addition to their immediate practical application to asymptotically exact inference, the results in this paper add to our understanding of how familiar issues in the literatures on moment inequalities and nonparametric estimation, such as moment selection and the curse of dimensionality, manifest themselves in the use of one sided KS statistics for conditional moment inequalities. Under the conditions in this paper, the asymptotic distribution of the KS statistic depends on nonbinding moments, which are determined through a balancing of a bias term and a variance term in a way that is similar to the objective functions for the point estimators considered by Kim and Pollard (1990). The dimension of the conditioning variables and the smoothness of the conditional mean determine which moments matter asymptotically and which types of local alternatives the KS statistic can detect.

# Appendix

This appendix contains proofs of the theorems in this paper. The proofs are organized into subsections according to the section containing the theorem in the body of the paper. In cases where a result follows immediately from other theorems or arguments in the body of the paper, I omit a separate proof. Statements involving convergence in distribution in which random elements in the converging sequence are not measurable with respect to the relevant Borel sigma algebra are in the sense of outer weak convergence (see van der Vaart and Wellner, 1996). For notational convenience, I use $d = d_X$ throughout this appendix.

## Asymptotic Distribution of the KS Statistic

In this subsection of the appendix, I prove Theorem 1. For notational convenience, let $Y_i = m(W_i, \theta)$ and $Y_{i,J(m)} = m_{J(m)}(W_i, \theta)$ and let $d = d_X$ and $k = d_Y$ throughout this subsection.

The asymptotic distribution comes from the behavior of the objective function $E_n Y_{i,j} I(s < X_i < s + t)$ for $(s, t)$ near $x_m$ such that $j \in J(m)$. The bulk of the proof involves showing that the objective function doesn't matter for $(s, t)$ outside of neighborhoods of $x_m$ with $j \in J(m)$ where these neighborhoods shrink at a fast enough rate. First, I derive the limiting distribution over such shrinking neighborhoods and the rate at which they shrink.

**Theorem 15.** *Let $h_n = n^{-\alpha}$ for some $0 < \alpha < 1/d$. Let*

$$\mathbb{G}_{n,x_m}(s,t) = \frac{\sqrt{n}}{h_n^{d/2}}(E_n - E)Y_{i,J(m)} I(h_n s < X_i - x_m < h_n(s+t))$$

*and let $g_{n,x_m}(s,t)$ have $j$th element*

$$g_{n,x_m,j}(s,t) = \frac{1}{h_n^{d+2}} EY_{i,j} I(h_n s < X_i - x_m < h_n(s+t))$$

*if $j \in J(m)$ and zero otherwise. Then, for any finite $M$, $(\mathbb{G}_{n,x_1}(s,t), \ldots, \mathbb{G}_{n,x_\ell}(s,t)) \xrightarrow{d} (\mathbb{G}_{P,x_1}(s,t), \ldots, \mathbb{G}_{P,x_\ell}(s,t))$ taken as random processes on $\|(s,t)\| \leq M$ with the supremum norm and $g_{n,x_m}(s,t) \to g_{P,x_m}(s,t)$ uniformly in $\|(s,t)\| \leq M$ where $\mathbb{G}_{P,x_m}(s,t)$ and $g_{P,x_m}(s,t)$ are defined as in Theorem 1 for $m$ from $1$ to $\ell$.*

*Proof.* The convergence in distribution in the first statement follows from verifying the conditions of Theorem 2.11.22 in van der Vaart and Wellner (1996). To derive the covariance kernel, note that

$$\begin{aligned}
&cov(\mathbb{G}_{n,x_m}(s,t), \mathbb{G}_{n,x_m}(s',t')) \\
&= h_n^{-d} EY_{i,J(m)} Y'_{i,J(m)} I\left\{h_n(s \vee s') < X - x_m < h_n\left[(s+t) \wedge (s'+t')\right]\right\} \\
&- h_n^{-d}\left\{EY_{i,J(m)} I\left[h_n s < X - x_m < h_n(s+t)\right]\right\}\left\{EY'_{i,J(m)} I\left[h_n s' < X - x_m < h_n(s'+t')\right]\right\}.
\end{aligned}$$

The second term goes to zero as $n \to \infty$. The first is equal to the claimed covariance kernel plus the error term

$$h_n^{-d} \int_{h_n(s \vee s') < x - x_m < h_n[(s+t) \wedge (s'+t')]} \left[E(Y_{i,J(m)} Y'_{i,J(m)} | X = x) f_X(x) - E(Y_{i,J(m)} Y'_{i,J(m)} | X = x_m) f_X(x_m)\right] dx,$$

which is bounded by

$$\begin{aligned}
&\left\{\max_{\|x - x_m\| \leq 2h_n M} \left[E(Y_{i,J(m)} Y'_{i,J(m)} | X = x) f_X(x) - E(Y_{i,J(m)} Y'_{i,J(m)} | X = x_m) f_X(x_m)\right]\right\} \\
&\times h_n^{-d} \int_{h_n(s \vee s') < x - x_m < h_n[(s+t) \wedge (s'+t')]} dx \\
&= \left\{\max_{\|x - x_m\| \leq 2h_n M} \left[E(Y_{i,J(m)} Y'_{i,J(m)} | X = x) f_X(x) - E(Y_{i,J(m)} Y'_{i,J(m)} | X = x_m) f_X(x_m)\right]\right\} \\
&\times \int_{(s \vee s') < x - x_m < (s+t) \wedge (s'+t')} dx.
\end{aligned}$$

This goes to zero as $n \to \infty$ by continuity of $E(Y_{i,J(m)} Y'_{i,J(m)} | X = x)$ and $f_X(x)$. For $m \neq r$

and $\|(s,t)\| \leq M$, $\|(s',t')\| \leq M$, $cov(\mathbb{G}_{n,x_m}(s,t), \mathbb{G}_{n,x_r}(s',t'))$ is eventually equal to

$$-h_n^{-d}\left\{EY_{i,J(m)}I\left[h_n s < X - x_m < h_n(s+t)\right]\right\}\left\{EY'_{i,J(r)}I\left[h_n s' < X - x_r < h_n(s'+t')\right]\right\},$$

which goes to zero, so the processes for different elements of $\mathcal{X}_0$ are independent as claimed.

For the claim regarding $g_{n,x_m}(s,t)$, first note that the assumptions imply that, for $j \in J(m)$, the first derivative of $x \mapsto E(Y_{i,j}|X=x)$ at $x = x_m$ is 0, and that this function has a second order Taylor expansion:

$$E(Y_{i,j}|X=x) = \frac{1}{2}(x-x_m)'V_j(x_m)(x-x_m) + R_n(x)$$

where

$$R_n(x) = \frac{1}{2}(x-x_m)'V_j(x^*(x))(x-x_m) - \frac{1}{2}(x-x_m)'V_j(x_m)(x-x_m)$$

and $V_j(x^*)$ is the second derivative matrix evaluated at some $x^*(x)$ between $x_m$ and $x$.

We have

$$\begin{aligned} g_{n,x_m,j}(s,t) = &\frac{1}{2h_n^{d+2}}\int_{h_n s<x-x_m<h_n(s+t)} (x-x_m)'V_j(x_m)(x-x_m)f_X(x_m)\,dx \\ &+ \frac{1}{2h_n^{d+2}}\int_{h_n s<x-x_m<h_n(s+t)} (x-x_m)'V_j(x_m)(x-x_m)[f_X(x)-f_X(x_m)]\,dx \\ &+ \frac{1}{h_n^{d+2}}\int_{h_n s<x-x_m<h_n(s+t)} R_n(x)f_X(x)\,dx. \end{aligned}$$

The first term is equal to $g_{P,x_m,j}(s,t)$ by a change of variable $x$ to $h_n x + x_m$ in the integral. The second term is bounded by $g_{P,x_m,j}(s,t)\sup_{\|x-x_m\|\leq 2h_n M}[f_X(x)-f_X(x_m)]/f_X(x_m)$, which goes to zero uniformly in $\|(s,t)\| \leq M$ by continuity of $f_X$. The third term is equal to (using the same change of variables)

$$\frac{1}{2}\int_{s<x<s+t}[x'V_j(x^*(h_n x + x_m))x - x'V_j(x_m)x]f_X(h_n x + x_m)\,dx.$$

This is bounded by a constant times $\sup_{\|x\|\leq M}|x'V_j(x^*(h_n x+x_m))x - x'V_j(x_m)x|$, which goes to zero as $n \to \infty$ by continuity of the second derivatives. □

Thus, if we let $h_n$ be such that $\sqrt{n}/h_n^{d/2} = 1/h_n^{d+2} \iff h_n = n^{-1/(d+4)}$ and scale up by

$\sqrt{n}/h_n^{d/2} = 1/h_n^{d+2} = n^{(d+2)/(d+4)}$, we will have

$$\begin{aligned}
&n^{(d+2)/(d+4)}(E_n Y_{i,J(1)} I(h_n s < X - x_1 < h_n(s+t)), \ldots, E_n Y_{i,J(\ell)} I(h_n s < X - x_\ell < h_n(s+t))\\
&= (\mathbb{G}_{n,x_1}(s,t) + g_{n,x_1}(s,t), \ldots, \mathbb{G}_{n,x_\ell}(s,t) + g_{n,x_\ell}(s,t))\\
&\xrightarrow{d} (\mathbb{G}_{P,x_1}(s,t) + g_{P,x_1}(s,t), \ldots, \mathbb{G}_{P,x_m}(s,t) + g_{P,x_m}(s,t))
\end{aligned}$$

taken as stochastic processes in $\{\|(s,t)\| \leq M\}$ with the supremum norm. From now on, let $h_n = n^{-1/(d+4)}$ so that this will hold.

We would like to show that the infimum of these stochastic processes over all of $\mathbb{R}^{2d}$ converges to the infimum of the limiting process over all of $\mathbb{R}^{2d}$, but this does not follow immediately since we only have uniform convergence on compact sets. Another way of thinking about this problem is that convergence in distribution in $\{\|(s,t)\| \leq M\}$ with the supremum norm for any $M$ implies convergence in distribution in $\mathbb{R}^{2d}$ with the topology of uniform convergence on compact sets (see Kim and Pollard, 1990), but the infimum over all of $\mathbb{R}^{2d}$ is not a continuous mapping on this space since uniform convergence on all compact sets does not imply convergence of the infimum over all of $\mathbb{R}^{2d}$. To get the desired result, the following lemma will be useful. The idea is to show that values of $(s,t)$ far away from zero won't matter for the limiting distribution, and then use convergence for fixed compact sets.

**Lemma 2.** *Let $\mathbb{H}_n$ and $\mathbb{H}_P$ be random functions from $\mathbb{R}^{k_1}$ to $\mathbb{R}^{k_2}$ such that, (i) for all $M$, $\mathbb{H}_n \xrightarrow{d} \mathbb{H}_P$ when $\mathbb{H}_n$ and $\mathbb{H}_P$ are taken as random processes on $\{t \in \mathbb{R}^{k_1} | \|t\| \leq M\}$ with the supremum norm, (ii) for all $r < 0$, $\varepsilon > 0$, there exists an $M$ such that $P\left(\inf_{\|t\|>M} \mathbb{H}_{P,j}(t) \leq r \text{ some } j\right) < \varepsilon$ and an $N$ such that $P\left(\inf_{\|t\|>M} \mathbb{H}_{n,j}(t) \leq r \text{ some } j\right) < \varepsilon$ for all $n \geq N$ and (iii) $\inf_t \mathbb{H}_{n,j}(t) \leq 0$ and $\inf_t \mathbb{H}_{P,j}(t) \leq 0$ with probability one. Then $\inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_n(t) \xrightarrow{d} \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t)$.*

*Proof.* First, by the Cramer-Wold device, it suffices to show that, for all $w \in \mathbb{R}^{k_2}$, $w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_n(t) \xrightarrow{d} w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t)$. For this, it suffices to show that for all $r \in \mathbb{R}$, $\liminf_n P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_n(t) < r\right) \geq P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t) < r\right)$ and $\limsup_n P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_n(t) \leq r\right) \leq P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t) \leq r\right)$ since, arguing along the lines of the Portmanteau Lemma, when $r$ is a continuity point of the limiting distribution, we will have

$$\begin{aligned}
&P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t) \leq r\right) = P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t) < r\right) \leq \liminf_n P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_n(t) < r\right)\\
&\leq \liminf_n P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_n(t) \leq r\right) \leq \limsup_n P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_n(t) \leq r\right) \leq P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t) \leq r\right).
\end{aligned}$$

Given $\varepsilon > 0$, let $M$ and $N$ be as in the assumptions of the lemma, but with $r$ replaced by $r/(k_2 \max_i |w_i|)$. Then

$$P\left(w' \inf_{\|t\|\geq M} \mathbb{H}_P(t) < r\right) \leq P\left((k_2 \max_i |w_i|) \inf_{\|t\|\geq M} \mathbb{H}_{P,j}(t) < r \text{ some } j\right) \leq \varepsilon$$

so that $P\left(w' \inf_{\|t\|\leq M} \mathbb{H}_P(t) < r\right) + \varepsilon \geq P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t) < r\right)$ and, for $n \geq N$,

$$P\left(w' \inf_{\|t\|\geq M} \mathbb{H}_n(t) \leq r\right) \leq P\left((k_2 \max_i |w_i|) \inf_{\|t\|\geq M} \mathbb{H}_{n,j}(t) \leq r \text{ some } j\right) \leq \varepsilon$$

so that $P\left(w' \inf_{\|t\|\leq M} \mathbb{H}_n(t) \leq r\right) + \varepsilon \geq P\left(w' \inf_{t\in\mathbb{R}} \mathbb{H}_n(t) \leq r\right)$. Thus, by convergence in distribution of the infima over $\|t\| \leq M$,

$$\liminf_n P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_n(t) < r\right) \geq \liminf_n P\left(w' \inf_{\|t\|\leq M} \mathbb{H}_n(t) < r\right) \geq P\left(w' \inf_{\|t\|\leq M} \mathbb{H}_P(t) < r\right)$$
$$\geq P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t) < r\right) - \varepsilon$$

and

$$\limsup_n P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_n(t) \leq r\right) \leq \limsup_n P\left(w' \inf_{\|t\|\leq M} \mathbb{H}_n(t) \leq r\right) + \varepsilon$$
$$\leq P\left(w' \inf_{\|t\|\leq M} \mathbb{H}_P(t) \leq r\right) + \varepsilon \leq P\left(w' \inf_{t\in\mathbb{R}^{k_1}} \mathbb{H}_P(t) \leq r\right) + \varepsilon.$$

Since $\varepsilon$ was arbitrary, this gives the desired result.

□

Technically, this lemma does not apply to

$$(\mathbb{G}_{n,x_1}(s,t) + g_{n,x_1}(s,t), \ldots, \mathbb{G}_{n,x_\ell}(s,t) + g_{n,x_\ell}(s,t))$$

since, for $m \neq r$, $\mathbb{G}_{n,x_m}(s,t) + g_{n,x_m}(s,t)$ evaluated at some increasing values of $(s,t)$ may actually be equal to $\mathbb{G}_{n,x_r}(s',t') + g_{n,x_r}(s',t')$ for some small values of $(s',t')$, since, once the local indices are large enough, the original indices overlap. Instead, noting that, for any

$\eta > 0$,

$$\begin{aligned}
&n^{(d+2)/(d+4)} \inf_{s,t} E_n Y_i I(s < X_i < s+t) \\
&= \Bigg( \min_{m \text{ s.t. } 1\in J(m)} \inf_{\|(s,t)\|\le \eta/h_n} \mathbb{G}_{n,x_m,1}(s,t) + g_{n,x_m,1}(s,t)), \ldots, \\
&\min_{m \text{ s.t. } k\in J(m)} \inf_{\|(s,t)\|\le \eta/h_n} \mathbb{G}_{n,x_m,k}(s,t) + g_{n,x_m,k}(s,t) \Bigg) \\
&\wedge \Bigg( n^{(d+2)/(d+4)} \inf_{\|(s-x_m,t)\|>\eta \text{ all } m \text{ s.t. } 1\in J(m)} E_n Y_{i,1} I(s < X_i < s+t), \ldots, \\
&n^{(d+2)/(d+4)} \inf_{\|(s-x_m,t)\|>\eta \text{ all } m \text{ s.t. } k\in J(m)} E_n Y_{i,k} I(s < X_i < s+t) \Bigg) \\
&\equiv Z_{n,1} \wedge Z_{n,2},
\end{aligned}$$

I show that, for some $\eta > 0$, $Z_{n,2} \overset{p}{\to} 0$ using a separate argument, and use Lemma 2 to show that, for the same $\eta$,

$$\begin{aligned}
&(\inf_{s,t}[\mathbb{G}_{n,x_1}(s,t) + g_{n,x_1}(s,t)]I(\|(s,t)\| \le \eta/h_n), \ldots, \inf_{s,t}[\mathbb{G}_{n,x_\ell}(s,t) + g_{n,x_\ell}(s,t)]I(\|(s,t)\| \le \eta/h_n)) \\
&\overset{d}{\to} (\inf_{s,t} \mathbb{G}_{P,x_1}(s,t) + g_{P,x_1}(s,t), \ldots, \inf_{s,t} \mathbb{G}_{P,x_\ell}(s,t) + g_{P,x_\ell}(s,t)),
\end{aligned}$$

from which it follows that $Z_{n,1} \overset{d}{\to} Z$ for $Z$ defined as in Theorem 1 by the continuous mapping theorem.

Part (i) of Lemma 2 follows from Theorem 15 (the $I(\|(s,t)\| \le \eta/h_n)$ term does not change this, since it is equal to one for $\|(s,t)\| \le M$ eventually). Part (iii) follows since the processes involved are equal to zero when $t = 0$. To verify part (ii), first note that it suffices to verify part (ii) of the lemma for $\mathbb{G}_{n,x_m,j}(s,t) + g_{n,x_m,j}(s,t)$ and $\mathbb{G}_{P,x_m,j}(s,t) + g_{P,x_m,j}(s,t)$ for each $m$ and $j$ individually. Part (ii) of the lemma holds trivially for $m$ and $j$ such that $j \notin J(m)$, so we need to verify this part of the lemma for $m$ and $j$ such that $j \in J(m)$.

The next two lemmas provide bounds that will be used to verify condition (ii) of Lemma 2 for $\mathbb{G}_{n,x_m,j}(s,t)+g_{n,x_m,j}(s,t)$ and $\mathbb{G}_{P,x_m,j}(s,t)+g_{P,x_m,j}(s,t)$ for $m$ and $j$ with $j \in J(m)$. To do this, the bounds in the lemmas are applied to sets of $(s,t)$ with $\|(s,t)\|$ increasing. The idea is similar to the "peeling" argument of, for example, Kim and Pollard (1990), but different arguments are required to deal with values of $(s,t)$ for which, even though $\|s\|$ is large, $\prod_i t_i$ is small so that the objective function on average uses only a few observations, which may happen to be negative. To get bounds on the suprema of the limiting and finite sample processes where $t$ may be small relative to $s$, the next two lemmas bound the supremum by a maximum over $s$ in a finite grid of suprema over $t$ with $s$ fixed, and then use exponential bounds on suprema of the processes with fixed $s$.

**Lemma 3.** *Fix $m$ and $j$ with $j \in J(m)$. For some $C > 0$ that depends only on $d$, $f_X(x_m)$ and $E(Y_{i,j}^2|X = x_m)$, we have, for any $B \geq 1$, $\varepsilon > 0$, $w > 0$,*

$$P\left(\sup_{\|(s,t)\| \leq B, \prod_i t_i \leq \varepsilon} |\mathbb{G}_{P,x_m,j}(s,t)| \geq w\right) \leq 2\left\{3B[B^d/(\varepsilon \wedge 1)] + 2\right\}^{2d} \exp\left(-C\frac{w^2}{\varepsilon}\right)$$

*for $\frac{w^2}{\varepsilon}$ greater than some constant that depends only on $d$, $f_X(x_m)$ and $E(Y_{i,j}^2|X = x_m)$.*

*Proof.* Let $\mathbb{G}(s,t) = \mathbb{G}_{P,x_m,j}(s,t)$. We have, for any $s_0 \leq s \leq s+t \leq s_0+t_0$,

$$\begin{aligned}
&\mathbb{G}(s,t) = \mathbb{G}(s_0, t+s-s_0) \\
&+ \sum_{1\leq j\leq d} (-1)^j \sum_{1\leq i_1<i_2<\ldots<i_j\leq d} \\
&\mathbb{G}(s_0, (t_1+s_1-s_{0,1}, \ldots, t_{i_1-1}+s_{i_1-1}-s_{0,i_1-1}, s_{i_1}-s_{0,i_1}, t_{i_1+1}+s_{i_1+1}-s_{0,i_1+1}, \\
&\ldots, t_{i_j-1}+s_{i_j-1}-s_{0,i_j-1}, s_{i_j}-s_{0,i_j}, t_{i_j+1}+s_{i_j+1}-s_{0,i_j+1}, \ldots, t_d+s_d-s_{0,d})).
\end{aligned}$$

Thus, since there are $2^d$ terms in the above display, each with absolute value bounded by $\sup_{t\leq t_0} |\mathbb{G}(s_0,t)|$,

$$\sup_{s_0\leq s\leq s+t\leq s_0+t_0} |\mathbb{G}(s,t)| \leq 2^d \sup_{t\leq t_0} |\mathbb{G}(s_0,t)| \overset{d}{=} 2^d \sup_{t\leq t_0} |\mathbb{G}(0,t)|.$$

Let $A$ be a grid of meshwidth $(\varepsilon\wedge 1)/B^d$ covering $[-B, 2B]^d$. For any $(s,t)$ with $\|(s,t)\| \leq B$ and $\prod_i t_i \leq \varepsilon$, there are $s_0$ and $t_0$ with $s_0, s_0+t_0 \in A$ such that $s_0 \leq s \leq s+t \leq s_0+t_0$, and $\prod_i t_{0,i} \leq \prod_i (t_i + (\varepsilon \wedge 1)/B^d) = \sum_{j=0}^d [(\varepsilon \wedge 1)/B^d]^j \sum_{I\in\{1,\ldots,d\},|I|=d-j} \prod_{i\in I} t_i \leq \prod_i t_i + \sum_{j=1}^d [(\varepsilon \wedge 1)/B^d]^j \binom{d}{d-j} B^{d-j} \leq \varepsilon + \varepsilon \sum_{j=1}^d B^{-dj} \binom{d}{d-j} B^{d-j} \leq 2^d\varepsilon$. For this $s_0, t_0$, we will then have, by the above display, $|\mathbb{G}(s,t)| \leq 2^d \sup_{t\leq t_0} |\mathbb{G}(s_0,t)|$.

This gives

$$\sup_{\|(s,t)\|\leq B, \prod_i t_i\leq\varepsilon} |\mathbb{G}(s,t)| \leq 2^d \max_{s_0, s_0+t_0\in A, \prod_i t_{0,i}\leq 2^d\varepsilon} \sup_{t\leq t_0} |\mathbb{G}(s_0,t)|,$$

so that

$$
\begin{aligned}
&P\left(\sup_{\|(s,t)\|\le B, \prod_i t_i \le \varepsilon} |\mathbb{G}(s,t)| \ge w\right) \le |A|^2 \max_{s_0, s_0+t_0 \in A, \prod_i t_{0,i} \le 2^d \varepsilon} P\left(2^d \sup_{t \le t_0} |\mathbb{G}(s_0,t)| \ge w\right) \\
&= |A|^2 \max_{s_0, s_0+t_0 \in A, \prod_i t_{0,i} \le 2^d \varepsilon} P\left(2^d \sup_{t\le 1} \left(\prod_i t_{0,i}\right)^{1/2} |\mathbb{G}(0,t)| \ge w\right) \\
&\le |A|^2 P\left(\sup_{t \le 1} |\mathbb{G}(0,t)| \ge \frac{w}{2^d 2^{d/2} \varepsilon^{1/2}}\right).
\end{aligned}
$$

The result then follows using the fact that $|A| \le \left\{3B[B^d/(\varepsilon \wedge 1)] + 2\right\}^d$ and using Theorem 2.1 (p.43) in Adler (1990) to bound the probability in the last line of the display (the theorem in Adler (1990) shows that the probability in the above display is bounded by $2\exp(-K_1 w^2/\varepsilon + K_2 w/\varepsilon^{1/2} + K_3)$ for some constants $K_1$, $K_2$, and $K_3$ with $K_1 > 0$ that depend only on $d$, $f_X(x_m)$ and $E(Y_{i,j}^2|X = x_m)$ and this expression is less than $2\exp(-(K_1/2)w^2/\varepsilon)$ for $w^2/\varepsilon$ greater than some constant that depends only on $K_1$, $K_2$, and $K_3$).

□

**Lemma 4.** *Fix $m$ and $j$ with $j \in J(m)$. For some $C > 0$ that depends only on the distribution of $(X,Y)$ and some $\eta > 0$, we have, for any $1 \le B \le h_n^{-1}\eta$, $w > 0$ and $\varepsilon \ge n^{-4/(d+4)}(1 + \log n)^2$,*

$$
P\left(\sup_{\|(s,t)\|\le B, \prod_i t_i \le \varepsilon} |\mathbb{G}_{n,x_m,j}(s,t)| \ge w\right) \le 2\left\{3B[B^d/(\varepsilon \wedge 1)] + 2\right\}^{2d} \exp\left(-C\frac{w}{\varepsilon^{1/2}}\right).
$$

*Proof.* Let $\mathbb{G}_n(s,t) = \mathbb{G}_{n,x_m,j}(s,t)$. By the same argument as in the previous lemma with $\mathbb{G}$ replaced by $\mathbb{G}_n$, we have

$$
\sup_{s_0 \le s \le s+t \le s_0+t_0} |\mathbb{G}_n(s,t)| \le 2^d \sup_{t \le t_0} |\mathbb{G}_n(s_0,t)|.
$$

As in the previous lemma, let $A$ be a grid of meshwidth $(\varepsilon \wedge 1)/B^d$ covering $[-B, 2B]^d$. Arguing as in the previous lemma, we have, for any $(s,t)$ with $\|(s,t)\| \le B$ and $\prod_i t_i \le \varepsilon$, there exists some $s_0, t_0$ with $s_0, s_0 + t_0 \in A$ such that $\Pi_i t_{0,i} \le 2^d\varepsilon$ and $|\mathbb{G}_n(s,t)| \le$

$2^d \sup_{t\le t_0} |\mathbb{G}_n(s_0,t)|$. Thus,

$$\sup_{\|(s,t)\|\le B, \prod_i t_i \le \varepsilon} |\mathbb{G}_n(s,t)| \le 2^d \max_{s_0, s_0+t_0\in A, \prod_i t_{0,i}\le 2^d\varepsilon} \sup_{t\le t_0} |\mathbb{G}_n(s_0,t)|$$
$$= 2^d \max_{s_0, s_0+t_0\in A, \prod_i t_{0,i}\le 2^d\varepsilon} \sup_{t\le t_0} \frac{\sqrt{n}}{h_n^{d/2}} |(E_n - E) Y_{i,j} I(h_n s_0 \le X_i - x_m \le h_n(s_0+t))|.$$

This gives

$$P\left(\sup_{\|(s,t)\|\le B, \prod_i t_i \le 2^d\varepsilon} |\mathbb{G}_n(s,t)| \ge w\right)$$
$$\le |A|^2 \max_{s_0, s_0+t_0\in A, \prod_i t_{0,i}\le 2^d\varepsilon} P\left(2^d \sup_{t\le t_0} \frac{\sqrt{n}}{h_n^{d/2}} |(E_n - E) Y_{i,j} I(h_n s_0 \le X_i - x_m \le h_n(s_0+t))| \ge w\right).$$

We have, for some universal constant $K$ and all $n$ with $\varepsilon \ge n^{-4/(d+4)}(1+\log n)^2$, letting $\mathcal{F}_n = \{(x,y) \mapsto y_j I(h_n s_0 \le x - x_m \le h_n(s_0+t)) | t \le t_0\}$ and defining $\|\cdot\|_{P,\psi_1}$ to be the Orlicz norm defined on p.90 of van der Vaart and Wellner (1996) for $\psi_1(x) = \exp(x) - 1$,

$$\|2^d \sup_{f\in\mathcal{F}_n} |\sqrt{n}(E_n - E) f(X_i, Y_i)|\|_{P,\psi_1}$$
$$\le K\left[E \sup_{f\in\mathcal{F}_n} |\sqrt{n}(E_n - E) f(X_i, Y_i)| + n^{-1/2}(1+\log n) \| |Y_{i,j}| I(h_n s_0 \le X_i - x_m \le h_n(s_0+t_0)) \|_{P,\psi_1}\right]$$
$$\le K\left[J(1, \mathcal{F}_n, L^2) \left\{E[|Y_{i,j}| I(h_n s_0 < X_i - x_m < h_n(s_0+t_0))]^2\right\}^{1/2} + n^{-1/2}(1+\log n)\|Y\|_{P,\psi_1}\right]$$
$$\le K\left[J(1, \mathcal{F}_n, L^2) \overline{f}^{1/2} \overline{Y} h_n^{d/2} 2^{d/2} \varepsilon^{1/2} + n^{-1/2}(1+\log n)\|Y_{i,j}\|_{P,\psi_1}\right]$$
$$\le K\left[J(1, \mathcal{F}_n, L^2) \overline{f}^{1/2} \overline{Y} 2^{d/2} + \|Y_{i,j}\|_{P,\psi_1}\right] h_n^{d/2} \varepsilon^{1/2}.$$

The first inequality follows by Theorem 2.14.5 in van der Vaart and Wellner (1996). The second uses Theorem 2.14.1 in van der Vaart and Wellner (1996). The fourth inequality uses the fact that $h_n^{d/2}\varepsilon^{1/2} = n^{-d/[2(d+4)]}\varepsilon^{1/2} \ge n^{-1/2}(1+\log n)$ once $\varepsilon^{1/2} \ge n^{-1/2+d/[2(d+4)]}(1+\log n) = n^{-2/(d+4)}(1+\log n)$. Since each $\mathcal{F}_n$ is contained in the larger class $\mathcal{F}$ defined in the same way but replacing $s_0$ with $s$, and allowing $(s,t)$ to vary over all of $\mathbb{R}^{2d}$, we can replace $\mathcal{F}_n$ by $\mathcal{F}$ on the last line of this display. Since $J(1, \mathcal{F}, L^2)$ and $\|Y_{i,j}\|_{\psi_1}$ are finite ($\mathcal{F}$ is a VC class and $Y_{i,j}$ is bounded), the bound is equal to $C^{-1}\varepsilon^{1/2}h_n^{d/2}$ for a constant $C$ that depends only on the distribution of $(X_i, Y_i)$.

This bound along with Lemma 8.1 in Kosorok (2008) implies

$$\begin{aligned}
&P\left(2^d \sup_{t\le t_0} \frac{\sqrt{n}}{h_n^{d/2}} |(E_n - E)Y_{i,j} I(h_n s_0 \le X_i - x_m \le h_n(s_0+t))| \ge w\right)\\
&= P\left(2^d \sup_{f\in\mathcal{F}_n} |\sqrt{n}(E_n - E)f(X_i,Y_i)| \ge w h_n^{d/2}\right)\\
&\le 2\exp\left(-\frac{w h_n^{d/2}}{\|2^d \sup_{f\in\mathcal{F}_n} |\sqrt{n}(E_n-E)f(X_i,Y_i)|\|_{P,\psi_1}}\right)\\
&\le 2\exp\left(-\frac{w h_n^{d/2}}{C^{-1} h_n^{d/2} \varepsilon^{1/2}}\right) = 2\exp\left(-Cw/\varepsilon^{1/2}\right).
\end{aligned}$$

The result follows using this and the fact that $|A| \le \left\{3B[B^d/(\varepsilon\wedge 1)] + 2\right\}^d$. □

The following theorem verifies the part of condition (ii) of Lemma 2 concerning the limiting process $\mathbb{G}_{P,x_m,j}(s,t) + g_{P,x_m,j}(s,t)$.

**Theorem 16.** *Fix $m$ and $j$ with $j \in J(m)$. For any $r < 0$, $\varepsilon > 0$ there exists an $M$ such that*

$$P\left(\inf_{\|(s,t)\|>M} \mathbb{G}_{P,x_m,j}(s,t) + g_{P,x_m,j}(s,t) \le r\right) < \varepsilon.$$

*Proof.* Let $\mathbb{G}(s,t) = \mathbb{G}_{P,x_m,j}(s,t)$ and $g(s,t) = g_{P,x_m,j}(s,t)$. Let $S_k = \{k \le \|(s,t)\| \le k+1\}$ and let $S_k^L = S_k \cap \{\prod_i t_i \le (k+1)^{-\delta}\}$ for some fixed $\delta$. By Lemma 3,

$$\begin{aligned}
P\left(\inf_{S_k^L} \mathbb{G}(s,t) + g(s,t) \le r\right) &\le P\left(\sup_{S_k^L} |\mathbb{G}(s,t)| \ge |r|\right)\\
&\le 2\left\{3(k+1)[(k+1)^d/k^{-\delta}] + 2\right\}^{2d} \exp\left(-Cr^2(k+1)^\delta\right)
\end{aligned}$$

for $k$ large enough where $C$ depends only on $d$. This bound is summable over $k$.

For any $\alpha$ and $\beta$ with $\alpha < \beta$, let $S_k^{\alpha,\beta} = S_k \cap \{(k+1)^\alpha < \prod_i t_i \le (k+1)^\beta\}$. We have, for some $C_1 > 0$ that depends only on $d$ and $V_j(x_m)$, $g(s,t) \ge C_1\|(s,t)\|^2 \prod_i t_i$. (To see this, note that $g(s,t)$ is greater than or equal to a constant times $\int_{s_1}^{s_1+t_1} \cdots \int_{s_d}^{s_d+t_d} \|x\|^2 dx_d \cdots dx_1 = \left(\Pi_{i=1}^d t_i\right) \sum_{i=1}^d (s_i^2 + t_i^2/3 + s_i t_i)$, and the sum can be bounded below by a constant times $\|(s,t)\|^2$ by minimizing over $s_i$ for fixed $t_i$ using calculus. The claimed expression for the integral follows from evaluating the inner integral to get an expression involving the integral for $d-1$, and then using induction.) Using this and Lemma 3,

$$P\left(\inf_{S_k^{\alpha,\beta}} \mathbb{G}(s,t)+g(s,t)\le r\right)\le P\left(\sup_{S_k^{\alpha,\beta}}|\mathbb{G}(s,t)|\ge C_1k^{2+\alpha}\right)$$
$$\le 2\left\{3(k+1)[(k+1)^d/((k+1)^\beta\wedge 1)]+2\right\}^{2d}\exp\left(-CC_1^2\frac{k^{4+2\alpha}}{(k+1)^\beta}\right).$$

This is summable over $k$ if $4+2\alpha-\beta>0$.

Now, note that, since $\prod_i t_i\le (k+1)^d$ on $S_k$, we have, for any $-\delta<\alpha_1<\alpha_2<\ldots<\alpha_{\ell-1}<\alpha_\ell=d$, $S_k=S_k^L\cup S_k^{-\delta,\alpha_1}\cup S_k^{\alpha_1,\alpha_2}\cup\ldots\cup S_k^{\alpha_{\ell-1},\alpha_\ell}$. If we choose $\delta<3/2$ and $\alpha_i=i$ for $i\in\{1,\ldots,d\}$, the arguments above will show that the probability of the infimum being less than or equal to $r$ over $S_k^L$, $S_k^{-\delta,\alpha_1}$ and each $S_k^{\alpha_i,\alpha_{i+1}}$ is summable over $k$, so that $P\left(\inf_{S_k}\mathbb{G}(s,t)+g(s,t)\le r\right)$ is summable over $k$, so setting $M$ so that the tail of this sum past $M$ is less than $\varepsilon$ gives the desired result. □

The following theorem verifies condition (ii) of Lemma 2 for the sequence of finite sample processes $\mathbb{G}_{n,x_m,j}(s,t)+g_{n,x_m,j}(s,t)$ with $\eta/h_n\ge\|(s,t)\|$. As explained above, the case where $\eta/h_n\le\|(s,t)\|$ is handled by a separate argument.

**Theorem 17.** *Fix $m$ and $j$ with $j\in J(m)$. There exists an $\eta>0$ such that for any $r<0$, $\varepsilon>0$, there exists an $M$ and $N$ such that, for all $n\ge N$,*

$$P\left(\inf_{M<\|(s,t)\|\le\eta/h_n}\mathbb{G}_{n,x_m,j}(s,t)+g_{n,x_m,j}(s,t)\le r\right)<\varepsilon.$$

*Proof.* Let $\mathbb{G}_n(s,t)=\mathbb{G}_{n,x_m,j}(s,t)$ and $g_n(s,t)=g_{n,x_m,j}(s,t)$. Let $\eta$ be small enough that the assumptions hold for $\|x-x_m\|\le 2\eta$ and that, for some constant $C_2$, $E(Y_{i,j}|X_i=x)\ge C_2\|x-x_m\|^2$ for $\|x-x_m\|\le 2\eta$. This implies that, for $\|(s,t)\|\le h_n^{-1}\eta$,

$$g_n(s,t)\ge\frac{C_2}{h_n^{d+2}}\int_{h_ns<x-x_m<h_n(s+t)}\|x-x_m\|^2f_X(x)\,dx$$
$$\ge\frac{C_2\underline{f}}{h_n^{d+2}}\int_{h_ns<x-x_m<h_n(s+t)}\|x-x_m\|^2\,dx=C_2\underline{f}\int_{s<x<s+t}\|x\|^2\,dx_d\cdots dx_1\ge C_3\|(s,t)\|^2\prod_i t_i$$

where $C_3$ is a constant that depends only on $\underline{f}$ and $d$ and the last inequality follows from bounding the integral as explained in the proof of the previous theorem.

As in the proof of the previous theorem, let $S_k=\{k\le\|(s,t)\|\le k+1\}$ and let

$S_k^L = S_k \cap \{\prod_i t_i \le (k+1)^{-\delta}\}$ for some fixed $\delta$. We have, using Lemma 4,

$$P\left(\inf_{S_k^L} \mathbb{G}_n(s,t) + g_n(s,t) \le r\right) \le P\left(\sup_{S_k^L} |\mathbb{G}_n(s,t)| \ge |r|\right)$$
$$\le 2\left\{3(k+1)[(k+1)^d/k^{-\delta}] + 2\right\}^{2d} \exp\left(-C\frac{|r|}{(k+1)^{-\delta/2}}\right)$$

for $(k+1)^{-\delta} \ge n^{-4/(d+4)}(1+\log n)^2 \iff k+1 \le n^{4/[\delta(d+4)]}(1+\log n)^{-2/\delta}$ so, if $\delta < 4$, this will hold eventually for all $(k+1) \le h_n^{-1}\eta$ (once $h_n^{-1}\eta \le n^{4/[\delta(d+4)]}(1+\log n)^{-2/\delta} \iff \eta \le n^{(4/\delta-1)/(d+4)}(1+\log n)^{-2/\delta}$). The bound is summable over $k$ for any $\delta > 0$.

Again following the proof of the previous theorem, for $\alpha < \beta$, define $S_k^{\alpha,\beta} = S_k \cap \{(k+1)^\alpha < \prod_i t_i \le (k+1)^\beta\}$. We have, again using Lemma 4,

$$P\left(\inf_{S_k^{\alpha,\beta}} \mathbb{G}_n(s,t) + g_n(s,t) \le r\right) \le P\left(\sup_{S_k^{\alpha,\beta}} |\mathbb{G}_n(s,t)| \ge C_3 k^{2+\alpha}\right)$$
$$\le 2\left\{3(k+1)[(k+1)^d/(k^\alpha \wedge 1)] + 2\right\}^{2d} \exp\left(-C\frac{C_3 k^{2+\alpha}}{(k+1)^{\beta/2}}\right)$$

for $(k+1)^\beta \ge n^{-4/(d+4)}$ (which will hold once the same inequality holds for $\delta$ for $-\delta < \beta$) and $k+1 \le h_n^{-1}\eta$. The bound is summable over $k$ for any $\alpha, \beta$ with $4 + 2\alpha - \beta > 0$.

Thus, noting as in the previous theorem that, for any $-\delta < \alpha_1 < \alpha_2 < \ldots < \alpha_{\ell-1} < \alpha_\ell = d$, $S_k = S_k^L \cup S_k^{-\delta,\alpha_1} \cup S_k^{\alpha_1,\alpha_2} \cup \ldots \cup S_k^{\alpha_{\ell-1},\alpha_\ell}$, if we choose $\delta < 3/2$ and $\alpha_i = i$ for $i \in \{1, \ldots, d\}$ the probability of the infimum being less than or equal to $r$ over the sets indexed by $k$ for any $k \le h_n^{-1}\eta$ is bounded uniformly in $n$ by a sequence that is summable over $k$ (once $\eta \le n^{(4/\delta-1)/(d+4)}(1+\log n)^{-2/\delta}$). Thus, if we choose $M$ such that the tail of this sum past $M$ is less than $\varepsilon$ and let $N$ be large enough so that $\eta \le N^{(4/\delta-1)/(d+4)}(1+\log N)^{-2/\delta}$, we will have the desired result.

□

To complete the proof of Theorem 1, we need to show that

$$Z_{n,2} \equiv \left(n^{(d+2)/(d+4)} \inf_{\|(s-x_m,t)\|>\eta \text{ all } m \text{ s.t. } 1 \in J(m)} E_n Y_{i,1} I(s < X_i < s+t), \ldots,\right.$$
$$\left. n^{(d+2)/(d+4)} \inf_{\|(s-x_m,t)\|>\eta \text{ all } m \text{ s.t. } k \in J(m)} E_n Y_{i,k} I(s < X_i < s+t)\right) \xrightarrow{p} 0.$$

This follows from the next two lemmas.

**Lemma 5.** *Under Assumptions 1 and 2, for any $\eta > 0$, there exists some $\underline{B} > 0$ such that $EY_{i,j}I(s < X_i < s+t) \geq \underline{B}P(s < X_i < s+t)$ for all $(s,t)$ with $\|(s - x_m, t)\| > \eta$ for all $m$ with $j \in J(m)$.*

*Proof.* Given $\eta > 0$, we can make $\eta$ smaller without weakening the result, so let $\eta$ be small enough that $\|x_m - x_r\|_\infty > 2\eta$ for all $m \neq r$ with $j \in J(m) \cap J(r)$ and $f_X$ satisfies $0 < \underline{f} \leq f_X(x) \leq \overline{f} < \infty$ for some $\overline{f}$ and $\underline{f}$ on $\{x \| \|x - x_m\|_\infty \leq \eta\}$. If $\|(s - x_m, t)\| > \eta$, then $\|(s - x_m, s + t - x_m)\|_\infty > \eta/(4d)$, so it suffices to show that $EY_{i,j}I(s < X_i < s+t) \geq \underline{B}P(s < X_i < s+t)$ for all $(s,t)$ with $\|(s - x_m, s + t - x_m)\|_\infty > \eta/(4d)$. Let $\underline{\mu} > 0$ be such that $E(Y_{i,j}|X_i = x) > \underline{\mu}$ when $\|x - x_m\|_\infty \geq \eta/(8d)$ for $m$ with $j \in J(m)$. For notational convenience, let $\delta = \eta/(4d)$.

For $m$ with $j \in J(m)$, let $B(x_m, \delta) = \{x \| \|x - x_m\|_\infty \leq \delta\}$ and $B(x_m, \delta/2) = \{x \| \|x - x_m\|_\infty \leq \delta/2\}$. First, I show that, for any $(s,t)$ with $\|(s - x_m, s + t - x_m)\|_\infty \geq \delta$, $P(\{s < X_i < s+t\} \cap B(x_m, \delta) \backslash B(x_m, \delta/2)) \geq (1/3)(\underline{f}/\overline{f})P(\{s < X_i < s+t\} \cap B(x_m, \delta/2))$. Intuitively, this holds because, taking any box with a corner outside of $B(x_m, \delta)$, this box has to intersect with a substantial proportion of $B(x_m, \delta) \backslash B(x_m, \delta/2)$ in order to intersect with $B(x_m, \delta/2)$.

Formally, we have $\{s < x < s+t\} \cap B(x_m, \delta) = \{s \vee (x_m - \delta) < x < (s+t) \wedge (x_m + \delta)\}$, so that, letting $\lambda$ be the Lebesgue measure on $\mathbb{R}^d$, $\lambda(\{s < x < s+t\} \cap B(x_m, \delta)) = \prod_i [(s_i + t_i) \wedge (x_{m,i} + \delta) - s_i \vee (x_{m,i} - \delta)]$. Similarly, $\lambda(\{s < x < s+t\} \cap B(x_m, \delta/2)) = \prod_i [(s_i + t_i) \wedge (x_{m,i} + \delta/2) - s_i \vee (x_{m,i} - \delta/2)]$. For all $i$, $[(s_i + t_i) \wedge (x_{m,i} + \delta/2) - s_i \vee (x_{m,i} - \delta/2)] \leq [(s_i + t_i) \wedge (x_{m,i} + \delta) - s_i \vee (x_{m,i} - \delta)]$. For some $r$, we must have $s_r \leq x_{m,r} - \delta$ or $s_r + t_r \geq x_{m,r} + \delta$. For this $r$, we will have $[(s_r + t_r) \wedge (x_{m,r} + \delta/2) - s_r \vee (x_{m,r} - \delta/2)] \leq 2[(s_r + t_r) \wedge (x_{m,r} + \delta) - s_r \vee (x_{m,r} - \delta)]/3$. Thus, $\lambda(\{s < x < s+t\} \cap B(x_m, \delta/2)) \leq 2\lambda(\{s < x < s+t\} \cap B(x_m, \delta))/3$. It then follows that $\lambda(\{s < x < s+t\} \cap B(x_m, \delta) \backslash B(x_m, \delta/2)) \geq (1/3)\lambda(\{s < x < s+t\} \cap B(x_m, \delta))$, so that $P(\{s < x < s+t\} \cap B(x_m, \delta) \backslash B(x_m, \delta/2)) \geq (1/3)(\underline{f}/\overline{f})P(\{s < x < s+t\} \cap B(x_m, \delta))$.

Now, we use the fact that $E(Y_{i,j}|X_i)$ is bounded away from zero outside of $B(x_m, \delta/2)$, and that the proportion of $\{s < x < s+t\}$ that intersects with $B(x_m, \delta/2)$ can't be too large. We have, for any $(s,t)$ with $\|(s - x_m, s + t - x_m)\|_\infty \geq \delta$,

$$\begin{aligned}
&EY_{i,j}I(s < X_i < s+t) \geq \underline{\mu}P(\{s < X_i < s+t\} \backslash [\cup_m B(x_m, \delta/2)]) \\
&= \underline{\mu}P(\{s < X_i < s+t\} \backslash [\cup_m B(x_m, \delta)]) + \underline{\mu}\sum_m P(\{s < X_i < s+t\} \cap B(x_m, \delta) \backslash B(x_m, \delta/2)) \\
&\geq \underline{\mu}P(\{s < X_i < s+t\} \backslash [\cup_m B(x_m, \delta)]) + \underline{\mu}\sum_m (1/3)(\underline{f}/\overline{f})P(\{s < X_i < s+t\} \cap B(x_m, \delta)) \\
&\geq \underline{\mu}(1/3)(\underline{f}/\overline{f})P(s < X_i < s+t)
\end{aligned}$$

where the unions are taken over $m$ such that $j \in J(m)$. The equality in the second line follows because the sets $B(x_m, \delta)$ are disjoint.

□

**Lemma 6.** *Let $S$ be any set in $\mathbb{R}^{2d}$ such that, for some $\underline{\mu} > 0$ and all $(s,t) \in S$,$EY_{i,j}I(s < X_i < s+t) \geq \underline{\mu}P(s < X_i < s+t)$. Then, under Assumption 2, for any sequence $a_n \to \infty$ and $r < 0$,*

$$\inf_{(s,t)\in S} \frac{n}{a_n \log n} E_n Y_{i,j} I(s < X_i < s+t) > r$$

*with probability approaching 1.*

*Proof.* For $(s,t) \in S$,

$$\begin{aligned}
&\frac{n}{a_n \log n} E_n Y_{i,j} I(s < X_i < s+t) \leq r \\
&\implies \frac{n}{a_n \log n}(E_n - E)Y_{i,j}I(s < X_i < s+t) \leq r - \frac{n}{a_n \log n} EY_{i,j}I(s < X_i < s+t) \\
&\leq r - \frac{n}{a_n \log n}\underline{\mu}P(s < X_i < s+t) \leq -\left\{|r| \vee \left[\frac{n}{a_n \log n}\underline{\mu}P(s < X_i < s+t)\right]\right\} \\
&\implies \left[\frac{\frac{a_n \log n}{n}}{\frac{a_n \log n}{n} \vee P(s < X_i < s+t)}\right]^{1/2} |(E_n - E)Y_{i,j}I(s < X_i < s+t)| \\
&\geq \left[\frac{\frac{a_n \log n}{n}}{\frac{a_n \log n}{n} \vee P(s < X_i < s+t)}\right]^{1/2} \left\{\left[\frac{a_n \log n}{n}|r|\right] \vee \left[\underline{\mu}P(s < X_i < s+t)\right]\right\}.
\end{aligned}$$

If $\frac{a_n \log n}{n} \geq P(s < X_i < s+t)$, then the last line is greater than or equal to $\frac{a_n \log n}{n}|r|$. If $\frac{a_n \log n}{n} \leq P(s < X_i < s+t)$, the last line is greater than or equal to $\left[\frac{\frac{a_n \log n}{n}}{P(s<X_i<s+t)}\right]^{1/2} \underline{\mu}P(s < X_i < s+t) = \left(\frac{a_n \log n}{n}\right)^{1/2} \underline{\mu}\sqrt{P(s < X_i < s+t)} \geq \underline{\mu}\frac{a_n \log n}{n}$. Thus,

$$\begin{aligned}
&P\left(\inf_{(s,t)\in S} \frac{n}{a_n \log n} E_n Y_{i,j} I(s < X_i < s+t) \leq r\right) \\
&\leq P\left(\sup_{(s,t)\in S} \left[\frac{\frac{a_n \log n}{n}}{\frac{a_n \log n}{n} \vee P(s < X_i < s+t)}\right]^{1/2} |(E_n - E)Y_{i,j}I(s < X_i < s+t)| \geq (|r| \wedge \underline{\mu})\frac{a_n \log n}{n}\right).
\end{aligned}$$

This converges to zero by Theorem 37 in Pollard (1984) with, in the notation of that theorem,

$\mathcal{F}_n$ the class of functions of the form

$$\left[\frac{\frac{a_n \log n}{n}}{\overline{Y}^2 \frac{a_n \log n}{n} \vee P(s < X_i < s+t)}\right]^{1/2} Y_{i,j} I(s < X_i < s+t)$$

with $(s,t) \in S$, $\delta_n = \left(\frac{n}{a_n \log n}\right)^{1/2}$ and $\alpha_n = 1$. To verify the conditions of the lemma, the covering number bound holds because each $\mathcal{F}_n$ is contained in the larger class $\mathcal{F}$ of functions of the form $wY_{i,j}I(s < X_i < s+t)$ where $(s,t)$ ranges over $S$ and $w$ ranges over $\mathbb{R}$, and this larger class is a VC subgraph class. The supremum bound on functions in $\mathcal{F}_n$ holds by Assumption 2. To verify the bound on the $L^2$ norm of functions in $\mathcal{F}_n$, note that

$$E\left\{\left[\frac{\frac{a_n \log n}{n}}{\overline{Y}^2 \frac{a_n \log n}{n} \vee P(s < X_i < s+t)}\right]^{1/2} Y_{i,j} I(s < X_i < s+t)\right\}^2$$
$$\leq \frac{\frac{a_n \log n}{n}}{\frac{a_n \log n}{n} \vee P(s < X_i < s+t)} P(s < X_i < s+T) \leq \frac{a_n \log n}{n} = \delta_n^2$$

since $ab/(a \vee b) \leq a$ for any $a, b > 0$.

□

By Lemma 5, $\{\|(s - x_m, t)\| > \eta \text{ all } m \text{ s.t. } j \in J(m)\}$ satisfies the conditions of Lemma 6, so $E_n Y_{i,j} I(s < X_i < s+t)$ converges to zero at a $n/(a_n \log n)$ rate for any $a_n \to \infty$, which can be made faster than the $n^{(d+2)/(d+4)}$ rate needed to show that $Z_{n,2} \xrightarrow{p} 0$. This completes the proof of Theorem 1.

## Inference

I use the following lemma in the proof of Theorem 2

**Lemma 7.** *Let $\mathbb{H}$ be a Gaussian random process with sample paths that are almost surely in the set $C(\mathbb{T}, \mathbb{R}^k)$ of continuous functions with respect to some semimetric on the index set $\mathbb{T}$ with a countable dense subset $\mathbb{T}_0$. Then, for any set $A \in \mathbb{R}^k$ with Lebesgue measure zero, $P(\inf_{t \in \mathbb{T}} \mathbb{H}(t) \in A) \leq P(\inf_{t \in \mathbb{T}, \det var(\mathbb{H}(t)) < \varepsilon} \mathbb{H}(t) \in A$ for all $\varepsilon > 0)$.*

*Proof.* First, note that, if the infimum over $\mathbb{T}$ is in $A$, then, since $\{t \in \mathbb{T} | \det var(\mathbb{H}(t)) \geq \varepsilon\}$ and $\{t \in \mathbb{T} | \det var(\mathbb{H}(t)) < \varepsilon\}$ partition $T$, the infimum over one of these sets must be in $A$. By Proposition 3.2 in Pitt and Tran (1979), the infimum of $\mathbb{H}(t)$ over the former set has a distribution that is continuous with respect to the Lebesgue measure, so the probability of the infimum of $\mathbb{H}(t)$ over this set being in $A$ is zero. Thus, $P\left(\inf_{t\in\mathbb{T}}\mathbb{H}(t)\in A\right)\leq P\left(\inf_{t\in\mathbb{T},\det var(\mathbb{H}(t))<\varepsilon}\mathbb{H}(t)\in A\right)$. Taking $\varepsilon$ to zero along a countable sequence gives the result. □

*Proof of Theorem 2.* For $m$ from 1 to $\ell$, let $\{j_{m,1},\ldots,j_{m,|J(m)|}\}=J(m)$. Then, letting

$$\begin{aligned}\tilde{Z}\equiv&(\inf_{s,t}\mathbb{G}_{P,x_1,j_{1,1}}(s,t)+g_{P,x_1,j_{1,1}}(s,t),\ldots,\inf_{s,t}\mathbb{G}_{P,x_1,j_{1,|J(1)|}}(s,t)+g_{P,x_1,j_{1,|J(1)|}}(s,t),\ldots,\\&\inf_{s,t}\mathbb{G}_{P,x_\ell,j_{\ell,1}}(s,t)+g_{P,x_\ell,j_{\ell,1}}(s,t),\ldots,\inf_{s,t}\mathbb{G}_{P,x_\ell,j_{\ell,|J(\ell)|}}(s,t)+g_{P,x_\ell,j_{\ell,|J(\ell)|}}(s,t)),\end{aligned}$$

each element of $Z$ is the minimum of the elements of some subvector of $\tilde{Z}$, where the subvectors corresponding to different elements of $Z$ do not overlap. Thus, it suffices to show that $\tilde{Z}$ has an absolutely continuous distribution. For this, it suffices to show that, for each $m$,

$$(\inf_{s,t}\mathbb{G}_{P,x_m,j_{m,1}}(s,t)+g_{P,x_m,j_{m,1}}(s,t),\ldots,\inf_{s,t}\mathbb{G}_{P,x_m,j_{m,|J(m)|}}(s,t)+g_{P,x_m,j_{m,|J(m)|}}(s,t))$$

has an absolutely continuous distribution, since these are independent across $m$.

To this end, fix $m$ and let $\mathbb{H}(s,t)$ be the random process with sample paths in $C(\mathbb{R}^{2d},\mathbb{R}^{|J(m)|})$ defined by

$$\mathbb{H}(s,t)=(\mathbb{G}_{P,x_m,j_{m,1}}(s,t)+g_{P,x_m,j_{m,1}}(s,t),\ldots,\mathbb{G}_{P,x_m,j_{m,|J(m)|}}(s,t)+g_{P,x_m,j_{m,|J(m)|}}(s,t)).$$

By Assumption 4 $var(\mathbb{H}(s,t))=M\prod_i t_i$ for some positive definite matrix $M$, so that $\det var(\mathbb{H}(s,t))=(\det M)\left(\prod_i t_i\right)^{|J(m)|}$. Thus, $\inf_{(s,t)\in\mathbb{R}^{2d},\det var(\mathbb{H}(s,t))<\varepsilon}\mathbb{H}(s,t)\in A$ for all $\varepsilon>0$ iff. $\inf_{(s,t)\in\mathbb{R}^{2d},\prod_i t_i<\varepsilon}\mathbb{H}(s,t)\in A$ for all $\varepsilon>0$ so, by Lemma 7, $P(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A)\leq P(\inf_{(s,t)\in\mathbb{R}^{2d},\prod_i t_i<\varepsilon}\mathbb{H}(s,t)\in A$ for all $\varepsilon>0)$. For each $j$, $\prod_i t_i$ is equal to $var(\mathbb{H}_j(s,t))=\rho_j(0,(s,t))$ times some constant, where $\rho_j$ is the covariance semimetric for component $j$ given by $\rho_j((s,t),(s',t'))=var(\mathbb{H}_j(s,t)-\mathbb{H}_j(s',t'))$. Thus, there exists a constant $C$ such that $\prod_i t_i\leq\varepsilon$ implies $\rho_j(0,(s,t))<C\varepsilon$ for all $j$, so that $P(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A)\leq P(\inf_{(s,t)\in\mathbb{R}^{2d},\rho_j(0,(s,t))<C\varepsilon\text{ all }j}\mathbb{H}(s,t)\in A$ for all $\varepsilon>0)$.

Since the sample paths of $\mathbb{H}$ are almost surely continuous with respect to the semimetric $\max_j\rho_j((s,t),(s',t'))$ on the set $\|(s,t)\|\leq M$ for any finite $M$, $\inf_{\|(s,t)\|\leq M,\rho_j(0,(s,t))<C\varepsilon\text{ all }j}\mathbb{H}(s,t)\in A$ for all $\varepsilon>0$ implies that $\mathbb{H}(0)=0$ is a limit point of $A$ on this probability one set. Thus, for any set $A$ that does not have zero as a limit point, $P(\inf_{\|(s,t)\|\leq M}\mathbb{H}(s,t)\in A)=0$ for any finite $M$. Applying this to $A\backslash B_\eta(0)$ where $B_\eta(0)$ is the $\eta$-ball around 0 in $\mathbb{R}^{|J(m)|}$, we have

$$P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A\right)=P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A\cap B_\eta(0)\right)+P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A\backslash B_\eta(0)\right)$$
$$\le P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A\cap B_\eta(0)\right)+P\left(\inf_{\|(s,t)\|\le M}\mathbb{H}(s,t)\in A\backslash B_\eta(0)\right)$$
$$+P\left(\inf_{\|(s,t)\|>M}\mathbb{H}(s,t)\in A\backslash B_\eta(0)\right)$$
$$=P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A\cap B_\eta(0)\right)+P\left(\inf_{\|(s,t)\|>M}\mathbb{H}(s,t)\in A\backslash B_\eta(0)\right).$$

Noting that $P\left(\inf_{\|(s,t)\|>M}\mathbb{H}(s,t)\in A\backslash B_\eta(0)\right)$ can be made arbitrarily small by making $M$ large, this shows that $P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A\right)=P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A\cap B_\eta(0)\right)$ Taking $\eta$ to zero along a countable sequence, this shows that $P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A\right)\le P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)\in A\cap\{0\}\right)$ so that $\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{H}(s,t)$ has an absolutely continuous distribution with a possible atom at zero.

To show that there can be no atom at zero, we argue as follows. Fix $j\in J(m)$. The component of $\mathbb{H}$ corresponding to this $j$ is $\mathbb{G}_{P,x_m,j}(s,t)+g_{P,x_m,j}(s,t)$. For some constant $K$, for any $k\ge 0$, letting $s_{i,k}=(i/k,0,\dots,0)$ and $t_k=(1/k,1,\dots,1)$, we will have $g_{P,x_m,j}(s_{i,k},t_k)\le K/k$ for $i\le k$, so that

$$P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{G}_{P,x_m,j}(s,t)+g_{P,x_m,j}(s,t)=0\right)=P\left(\inf_{(s,t)\in\mathbb{R}^{2d}}\mathbb{G}_{P,x_m,j}(s,t)+g_{P,x_m,j}(s,t)\ge 0\right)$$
$$\le P\left(\mathbb{G}_{P,x_m,j}(s_{i,k},t_k)+g_{P,x_m,j}(s_{i,k},t_k)\ge 0\text{ all }i\in\{0,\dots,k\}\right)$$
$$\le P\left(\mathbb{G}_{P,x_m,j}(s_{i,k},t_k)+K/k\ge 0\text{ all }i\in\{0,\dots,k\}\right)$$
$$=P\left(\sqrt{k}\mathbb{G}_{P,x_m,j}(s_{i,k},t_k)+K/\sqrt{k}\ge 0\text{ all }i\in\{0,\dots,k\}\right)$$
$$=P\left(\mathbb{G}_{P,x_m,j}(s_{i,1},t_1)+K/\sqrt{k}\ge 0\text{ all }i\in\{0,\dots,k\}\right).$$

The final line is the probability of $k+1$ iid normal random variables each being greater than or equal to $-K/\sqrt{k}$, which can be made arbitrarily small by making $k$ large. □

*proof of Theorem 3.* This follows immediately from the continuity of the asymptotic distribution (see Politis, Romano, and Wolf, 1999). □

*proof of Theorem 4.* It suffices to show that, for every subsequence, there exists a further subsequence along which the distribution of $\hat{Z}$ converges weakly to the distribution of $Z$.

Given a subsequence, let the further subsequence be such that the convergence in probability in Assumption 6 is with probability one.

For any fixed $B > 0$, the processes

$$\left[\hat{\mathbb{G}}_{P,x_k}(s,t) + \hat{g}_{P,x_k}(s,t)\right] I(\|(s,t)\| \leq B_n)$$

are, along this subsequence, Gaussian processes with mean functions and covariance kernels converging with probability one to those of the distribution being estimated uniformly in $\|(s,t)\| \leq B$. Thus, with probability one, the distributions of these processes converge weakly to the distribution of the process being estimated along this subsequence taken as random processes on $\|(s,t)\| \leq B$. Thus, to get the weak convergence of the elementwise infimum, we just need to verify part (ii) of Lemma 2. To this end, note that, along the further subsequence, the infimum of

$$\left[\hat{\mathbb{G}}_{P,x_k,j}(s,t) + \hat{g}_{P,x_k,j}(s,t)\right] I(\|(s,t)\| \leq B_n)$$

is eventually bounded from below (in the stochastic dominance sense) by the infimum of a process defined the same way as

$$\mathbb{G}_{P,x_k,j}(s,t) + g_{P,x_k,j}(s,t),$$

but with $E(m_{J(k)}(W_i,\theta)m_{J(k)}(W_i,\theta)'|X = x_k)$ replaced by $2E(m_{J(k)}(W_i,\theta)m_{J(k)}(W_i,\theta)'|X = x_k)$, and $V(x_k)$ replaced by $V(x_k)/2$. Once $n$ is large enough that this holds along this further subsequence, part (ii) of Lemma 2 will hold by Lemma 16 applied to this process. □

*proof of Corollary 2.* By Theorem 4, the distribution of $S(\hat{Z})$ converges weakly conditionally in probability to the distribution of $S(Z)$, and by Theorem 1, $n^{(d_X+2)/(d_X+4)}S(T_n(\theta)) \xrightarrow{d} S(Z)$. $S(Z)$ has a continuous distribution by Theorem 2, so the result follows by standard arguments. □

## Other Shapes of the Conditional Mean

This section contains the proofs of the results in Section 5, which extend the results of Section 3 to other shapes of the conditional mean. First, I show how Assumption 1 implies Assumption 7 with $\gamma = 2$. Next, I prove Theorem 5, which gives an interpretation of Assumption 6 in terms of conditions on the number of bounded derivatives in the one dimensional case.

Finally, I prove Theorem 6, which derives the asymptotic distribution of the KS statistic under these assumptions. The proof is mostly the same as the proof of Theorem 1, and I present only the parts of the proof that differ, referring to the proof of Theorem 1 for the parts that do not need to be changed.

To see that, under part (ii) from Assumption 1, Assumption 7 will hold with $\gamma = 2$, note that, by a second order Taylor expansion, for some $x^*(x)$ between $x$ and $x_k$,

$$\frac{\bar{m}_j(\theta, x) - \bar{m}_j(\theta, x_k)}{\|x - x_k\|^2} = \frac{(x - x_k)V_j(x^*(x))(x - x_k)}{2\|x - x_k\|^2} = \frac{1}{2}\frac{x - x_k}{\|x - x_k\|}V_j(x^*(x))\frac{x - x_k}{\|x - x_k\|}.$$

Thus, letting $\psi_{j,k}(t) = \frac{1}{2}tV_j(x_k)t$ we have

$$\sup_{\|x - x_k\| \le \delta}\left\|\frac{\bar{m}_j(\theta, x) - \bar{m}_j(\theta, x_k)}{\|x - x_k\|^2} - \psi_{j,k}\left(\frac{x - x_k}{\|x - x_k\|}\right)\right\|$$
$$= \sup_{\|x - x_k\| \le \delta}\left\|\frac{1}{2}\frac{x - x_k}{\|x - x_k\|}V_j(x^*(x))\frac{x - x_k}{\|x - x_k\|} - \frac{1}{2}\frac{x - x_k}{\|x - x_k\|}V_j(x_k)\frac{x - x_k}{\|x - x_k\|}\right\|.$$

This goes to zero as $\delta \to 0$ by the continuity of the second derivative matrix.

The proof of Theorem 5 below shows that, in the one dimensional case, Assumption 1 follows more generally from conditions on higher order derivatives.

*proof of Theorem 5.* It suffices to consider the case where $d_Y = 1$. First, suppose that $\mathcal{X}_0$ has infinitely many elements. Let $\{x_k\}_{k=1}^{\infty}$ be a nonrepeating sequence of elements in $\mathcal{X}_0$. Since $\mathcal{X}_0$ is compact, this sequence must have a subsequence that converges to some $\tilde{x} \in \mathcal{X}_0$. If $\bar{m}(\theta, x)$ had a nonzero $r$th derivative at $\tilde{x}$ for some $r < p$, then, by Lemma 8 below, $\bar{m}(\theta, x)$ would be strictly greater than $\bar{m}(\theta, \tilde{x})$ for $x$ in some neighborhood of $\tilde{x}$, a contradiction. Thus, a $p$th order taylor expansion gives, using the notation $D_r(x) = \delta^r/\delta x^r \bar{m}(\theta, x)$ for $r \le p$, $\bar{m}(\theta, x) - \bar{m}(\theta, \tilde{x}) = D_p(x^*(x))(x - \tilde{x})^p/p! \le \bar{D}|x - \tilde{x}|^p/p!$ where $\bar{D}$ is a bound on the $p$th derivative and $x^*(x)$ is some value between $x$ and $\tilde{x}$.

If $\mathcal{X}_0$ has finitely many elements, then, for each $x_0 \in \mathcal{X}_0$, a $p$th order Taylor expansion gives $\bar{m}(\theta, x) - \bar{m}(\theta, x_0) = D_1(x_0)(x - x_0) + \frac{1}{2}D_2(x_0)(x - x_0)^2 + \cdots + \frac{1}{p!}D_p(x^*(x))(x - x_0)^p$. If, for some $r < p$, $D_r(x_0) \neq 0$ and $D_{r'}(x_0) = 0$ for $r' < r$, then Assumption 7 will hold at $x_0$ with $\gamma = r$. If not, we will have $\bar{m}(\theta, x) - \bar{m}(\theta, x_0) \le \bar{D}|x - x_0|^p/p!$ for all $x$. □

**Lemma 8.** *Suppose that $g : [\underline{x}, \overline{x}] \subseteq \mathbb{R} \to \mathbb{R}$ is minimized at some $x_0$. If the least nonzero derivative of $g$ is continuous at $x_0$, then, for some $\varepsilon > 0$, $g(x) > g(x_0)$ for $|x - x_0| \le \varepsilon$, $x \neq x_0$.*

*Proof.* Let $p$ be the least integer such that the $p$th derivative $g^{(p)}(x_0)$ is nonzero. By a $p$th order Taylor expansion, $g(x) - g(x_0) = g^{(p)}(x^*(x))(x - x_0)^p$ for some $x^*(x)$ between $x$ and $x_0$. By continuity of $g^{(p)}(x)$, $|g^{(p)}(x^*(x)) - g^{(p)}(x_0)| > |g^{(p)}(x_0)|/2$ for $x$ close enough to $x_0$, so that $g(x) - g(x_0) = g^{(p)}(x^*(x))(x - x_0)^p \geq |g^{(p)}(x_0)|/2|x - x_0|^p > 0$ (the $p$th derivative must have the same sign as $x - x_0$ if $p$ is odd in order for $g$ to be minimized at $x_0$). □

I now prove Theorem 6. I prove the theorem under the assumption that $\gamma(j,k) = \gamma$ for all $(j,k)$ with $j \in J(k)$. The general case follows from applying the argument to neighborhoods of each $x_k$, and getting faster rates of convergence for $(j,k)$ such that $\gamma(j,k) < \gamma$. The proof is the same as the proof of Theorem 1 with the following modifications.

First, Theorem 15 must be modified to the following theorem, with the new definition of $g_{P,x_k,j}(s,t)$.

**Theorem 18.** *Let $h_n = n^{-\beta}$ for some $0 < \beta < 1/d_X$. Let*

$$\mathbb{G}_{n,x_m}(s,t) = \frac{\sqrt{n}}{h_n^{d/2}}(E_n - E)Y_{i,J(m)}I(h_n s < X_i - x_m < h_n(s+t))$$

*and let $g_{n,x_m}(s,t)$ have $j$th element*

$$g_{n,x_m,j}(s,t) = \frac{1}{h_n^{d_X+\gamma}}EY_{i,j}I(h_n s < X_i - x_m < h_n(s+t))$$

*if $j \in J(m)$ and zero otherwise. Then, for any finite $M$, $(\mathbb{G}_{n,x_1}(s,t), \ldots, \mathbb{G}_{n,x_\ell}(s,t)) \xrightarrow{d} (\mathbb{G}_{P,x_1}(s,t), \ldots, \mathbb{G}_{P,x_\ell}(s,t))$ taken as random processes on $\|(s,t)\| \leq M$ with the supremum norm and $g_{n,x_m}(s,t) \to g_{P,x_m}(s,t)$ uniformly in $\|(s,t)\| \leq M$ where $\mathbb{G}_{P,x_m}(s,t)$ and $g_{P,x_m}(s,t)$ are defined as in Theorem 1 for $m$ from 1 to $\ell$.*

*Proof.* The proof of the first display is the same. For the proof of the claim regarding $g_{n,x_m}(s,t)$, we have

$$\begin{aligned}
g_{n,x_m,j}(s,t) &= \frac{1}{h_n^{d_X+\gamma}}\int_{h_n s < x - x_m < h_n(s+t)} \psi_{j,k}\left(\frac{x - x_m}{\|x - x_m\|}\right)\|x - x_m\|^\gamma f_X(x_m)\,dx \\
&+ \frac{1}{h_n^{d_X+\gamma}}\int_{h_n s < x - x_m < h_n(s+t)} \psi_{j,k}\left(\frac{x - x_m}{\|x - x_m\|}\right)\|x - x_m\|^\gamma [f_X(x) - f_X(x_m)]\,dx \\
&+ \frac{1}{h_n^{d_X+\gamma}}\int_{h_n s < x - x_m < h_n(s+t)} [\bar{m}_j(\theta, x) - \bar{m}_j(\theta, x_m) \\
&- \psi_{j,k}\left(\frac{x - x_m}{\|x - x_m\|}\right)\|x - x_m\|^\gamma\Bigg] f_X(x)\,dx.
\end{aligned}$$

The first term is equal to $g_{P,x_m,j}(s,t)$ by a change of variable $x$ to $h_n x + x_m$ in the integral. The second term is bounded by $g_{P,x_m,j}(s,t)\sup_{\|x-x_m\|\le 2h_n M}[f_X(x)-f_X(x_m)]/f_X(x_m)$, which goes to zero uniformly in $\|(s,t)\|\le M$ by continuity of $f_X$. The third term is equal to (using the same change of variables)

$$\int_{s<x<s+t}\left[\frac{\bar{m}_j(\theta,h_n x+x_m)-\bar{m}_j(\theta,x_m)}{h_n^\gamma}-\psi_{j,k}\left(\frac{x}{\|x\|}\right)\|x\|^\gamma\right]f_X(x)\,dx$$
$$=\int_{s<x<s+t}\|x\|^\gamma\left[\frac{\bar{m}_j(\theta,h_n x+x_m)-\bar{m}_j(\theta,x_m)}{\|h_n x\|^\gamma}-\psi_{j,k}\left(\frac{x}{\|x\|}\right)\right]f_X(x)\,dx.$$

For $\|(s,t)\|\le M$, this is bounded by a constant times

$$\sup_{\|x\|\le 2M}\left\|\frac{\bar{m}_j(\theta,h_n x+x_m)-\bar{m}_j(\theta,x_m)}{\|h_n x\|^\gamma}-\psi_{j,k}\left(\frac{x}{\|x\|}\right)\right\|,$$

which goes to zero as $n\to\infty$ by Assumption 7. □

The drift term and the mean zero term will be of the same order of magnitude if $\sqrt{n}/h_n^{d_X/2}=1/h_n^{d_X+\gamma}\Leftrightarrow h_n=n^{-1/(d_X+2\gamma)}$, so that

$$n^{(d_X+\gamma)/(d+2\gamma)}(E_nY_{i,J(1)}I(h_n s<X-x_1<h_n(s+t)),\dots,E_nY_{i,J(\ell)}I(h_n s<X-x_\ell<h_n(s+t))$$
$$=(\mathbb{G}_{n,x_1}(s,t)+g_{n,x_1}(s,t),\dots,\mathbb{G}_{n,x_\ell}(s,t)+g_{n,x_\ell}(s,t))$$
$$\xrightarrow{d}(\mathbb{G}_{P,x_1}(s,t)+g_{P,x_1}(s,t),\dots,\mathbb{G}_{P,x_m}(s,t)+g_{P,x_m}(s,t))$$

taken as stochastic processes in $\{\|(s,t)\|\le M\}$ with the supremum norm. From now on, let $h_n=n^{-1/(d+2\gamma)}$ so that this will hold.

Lemmas 3 and 4 hold as stated, except for the condition in Lemma 4 that $\varepsilon\ge n^{-4/(d+4)}(1+\log n)^2$ must be replaced by $\varepsilon\ge n^{2\gamma/(d+2\gamma)}(1+\log n)^2$ so that $h_n^{d/2}2^{d/2}\varepsilon^{1/2}\ge n^{-1/2}(1+\log n)$, which implies the fourth inequality in the last display in the proof of this lemma, holds for the sequence $h_n$ in the general case.

The next part of the proof that needs to be modified is the proofs of Theorems 16 and 17. For this, note that, for some constants $C_1$ and $\eta>0$

$$g_{P,x_m,j}(s,t)\ge C_1\|(s,t)\|^\gamma\prod_i t_i \tag{2}$$

and, for $\|(s,t)\| \leq \eta/h_n$,

$$g_{n,x_m,j}(s,t) \geq C_1 \|(s,t)\|^\gamma \prod_i t_i \tag{3}$$

for all $m$ and $j$. To see this, note that

$$g_{n,x_m,j}(s,t) = E\frac{1}{h_n^{d_X+\gamma}} EY_{i,j} I(h_n s < X_i - x_m < h_n(s+t))$$
$$= \frac{1}{h_n^{d_X+\gamma}} \int_{h_n s < x - x_m < h_n(s+t)} \bar{m}(\theta, x) f_X(x)\, dx = \int_{s<x<s+t} \frac{\bar{m}(\theta, h_n x + x_m)}{\|h_n x\|^\gamma} \|x\|^\gamma f_X(h_n x + x_m)\, dx$$

where the last equality follows from the change of variables $x$ to $h_n x + x_m$. For small enough $\eta$, this is greater than or equal to $\frac{1}{2}\int_{s<x<s+t} \underline{\psi} \|x\|^\gamma f_X(x_m)\, dx$ for $\|(s,t)\| \leq \eta/h_n$ by Assumption 7 and the continuity of $f_X$. By definition, $g_{P,x_m,j}(s,t)$ is also greater than or equal to a constant times $\int_{s<x<s+t} \|x\|^\gamma\, dx$. To see that this is greater than or equal to a constant times $\|(s,t)\|^\gamma \prod_i t_i$, note that the Euclidean norm is equivalent to the norm $(s,t) \mapsto \max_i \max\{|s_i|, |s_i + t_i|\}$ and let $i^*$ be an index such that $|s_{i^*}| = \max_i \max\{|s_i|, |s_i + t_i|\}$ or $|s_{i^*} + t_{i^*}| = \max_i \max\{|s_i|, |s_i + t_i|\}$. In the former case, we will have $\|x\| \geq |s_{i^*}|/2$ for $x$ on the set $\{s_{i^*} \leq x_{i^*} \leq s_{i^*} + |s_{i^*}|/2\} \cap \{s < x < s+t\}$, which has Lebesgue measure $\left(\prod_{i \neq i^*} t_i\right) \cdot |s_{i^*}|/2 \geq \left(\prod_{i \neq i^*} t_i\right) \cdot t_{i^*}/4$, so that $\int_{s<x<s+t} \|x\|^\gamma\, dx \geq (\max_i \max\{|s_i|, |s_i + t_i|\}/2)^\gamma \prod_i t_i/4$, and a symmetric argument holds in the latter case.

With these inequalities in hand, the modified proofs of Theorems 16 and 17 are as follows.

*proof of Theorem 16 for general case.* Let $\mathbb{G}(s,t) = \mathbb{G}_{P,x_m,j}(s,t)$ and $g(s,t) = g_{P,x_m,j}(s,t)$. Let $S_k = \{k \leq \|(s,t)\| \leq k+1\}$ and let $S_k^L = S_k \cap \{\prod_i t_i \leq (k+1)^{-\delta}\}$ for some fixed $\delta$. By Lemma 3,

$$P\left(\inf_{S_k^L} \mathbb{G}(s,t) + g(s,t) \leq r\right) \leq P\left(\sup_{S_k^L} |\mathbb{G}(s,t)| \geq |r|\right)$$
$$\leq \left\{3(k+1)[(k+1)^d/k^{-\delta}] + 2\right\}^{2d} \exp\left(-Cr^2(k+1)^\delta\right)$$

for $k$ large enough where $C$ depends only on $d$. Thus, the infimum over each $S_k^L$ is summable over $k$.

For any $\underline{\beta}$ and $\overline{\beta}$ with $\underline{\beta} < \overline{\beta}$, let $S_k^{\underline{\beta},\overline{\beta}} = S_k \cap \{(k+1)^{\underline{\beta}} < \prod_i t_i \leq (k+1)^{\overline{\beta}}\}$. Using Lemma

3 and (2),

$$P\left(\inf_{S_k^{\underline{\beta},\overline{\beta}}} \mathbb{G}(s,t) + g(s,t) \le r\right) \le P\left(\sup_{S_k^{\underline{\beta},\overline{\beta}}} |\mathbb{G}(s,t)| \ge C_1 k^{\gamma+\underline{\beta}}\right)$$
$$\le \left\{3(k+1)[(k+1)^d/((k+1)^{\overline{\beta}} \wedge 1)] + 2\right\}^{2d} \exp\left(-CC_1^2 \frac{k^{2\gamma+2\underline{\beta}}}{(k+1)^{\overline{\beta}}}\right).$$

This is summable over $k$ if $2\gamma + 2\underline{\beta} - \overline{\beta} > 0$.

Now, note that, since $\prod_i t_i \le (k+1)^d$ on $S_k$, we have, for any $-\delta < \beta_1 < \beta_2 < \ldots < \beta_{\ell-1} < \beta_\ell = d$, $S_k = S_k^L \cup S_k^{-\delta,\beta_1} \cup S_k^{\beta_1,\beta_2} \cup \ldots \cup S_k^{\beta_{\ell-1},\beta_\ell}$. If we choose $0 < \delta < \gamma$, $\beta_1 = 0$, $\beta_2 = \gamma$, and $\beta_{i+1} = (2\beta_i) \wedge d$ for $i \ge 2$, the arguments above will show that the probability of the infimum being less than or equal to $r$ over $S_k^L$, $S_k^{-\delta,\beta_1}$ and each $S_k^{\beta_i,\beta_{i+1}}$ is summable over $k$, so that $P\left(\inf_{S_k} \mathbb{G}(s,t) + g(s,t) \le r\right)$ is summable over $k$, so setting $M$ be such that the tail of this sum past $M$ is less than $\varepsilon$ gives the desired result. □

*proof of Theorem 17 for the general case.* Let $\mathbb{G}_n(s,t) = \mathbb{G}_{n,x_m,j}(s,t)$ and $g_n(s,t) = g_{n,x_m,j}(s,t)$. Let $\eta$ be small enough that (3) holds.

As in the proof of the previous theorem, let $S_k = \{k \le \|(s,t)\| \le k+1\}$ and let $S_k^L = S_k \cap \{\prod_i t_i \le (k+1)^{-\delta}\}$ for some fixed $\delta$. We have, using Lemma 4,

$$P\left(\inf_{S_k^L} \mathbb{G}_n(s,t) + g_n(s,t) \le r\right) \le P\left(\sup_{S_k^L} |\mathbb{G}_n(s,t)| \ge |r|\right)$$
$$\le \left\{6(k+1)[(k+1)^d/k^{-\delta}] + 2\right\}^{2d} \exp\left(-C\frac{|r|}{(k+1)^{-\delta/2}}\right)$$

for $(k+1)^{-\delta} \ge n^{-2\gamma/(d+2\gamma)}(1+\log n)^2 \iff k+1 \le n^{2\gamma/[\delta(d+2\gamma)]}(1+\log n)^{-2/\delta}$ so, if $\delta < 2\gamma$, this will hold eventually for all $(k+1) \le h_n^{-1}\eta$ (once $h_n^{-1}\eta \le n^{2\gamma/[\delta(d+2\gamma)]}(1+\log n)^{-2/\delta} \iff \eta \le n^{2\gamma/[\delta(d+2\gamma)]}n^{-1/(d+2\gamma)}(1+\log n)^{-2/\delta} = n^{(2\gamma/\delta-1)/(d+2\gamma)}(1+\log n)^{-2/\delta}$). The bound is summable over $k$ for any $\delta > 0$.

Again following the proof of the previous theorem, for $\underline{\beta} < \overline{\beta}$, define $S_k^{\underline{\beta},\overline{\beta}} = S_k \cap \{(k+1)^{\underline{\beta}} <$

$\prod_i t_i \leq (k+1)^{\overline{\beta}}\}$. We have, again using Lemma 4,

$$P\left(\inf_{S_k^{\underline{\beta},\overline{\beta}}} \mathbb{G}_n(s,t) + g_n(s,t) \leq r\right) \leq P\left(\sup_{S_k^{\underline{\beta},\overline{\beta}}} |\mathbb{G}_n(s,t)| \geq C_1 k^{\gamma+\underline{\beta}}\right)$$
$$\leq \left\{6(k+1)[(k+1)^d/(k^{\underline{\beta}} \wedge 1)] + 2\right\}^{2d} \exp\left(-C\frac{C_1 k^{\gamma+\underline{\beta}}}{(k+1)^{\overline{\beta}/2}}\right)$$

for $(k+1)^{\overline{\beta}} \geq n^{-2\gamma/(d+2\gamma)}(1+\log n)^2$ (which will hold once the same inequality holds for $\delta$ for $-\delta < \overline{\beta}$) and $k+1 \leq h_n^{-1}\eta$. The bound is summable over $k$ for any $\underline{\beta}, \overline{\beta}$ with $2\gamma + 2\underline{\beta} - \overline{\beta} > 0$.

Thus, noting as in the previous theorem that, for any $-\delta < \beta_1 < \beta_2 < \ldots < \beta_{\ell-1} < \beta_\ell = d$, $S_k = S_k^L \cup S_k^{-\delta,\beta_1} \cup S_k^{\beta_1,\beta_2} \cup \ldots \cup S_k^{\beta_{\ell-1},\beta_\ell}$, if we choose $0 < \delta < \gamma$, $\beta_1 = 0$, $\beta_2 = \gamma$, and $\beta_{i+1} = (2\beta_i) \wedge d$ for $i \geq 2$, the arguments above will show that the probability of the infimum being less than or equal to $r$ over the sets indexed by $k$ for any $k \leq h_n^{-1}\eta$ is bounded uniformly in $n$ by a sequence that is summable over $k$ (once $\eta \leq n^{(2\gamma/\delta-1)/(d+2\gamma)}(1+\log n)^{-2/\delta}$). Thus, if we choose $M$ such that the tail of this sum past $M$ is less than $\varepsilon$ and let $N$ be large enough so that $\eta \leq N^{(2\gamma/\delta-1)/(d+2\gamma)}(1+\log N)^{-2/\delta}$, we will have the desired result.

□

Lemmas 5 and 6 hold as stated with the same proofs, so the rest of the proof is the same as in the $\gamma = 2$ case. The $n/(a_n \log n)$ rate for $Z_{n,2}$ is still faster than the $n^{(d+\gamma)/(d+2\gamma)}$ rate for $a_n$ increasing slowly enough.

The proof of Theorem 2 for the limiting process is the same as before. The only place the drift term is used is in ensuring that the inequality $g_{P,x_m,j}(s_{i,k}, t_k) \leq K/k$ holds in the last display in the proof of the theorem, which is still the case.

## Testing Rate of Convergence Conditions: Subsampling

First, I collect results on the rate estimate $\hat{\beta}$ defined in (1). The next lemma bounds $\hat{\beta}$ when the statistic may not converge at a polynomial rate. Throughout the following, $S_n$ is a statistic on $\mathbb{R}$ with cdf $J_n(x)$ and quantile function $J_n^{-1}(t)$. $L_{n,b}(x|\tau)$ and $\tilde{L}_{n,b}(x|\tau)$ are defined as in the body of the paper, with $S(T_n(\theta))$ replaced by $S_n$.

**Lemma 9.** *Let $S_n$ be a statistic such that, for some sequence $\tau_n$ and $x > 0$, $\tau_n J_n^{-1}(t) \geq x$ for large enough $n$. Then, if $\tau_b S_n \xrightarrow{p} 0$ and $b/n \to 0$, we will have, for any $\varepsilon > 0$, $L_{n,b}^{-1}(t+\varepsilon|\tau) \geq x - \varepsilon$ with probability approaching one.*

*Proof.* It suffices to show $L_{n,b}(x-\varepsilon|\tau) \leq t+\varepsilon$ with probability approaching one. On the event $E_n \equiv \{|\tau_b S_{\mathcal{S}}| \leq \varepsilon\}$, which has probability approaching one, $L_{n,b}(x-\varepsilon|\tau) \leq \tilde{L}_{n,b}(x|\tau)$. We also have $E[L_{n,b}(x|\tau)] = P(\tau_b S_{\mathcal{S}} \leq x) = J_b(x/\tau_b) \leq t$ by assumption. Thus,

$$\begin{aligned}
P(L_{n,b}(x-\varepsilon|\tau) \leq t+\varepsilon) &\geq P\left(\left\{\tilde{L}_{n,b}(x|\tau) \leq t+\varepsilon\right\} \cap E_n\right) \\
&\geq P\left(\left\{\tilde{L}_{n,b}(x|\tau) \leq E[L_{n,b}(x|\tau)]+\varepsilon\right\} \cap E_n\right).
\end{aligned}$$

This goes to one by standard arguments. □

**Lemma 10.** *Let $\hat{\beta}_a$ be the estimator defined in Section 6.1, of any other estimator such that $\hat{\beta}_a = \frac{-\log L_{n,b_1}^{-1}(t|1)+\mathcal{O}_p(1)}{\log b_1 - \mathcal{O}_p(1)}$. Suppose that, for some $x_\ell > 0$ and $\beta_u$, $x_u n^{\beta_u} \leq J_n^{-1}(t-\varepsilon)$ eventually and $b_1^{\beta_u} S_n \xrightarrow{p} 0$. Then, for any $\varepsilon > 0$, we will have $\hat{\beta}_a \leq \hat{\beta}_u + \varepsilon$ with probability approaching one.*

*Proof.* We have

$$\hat{\beta}_a = -\frac{\log L_{n,b_1}^{-1}(t|1)}{\log b_1} + o_p(1) = \frac{\beta_u \log b_1 - \log L_{n,b_1}^{-1}(t|b^{\beta_u})}{\log b_1} + o_p(1) \leq \beta_u - \frac{\log(x_u/2)}{\log b_1} + o_p(1) \xrightarrow{p} \beta_u$$

where the inequality holds with probability approaching one by Lemma 9. □

The following lemma shows that the asymptotic distribution of the KS statistic is strictly increasing on its support, which is needed for the estimates of the rate of convergence in Politis, Romano, and Wolf (1999) to converge at a fast enough rate that they can be used in the subsampling procedure.

**Lemma 11.** *Under Assumptions 1, 2, 3, 4 and 5 with part (ii) of Assumption 1 replaced by Assumption 7, if $S$ is convex, then the the asymptotic distribution $S(Z)$ in Theorem 6 satisfies $P(S(Z) \in (a,\infty)) = 1$ for some $a$, and the cdf of $S(Z)$ is strictly increasing on $(a,\infty)$.*

*Proof.* First, note that, for any concave functions $f_1,\ldots,f_{d_Y}$, $f_i : V_i \to \mathbb{R}$, for some vector space $V_i$, $x \mapsto S(f_1(x_1),\ldots,f_{d_Y}(x_{d_Y}))$ is convex, since, for any $\lambda \in (0,1)$,

$$\begin{aligned}
&S(f_1(\lambda x_{a,1}+(1-\lambda)x_{b,1}),\ldots,f_k(\lambda x_{a,d_Y}+(1-\lambda)x_{b,d_Y})) \\
&\geq S(\lambda f_1(x_{a,1})+(1-\lambda)f_k(x_{b,1}),,\ldots,\lambda f_k(x_{a,d_Y})+(1-\lambda)f_k(x_{b,d_Y})) \\
&\geq \lambda S(f_1(x_{a,1}),\ldots,f_k(x_{a,d_Y}))+(1-\lambda)S(f_1(x_{b,1}),\ldots,f_k(x_{b,d_Y}))
\end{aligned}$$

where the first inequality follows since $S$ is decreasing in each argument and by concavity of the $f_k$s, and the second follows by convexity of $S$.

$S(Z)$ can be written as, for some random processes $\mathbb{H}_1(t),\dots,\mathbb{H}_{d_Y}(t)$ with continuous sample paths and $\mathbb{T} \equiv \mathbb{R}^{|\mathcal{X}_0|\cdot 2d_X}$, $S(\inf_{t\in\mathbb{T}}\mathbb{H}_1(t),\dots,\inf_{t\in\mathbb{T}}\mathbb{H}_{d_Y}(t))$. Since the infimum of a real valued function is a concave functional, this is a convex function of the sample paths of $(\mathbb{H}_1(t),\dots,\mathbb{H}_{d_Y}(t))$. The result follows from Theorem 11.1 in Davydov, Lifshits, and Smorodina (1998) as long as the vector of random processes can be given a topology for which this function is lower semi-continuous. In fact, this step can be done away with by noting that, for $\mathbb{T}_0$ a countable dense subset of $\mathbb{T}$ and $\mathbb{T}_\ell$ the first $\ell$ elements of this subset, $S(\inf_{t\in\mathbb{T}_\ell}\mathbb{H}_1(t),\dots,\inf_{t\in\mathbb{T}_\ell}\mathbb{H}_{d_Y}(t)) \overset{d}{\to} S(\inf_{t\in\mathbb{R}^{2d}}\mathbb{H}_1(t),\dots,\inf_{t\in\mathbb{R}^{2d}}\mathbb{H}_{d_Y}(t))$ as $\ell\to\infty$, so, letting $F_\ell$ be the cdf of $S(\inf_{t\in\mathbb{T}_\ell}\mathbb{H}_1(t),\dots,\inf_{t\in\mathbb{T}_\ell}\mathbb{H}_{d_Y}(t))$, applying Proposition 11.3 of Davydov, Lifshits, and Smorodina (1998) for each $F_\ell$ shows that $\Phi^{-1}(F_\ell(t))$ is concave for each $\ell$, so, by convergence in distribution, this holds for $S(Z)$ as well. □

The same result in Davydov, Lifshits, and Smorodina (1998) could also be used in the proof of Theorem 2 to show that the distribution of $S(Z)$ is continuous except possibly at the infimum of its support, but an additional argument would be needed to show that, if such an atom exists, it would have to be at zero. In the proof of Theorem 2, this is handled by using the results of Pitt and Tran (1979) instead.

We are now ready to prove Theorem 7.

*proof of Theorem 7.* First, suppose that Assumption 1 holds with part (ii) of Assumption 1 replaced by Assumption 7 for some $\underline{\gamma} < \gamma < \overline{\gamma}$ and $\mathcal{X}_0$ nonempty. By Theorem 6, $n^{(d_X+\gamma)/(d_X+2\gamma)}S(T_n(\theta))$ converges in distribution to a continous distribution. Thus, by Lemma 10, $\hat{\beta}_a \overset{p}{\to} (d_X+\gamma)/(d_X+2\gamma)$, so $\hat{\beta}_a > \underline{\beta} = (d_X+\overline{\gamma})/(d_X+2\overline{\gamma})$ with probability approaching one. On this event, the test uses the subsample estimate of the $1-\alpha$ quantile with rate estimate $\hat{\beta}\wedge\overline{\beta}$. By Theorem 8.2.1 in Politis, Romano, and Wolf (1999), $\hat{\beta}\wedge\overline{\beta} = (d_X+\gamma)/(d_X+2\gamma) + o_p((\log n)^{-1})$ as long as the asymptotic distribution of $n^{(d_X+\gamma)/(d_X+2\gamma)}S(T_n(\theta))$ is increasing on the smallest interval $(k_0,k_1)$ on which the asymptotic distribution has probability one. This holds by Lemma 11. By Theorem 8.3.1 in Politis, Romano, and Wolf (1999), the $o_p((\log n)^{-1})$ rate of convergence for the rate estimate $\hat{\beta}\wedge\overline{\beta}$ implies that the probability of rejecting converges to $\alpha$.

Next, suppose that Assumption 1 holds with part (ii) of Assumption 1 replaced by Assumption 7 for $\gamma = \overline{\gamma}$. The test that compares $n^{1/2}S(T_n(\theta))$ to a positive critical value will fail to reject with probability approaching one in this case, so, on an event with probability approaching one, the test will reject only if $\hat{\beta}_a \geq \underline{\beta}$ and the subsampling test with rate $\hat{\beta} \wedge \overline{\beta}$ rejects. Thus, the probability of rejecting is asymptotically no greater than the probability of rejecting with the subsampling test with rate $\hat{\beta} \wedge \overline{\beta}$, which has asymptotic level $\alpha$ under these conditions by the argument above.

Now, consider the case where, for some $x_0 \in \mathcal{X}_0$ and $B < \infty$, $\bar{m}_j(\theta, x) \leq B\|x - x_0\|^\gamma$ for some $\gamma > \bar{\gamma}$. Let $\tilde{m}_j(W_i, \theta) = m_j(W_i, \theta) + (B\|x - x_0\|^\gamma - \bar{m}_j(\theta, x))$. Then $\tilde{m}_j(W_i, \theta) \geq m_j(W_i, \theta)$, and $\tilde{m}_j(W_i, \theta)$ satisfies the assumptions of Theorems 6 and 2, so

$$n^{(d_X+\gamma)/(d_X+2\gamma)} S(T_n(\theta)) \geq n^{(d_X+\gamma)/(d_X+2\gamma)} S(0, \ldots, 0, \inf_{s,t} E_n \tilde{m}_j(W_i, \theta) I(s < X_i < s + t), 0, \ldots, 0)$$

and the latter quantity converges in distribution to a continuous random variable that is positive with probability one. Thus, by Lemma 10, for any $\varepsilon > 0$, $\hat{\beta}_a < (d_X + \gamma)/(d_X + 2\gamma) + \varepsilon$ with probability approaching one. For $\varepsilon$ small enough, this means that $\hat{\beta}_a < (d_X + \overline{\gamma})/(d_X + 2\overline{\gamma})$ with probability approaching one. Thus, the procedure uses an asymptotically level $\alpha$ test with probability approaching one.

The remaining case is where $\bar{m}_j(\theta, x)$ is bounded from below away from zero. If $m_j(W_i, \theta) \geq 0$ for all $j$ with probability one, $S(T_n(\theta))$ and the estimated $1 - \alpha$ quantile will both be zero, so the probability of rejecting will be zero, so suppose that $P(m_j(W_i, \theta) < 0) > 0$ for some $j$. Then, for some $\eta > 0$, we have $nS(T_n(\theta)) > \eta$ with probability approaching one. From Lemma 9 (applied with $t$ less that $1 - \alpha$ and $\tau_b = b$), it follows that $L_{n,b}^{-1}(1 - \alpha | b^{\hat{\beta} \wedge \overline{\beta}}) = b^{\hat{\beta} \wedge \overline{\beta} - 1} L_{n,b}^{-1}(1 - \alpha | b) \geq b^{\hat{\beta} \wedge \overline{\beta} - 1} \eta / 2$ with probability approaching one. By Lemma 6, $S(T_n(\theta))$ will converge at a $n \log n$ rate, so that $n^{\hat{\beta} \wedge \overline{\beta}} S(T_n(\theta)) < n^{\hat{\beta} \wedge \overline{\beta} - 1} (\log n)^2$ with probability approaching one. Thus, we will fail to reject with probability approaching one as long as $n^{\hat{\beta} \wedge \overline{\beta} - 1} (\log n)^2 \leq b^{\hat{\beta} \wedge \overline{\beta} - 1} \eta / 2 = n^{\chi_3 (\hat{\beta} \wedge \overline{\beta} - 1)} \eta / 2$ for large enough $n$, and this holds since $\chi_3 < 1$. A similar argument holds for $\tilde{L}_{n,b}^{-1}(1 - \alpha | b^{\hat{\beta} \wedge \overline{\beta}})$.

□

## Testing Rate of Convergence Conditions: Estimating the Second Derivative

*proof of Lemma 1.* Let $h(x) = \bar{m}_j(\theta, x) - \min_{x' \in D} \bar{m}_j(\theta, x)$ where $\bar{m}_j(\theta, x) = E(m_j(W_i, \theta) | X_i = x)$ for a continuous version of the conditional mean function. First, note that $\mathcal{X}_0^j$ is compact. Since each $x \in \mathcal{X}_0^j$ is a local minimizer of $h(x)$ such that the second derivative matrix is strictly positive definite at $x$, there is an open set $A(x)$ containing each $x \in \mathcal{X}_0^j$ such that $h(x) > 0$ on $A(x) \backslash x$. The sets $A(x)$ with $x$ ranging over $\mathcal{X}_0^j$ form a covering of $\mathcal{X}_0^j$ with open sets. Thus, there is a finite subcover $A(x_1), \ldots A(x_\ell)$ of $\mathcal{X}_0^j$. Since the only elements in $A(x_1) \cup \cdots \cup A(x_\ell)$ that are also in $\mathcal{X}_0^j$ are $x_1, \ldots, x_\ell$, this means that $\mathcal{X}_0^j = \{x_1, \ldots, x_\ell\}$. □

*proof of Theorem 12.* By the next lemma, we will have $\mathcal{X}_0^j \subseteq \hat{\mathcal{X}}_0^j \subseteq \cup_{k=1}^{\hat{\ell}_j} B_{\varepsilon_n}(\hat{x}_{j,k})$ and $\mathcal{X}_0^j \subseteq \hat{\mathcal{X}}_0^j \subseteq \cup_{k \text{ s.t. } j \in J(k)} B_{\varepsilon_n}(x_k)$ with probability approaching one. When this holds, we will have $\hat{\ell} \leq |\{k|j \in J(k)\}|$ by construction and, once $\varepsilon_n$ is less than the smallest distance between any two points in $\mathcal{X}_0^j$, we will also have $\hat{\ell}_j = |\{k|j \in J(k)\}|$ and, for each $k$ from 1 to $\hat{\ell}_j$, we will have, for some function $r(j,k)$ such that $r(j,\cdot)$, is bijective from $\{1, \ldots, \hat{\ell}_j\}$ to $\{k|j \in J(k)\}$, $x_{r(j,k)} \in B_{\varepsilon_n}(\hat{x}_{j,k})$ for each $j, k$. When this holds, all of the $\hat{x}_{j,k}$s with $r(j,k)$ equal will be in the same equivalence class, since the corresponding $\varepsilon_n$ neighborhoods will intersect. When $\varepsilon_n$ is small enough that $\varepsilon_n$ neighborhoods containing $x_r$ and $\varepsilon_n$ neighborhoods containing $x_s$ do not intersect for $r \neq s$, there will be exactly $\ell$ equivalence classes, each one corresponding to the $(j,k)$ indices such that $r(j,k)$ is the same. Let the labeling of the $\tilde{x}_s$s be such that, for all $s$, $\tilde{x}_s = \hat{x}_{j,k}$ for some $(j,k)$ such that $r(j,k) = s$. Then, for each $s$, we have, for some $(j,k)$ such that $r(j,k) = s$, $x_s = x_{r(j,k)} \in B_{\varepsilon_n}(\hat{x}_{j,k}) = B_{\varepsilon_n}(\tilde{x}_s)$ with probability approaching one so that $\tilde{x}_s \xrightarrow{p} x_s$. To verify that $\hat{J}(s) = J(s)$ with probability approaching one, note that, for $j \in J(s)$, we will have $x_s \in \mathcal{X}_0^j \subseteq \cup_k B_{\varepsilon_n}(\hat{x}_{j,k})$ and $x_s \in B_{\varepsilon_n}(\tilde{x}_s)$ eventually, and, when this holds, $[\cup_k B_{\varepsilon_n}(\hat{x}_{j,k})] \cap B_{\varepsilon_n}(\tilde{x}_s) \neq \emptyset$ so that $j \in \hat{J}(s)$. For $j \notin J(s)$, each $\hat{x}_{j,k}$ will eventually be within $\varepsilon_n$ of some $x_r$ with $r \neq s$, while indices $(j',k')$ in the equivalence class associated with $s$ will eventually have $\hat{x}_{j',k'}$ within $2\varepsilon$ of $x_s$, so that $(j,k)$ will not be in the equivalence class associated with $s$ for any $k$, and $j \notin \hat{J}(s)$. □

**Lemma 12.** *Suppose that* $\sup_{x \in D} \|\hat{\bar{m}}_j(\theta,x) - \bar{m}_j(\theta,x)\| = \mathcal{O}(a_n)$ *for some sequence* $a_n \to 0$. *Then, under Assumption 1, for any sequence* $b_n \to \infty$ *with* $b_n a_n \to 0$ *and* $\varepsilon_n$ *with* $\varepsilon_n \to 0$ *more slowly than* $\sqrt{b_n a_n}$, *the set* $\hat{\mathcal{X}}_0^j \equiv \{x|\hat{\bar{m}}_j(\theta,x) \leq b_n a_n\}$ *satisfies*

$$\mathcal{X}_0^j \subseteq \hat{\mathcal{X}}_0^j \subseteq \cup_{k \text{ s.t. } j \in J(k)} B_{\varepsilon_n}(x_k)$$

*Proof.* We will have $\mathcal{X}_0^j \subseteq \hat{\mathcal{X}}_0^j$ as soon as $\sup_{x \in D} \|\hat{\bar{m}}_j(\theta,x) - \bar{m}_j(\theta,x)\| \leq b_n a_n$, which happens with probability approaching one. To show that $\hat{\mathcal{X}}_0^j \subseteq \cup_{k \text{ s.t. } j \in J(k)} B_{\varepsilon_n}(x_k)$ eventually, suppose that, for some $\hat{x} \in \hat{\mathcal{X}}_0^j$, $\hat{x} \notin B_{\varepsilon_n}(x_k)$ for any $k$. Let $C$ and $\eta$ be such that $\bar{m}_j(\theta,x) \geq C \min_k \|x - x_k\|^2$ when $\|x - x_k\| \leq \eta$ for some $k$ (such a $C$ and $\eta$ exist by Assumption 1). Then, for any $\hat{x}$ such that $\hat{\bar{m}}_j(\theta,\hat{x}) \leq b_n a_n$, we must have, with probability approaching one,

$$C \min_k \|x - x_k\|^2 \leq \bar{m}_j(\theta,\hat{x}) \leq b_n a_n + \bar{m}_j(\theta,\hat{x}) - \hat{\bar{m}}_j(\theta,\hat{x}) \leq 2 b_n a_n$$

where the first inequality follows since $\hat{\mathcal{X}}_0^j$ is contained in $\{x|\|x-x_k\| \leq \eta$ some $k$ s.t. $j \in J(k)\}$ eventually. Since $\varepsilon_n \geq \sqrt{2b_n a_n/C}$ eventually, the first claim follows. □

## Local Alternatives

*proof of Theorem 13.* Everything is the same as in the proof of Theorem 1, but with the following modifications.

First, in the proof of Theorem 15, we need to show that, for all $j$,

$$\frac{\sqrt{n}}{\sqrt{h_n^d}}(E_n - E)[m_j(W_i, \theta_0 + a_n) - m_j(W_i, \theta_0)]I(h_n s < X_i - x_k < h_n(s+t))$$

converges to zero uniformly over $\|(s,t)\| < M$ for any fixed $M$. By Theorem 2.14.1 in van der Vaart and Wellner (1996), the $L^2$ norm of this is bounded up to a constant by $J(1, \mathcal{F}_n, L_2)\frac{1}{h_n^d}\sqrt{EF_n(X_i, W_i)^2}$, where $\mathcal{F}_n = \{(x,w) \mapsto [m_j(w, \theta_0 + a_n) - m_j(w, \theta_0)]I(h_n s < x - x_k < h_n(s+t))|(s,t) \in \mathbb{R}^{2d}\}$ and $F_n(x,w) = |m_j(w, \theta_0 + a_n) - m_j(w, \theta_0)|I(-h_n M\iota < x - x_k < 2h_n M\iota)$ is an envelope function for this class (here $\iota$ is a vector of ones). The covering numbers of the $\mathcal{F}_n$s are uniformly bounded by a polynomial, so that we just need to show that $\frac{1}{h_n^d}\sqrt{EF_n(X_i, W_i)^2}$ converges to zero. We have

$$\begin{aligned}
&\frac{1}{\sqrt{h_n^d}}\sqrt{EF_n(X_i, W_i)^2} \\
&= \frac{1}{\sqrt{h_n^d}}\sqrt{EE\{[m_j(W_i, \theta_0 + a_n) - m_j(W_i, \theta_0)]^2|X_i\}I(-h_n M\iota < X_i - x_k < 2h_n M\iota)} \\
&\leq \frac{1}{\sqrt{h_n^d}}\sqrt{EI(-h_n M\iota < X_i - x_k < 2h_n M\iota)} \sup_{\|x-x_k\| \leq \eta} E\{[m_j(W_i, \theta_0 + a_n) - m_j(W_i, \theta_0)]^2|X_i = x\}
\end{aligned}$$

where the first equality uses the law of iterated expectations and the second holds eventually with $\eta$ chosen so that the convergence in Assumption 13 is uniform over $\|x - x_k\| < \eta$. The first term is bounded eventually by $\overline{f}\int_{-M\iota<x<2M\iota} dx$ where $\overline{f}$ is a bound for the density of $X_i$ in a neighborhood of $x_k$ (this follows from the same change of variables as in other parts of the proof). The second term converges to zero by Assumption 13.

Next, in the proof of Theorem 15, we need to show that

$$\frac{1}{h_n^{d+2}}E[\bar{m}_j(\theta_0 + a_n, X_i) - \bar{m}_j(\theta_0, X_i)]I(h_n s < X_i - x_k < h_n(s+t)) \to f_X(x_k)\bar{m}_{\theta,j}(\theta_0, x_k)a\prod_i t_i$$

uniformly in $\|(s,t)\| \leq M$. We have

$$\frac{1}{h_n^{d+2}} E[\bar{m}_j(\theta_0 + a_n, X_i) - \bar{m}_j(\theta_0, X_i)] I(h_n s < X_i - x_k < h_n(s+t)) - f_X(x_k)\bar{m}_{\theta,j}(\theta_0, x_k) a \prod_i t_i$$
$$= \frac{1}{h_n^{d+2}} \int_{h_n s < x - x_k < h_n(s+t)} \left\{ [\bar{m}_j(\theta_0 + a_n, x) - \bar{m}_j(\theta_0, x)] f_X(x) - h_n^2 f_X(x_k) \bar{m}_{\theta,j}(\theta_0, x_k) a \right\} dx$$
$$= \int_{s<x<s+t} \left\{ h_n^{-2} [\bar{m}_j(\theta_0 + a_n, h_n x + x_k) - \bar{m}_j(\theta_0, h_n x + x_k)] f_X(h_n x + x_k) - f_X(x_k) \bar{m}_{\theta,j}(\theta_0, x_k) a \right\} dx$$

where the second equality comes from the change of variable $x \mapsto h_n x + x_k$. This will go to zero uniformly in $\|(s,t)\| \leq M$ as long as $\sup_{\|x\| \leq 2M} \|f_X(h_n x + x_k) - f_X(x_k)\|$ and

$$\sup_{\|x\| \leq 2M} \|h_n^{-2}[\bar{m}_j(\theta_0 + a_n, h_n x + x_k) - \bar{m}_j(\theta_0, h_n x + x_k)] - \bar{m}_{\theta,j}(\theta_0, x_k) a\|$$

both go to zero. $\sup_{\|x\| \leq 2M} \|f_X(h_n x + x_k) - f_X(x_k)\|$ goes to zero by continuity of $f_X$ at $x_k$. As for the other expression, since $a h_n^2 = a_n$, the mean value theorem shows that this is equal to $\bar{m}_{\theta,j}(\theta^*(a_n), h_n x + x_k) a - \bar{m}_{\theta,j}(\theta_0, x_k) a$ for some $\theta^*(a_n)$ between $\theta_0$ and $\theta_0 + a_n$. This goes to zero by Assumption 12.

In verifying the conditions of Lemma 2, we need to make sure the bounds, $g_{P,x_k,j,a}(s,t) \geq C\|(s,t)\|^2 \prod_i t_i$ and

$$g_{n,x_k,j,a}(s,t) \equiv \frac{1}{h_n^{d+2}} E m_j(W_i, \theta_0 + a_n) I(h_n s < X_i < h_n(s+t)) \geq C\|(s,t)\|^2 \prod_i t_i$$

still hold for $\|(s,t)\| \geq M$ for $M$ large enough and, for the latter function, $\|(s,t)\| \leq h_n^{-1}\eta$ for some $\eta > 0$ and $n$ greater than some $N$ that does not depend on $M$. We have

$$g_{P,x_k,j,a}(s,t) = g_{P,x_k,j}(s,t) + \bar{m}_{\theta,j}(\theta_0, x_k) a f_X(x_k) \prod_i t_i \geq C\|(s,t)\|^2 \prod_i t_i + \bar{m}_{\theta,j}(\theta_0, x_k) a f_X(x_k) \prod_i t_i$$
$$= \|(s,t)\|^2 [C + \bar{m}_{\theta,j}(\theta_0, x_k) a f_X(x_k) / \|(s,t)\|^2] \prod_i t_i$$

where the first inequality follows from the bound in the original proof. For $\|(s,t)\| \geq M$ for $M$ large enough, this is greater than or equal to $K\|(s,t)\|^2 \prod_i t_i$ for $K = C -$

$|\bar{m}_{\theta,j}(\theta_0,x_k)a|f_X(x_k)/M^2 > 0$. For $g_{n,x_k,j,a}(s,t)$, we have

$$\|g_{P,x_k,j,a}(s,t) - g_{P,x_k,j}(s,t)\| = \|\frac{1}{h_n^{d+2}}E[m_j(W_i,\theta_0+a_n) - m_j(W_i,\theta_0)]I(h_n s < X_i < h_n(s+t))\|$$
$$\leq \sup_{\|x-x_k\|\leq\eta} \|\frac{1}{h_n^2}[\bar{m}_j(\theta_0+a_n,x) - \bar{m}_j(\theta_0,x)]\|\|\frac{1}{h_n^d}EI(h_n s < X_i < h_n(s+t))\|.$$

By the mean value theorem, $\bar{m}_j(\theta_0+a_n,x) - \bar{m}_j(\theta_0,x) = \bar{m}_{j,\theta}(\theta^*(a_n),x)a_n$ for some $\theta^*(a_n)$ between $\theta_0$ and $\theta_0+a_n$. By continuity of the derivative as a function of $(\theta,x)$, for small enough $\eta$ and $n$ large enough, $\bar{m}_{j,\theta}(\theta^*(a_n),x)$ is bounded from above, so that $\|\frac{1}{h_n^2}[\bar{m}_j(\theta_0+a_n,x)-\bar{m}_j(\theta_0,x)]\|$ is bounded by a constant times $\|a_n\|/h_n^2 = \|a\|$. By continuity of $f_X$ at $x_k$, $\|\frac{1}{h_n^d}EI(h_n s < X_i < h_n(s+t))\|$ is bounded by some constant times $\prod_i t_i$ for $\|(s,t)\| \leq h_n^{-1}\eta$. Thus, for $M \leq \|(s,t)\| \leq h_n^{-1}\eta$ for the appropriate $M$ and $\eta$, we have, for some constant $C_1$,

$$g_{P,x_k,j,a}(s,t) \geq g_{P,x_k,j}(s,t) - C_1\prod_i t_i \geq C\|(s,t)\|^2\prod_i t_i - C_1\prod_i t_i$$
$$= \|(s,t)\|^2[C - C_1/\|(s,t)\|^2]\prod_i t_i$$

where the second inequality uses the bound from the original proof. For $M$ large enough, this gives the desired bound with the constant equal to $C - C_1/M > 0$.

In verifying the conditions of Lemma 2, we also need to make sure the argument in Lemma 4 still goes through when $m(W_i,\theta_0)$ is replaced by $m(W_i,\theta_0+a_n)$. To get the lemma to hold (with the constant $C$ depending only on the distribution of $X$ and the $\overline{Y}$ in Assumption 14), we can use the same proof, but with the classes of functions $\mathcal{F}_n$ defined to be $\mathcal{F}_n = \{(x,w) \mapsto m_j(w,\theta_0+a_n)I(h_n s_0 < x - x_k < h_n(s_0+t))|t \leq t_0\}$ ($J(1,\mathcal{F}_n,L^2)$ is bounded uniformly for these classes because the covering number of each $\mathcal{F}_n$ is bounded by the same polynomial), and using the envelope function $F_n(x,w) = \overline{Y}I(h_n s_0 < x - x_k < h_n(s_0+t_0))$ when applying Theorem 2.14.1 in van der Vaart and Wellner (1996).

□

*proof of Theorem 14.* First, note that, for any neighborhoods $B(x_k)$ of the elements of $\mathcal{X}_0$, $\sqrt{n}\inf_{s,t} E_n m_j(W_i,\theta_0+a_n)I(s < X < s+t) = \sqrt{n}\inf_{(s,s+t)\in\cup_{k \text{ s.t. } j\in J(k)}B(x_k)} E_n m_j(W_i,\theta_0+a_n)I(s < X_i < s+t) + o_p(1)$ since, if these neighborhoods are made small enough, we will have, for any $(s,s+t)$ not in one of these neighborhoods, $Em_j(W_i,\theta_0+a_n)I(s < X_i < s+t) \geq \underline{B}P(s < X_i < s+t)$ by an argument similar to the one in Lemma 5, so that an argument similar to the one in Lemma 6 will show that $\inf_{(s,s+t)\in\cup_{k \text{ s.t. } j\notin J(k)}B(x_k)} E_n m_j(W_i,\theta_0+a_n)I(s <$

$X_i < s+t)$ converges to zero at a faster than $\sqrt{n}$ rate (Assumption 12 guarantees that $E[m_j(W_i,\theta_0+a_n)|X]$ is eventually bounded away from zero outside of any neighborhood of $\mathcal{X}_0$ so that a similar argument applies).

Thus, the result will follow once we show that, for each $j$ and $k$ such that $j \in J(k)$,

$$\sqrt{n} \inf_{(s,s+t)\in B(x_k)} E_n m_j(W_i,\theta_0+a_n)I(s<X_i<s+t)$$
$$\xrightarrow{p} \inf_{s,t} f_X(x_k)\int_{s<x<s+t}\left(\frac{1}{2}x'Vx+\overline{m}_{\theta,j}(\theta_0,x_k)a\right)dx.$$

With this in mind, fix $j$ and $k$ with $j\in J(k)$.

Let $(s_n^*,t_n^*)$ minimize $E_n m_j(W_i,\theta_0+a_n)I(s<X<s+t)$ over $B(x_k)^2$ (and be chosen from the set of minimizers in a measurable way). First, I show that $\rho(0,(s_n^*,t_n^*))\xrightarrow{p}0$ where $\rho$ is the covariance semimetric $\rho((s,t),(s',t'))=var(m_j(W_i,\theta_0)I(s<x<s+t)-m_j(W_i,\theta_0)I(s'<x<s'+t'))$. To show this, note that, for any $\varepsilon>0$, $Em_j(W_i,\theta_0+a_n)I(s<X_i<s+t)$ is bounded from below away from zero for $\rho(0,(s,t))\geq\varepsilon$ for large enough $n$. To see this, note that, for $\rho(0,(s,t))\geq\varepsilon$, $\prod_i t_i\geq K$ for some constant $K$, so that $\|(s,t)\|\geq K^{1/d}$ and, for some constant $C$ and a bound $\overline{f}$ for $f_X$ on $B(x_k)$,

$$\begin{aligned}
&Em_j(W_i,\theta_0+a_n)I(s<X_i<s+t)\\
&=Em_j(W_i,\theta_0)I(s<X_i<s+t)+E[\bar{m}_j(\theta_0+a_n,X_i)-\bar{m}_j(\theta_0,X_i)]I(s<X_i<s+t)\\
&\geq C_1\|(s,t)\|^2\left(\prod_i t_i\right)-\sup_{x\in B(x_k)}\|\bar{m}_j(\theta_0+a_n,x)-\bar{m}_j(\theta_0,x)\|\overline{f}\left(\prod_i t_i\right)\\
&\geq\left[C_1\|(s,t)\|^2-\sup_{x\in B(x_k)}\|\bar{m}_j(\theta_0+a_n,x)-\bar{m}_j(\theta_0,x)\|\overline{f}\right]K.
\end{aligned}$$

By Assumption 13, $\sup_{x\in B(x_k)}\|\bar{m}_j(\theta_0+a_n,x)-\bar{m}_j(\theta_0,x)\|$ converges to zero, so the last term in this display will be positive and bounded away from zero for large enough $n$. Thus, we can write $\sqrt{n}E_n m_j(W_i,\theta_0+a_n)I(s<X_i<s+t)$ as the sum of $\sqrt{n}(E_n-E)m_j(W_i,\theta_0+a_n)I(s<X_i<s+t)$, which is $\mathcal{O}_p(1)$ uniformly in $(s,t)$, and $\sqrt{n}Em_j(W_i,\theta_0+a_n)I(s<X<s+t)$, which is bounded from below uniformly in $\rho(0,(s,t))\geq\varepsilon$ by a sequence of constants that go to infinity. Thus, $\inf_{\rho(0,(s,t))\geq\varepsilon}\sqrt{n}E_n m_j(W_i,\theta_0+a_n)I(s<X<s+t)$ is greater than zero with probability approaching one, so $\rho(0,(s^*,t^*))\xrightarrow{p}0$.

Thus, for some sequence of random variables $\varepsilon_n \xrightarrow{p} 0$,

$$\sqrt{n}\inf_{s,t} E_n m_j(W_i,\theta_0+a_n)I(s<X<s+t)$$
$$=\sqrt{n}\inf_{\rho(0,(s^*,t^*))\leq\varepsilon_n,(s,s+t)\in B(x_k)} E_n m_j(W_i,\theta_0+a_n)I(s<X<s+t).$$

This is equal to $\sqrt{n}\inf_{\rho(0,(s^*,t^*))\leq\varepsilon_n,(s,s+t)\in B(x_k)} Em_j(W_i,\theta_0+a_n)I(s<X<s+t)$ plus a term that is bounded by $\sqrt{n}\sup_{\rho(0,(s^*,t^*))\leq\varepsilon_n,(s,s+t)\in B(x_k)}|(E_n-E)E_n m_j(W_i,\theta_0+a_n)I(s< X<s+t)|$. By Assumption 13 and an argument using the maximal inequality in Theorem 2.14.1 in van der Vaart and Wellner (1996), $\sqrt{n}\sup_{(s,s+t)\in B(x_k)}|(E_n-E)[m_j(W_i,\theta_0+a_n)-m_j(W_i,\theta_0)]I(s<X_i<s+t)|$ converges in probability to zero. $\sqrt{n}(E_n-E)m_j(W_i,\theta_0)I(s< X_i<s+t)$ converges in distribution under the supremum norm to a mean zero Gaussian process $\mathbb{H}(s,t)$ with covariance kernel $cov(\mathbb{H}(s,t),\mathbb{H}(s',t'))=cov(m_j(W_i,\theta_0)I(s< X_i<s+t),m_j(W_i,\theta_0)I(s'<X_i<s'+t'))$ and almost sure $\rho$ continuous sample paths. Since $(z,\varepsilon)\mapsto\sup_{\rho(0,(s,t))\leq\varepsilon}|z(s,t)|$ is continuous in $C(\mathbb{R}^{2d_X},\rho)\times\mathbb{R}$ (where $C(\mathbb{R}^{2d_X},\rho)$ is the space of $\rho$ continuous functions on $\mathbb{R}^{2d}$) under the product norm of the supremum norm and the Euclidean norm, by the continuous mapping theorem, $\sup_{\rho(0,(s,t))\leq\varepsilon_n}|\sqrt{n}(E_n-E)m_j(W_i,\theta_0)I(s<X_i<s+t)|\xrightarrow{d}\sup_{\rho(0,(s,t))\leq 0}\mathbb{H}(s,t)=0$ (the last step follows since $var(\mathbb{H}(s,t))=0$ whenever $\rho(0,(s,t))=0$).

Thus,

$$\sqrt{n}\inf_{(s,s+t)\in B(x_k)} E_n m_j(W_i,\theta_0+a_n)I(s<X_i<s+t)$$
$$=\sqrt{n}\inf_{\rho(0,(s,t))<\varepsilon_n,(s,s+t)\in B(x_k)} Em_j(W_i,\theta_0+a_n)I(s<X_i<s+t)+o_p(1)$$
$$=\sqrt{n}\inf_{\rho(0,(s,t))<\varepsilon_n,(s,s+t)\in B(x_k)}\int_{s<x<s+t}\bar{m}_j(\theta_0+a_n,x)f_X(x)\,dx+o_p(1).$$

By Assumption 12, the integrand is positive eventually for $\|(s-x_k,t)\|\geq\eta$ for any $\eta>0$, and once this holds, the infimum will be achieved on $\|(s-x_k,t)\|<\eta$. Using a first order Taylor expansion in the first argument of $\bar{m}_j(\theta_0+a_n,x)$ and a second order Taylor expansion in the second argument the integrand is equal to

$$\left[\frac{1}{2}(x-x_k)V(x^*(x))(x-x_k)'+\bar{m}_{\theta,j}(\theta^*(a_n),x)a_n\right]f_X(x)$$

for some $x^*(x)$ between $x$ and $x_k$ and $\theta^*(a_n)$ between $\theta_0$ and $\theta_0+a_n$. For $\eta$ small enough, continuity of the derivatives at $(\theta_0,x_k)$ guarantees that this is bounded from below by $C_1\|x-$

$x_k\|^2 - C_2 a_n$ for some constants $C_1$ and $C_2$, so the integrand is positive for $x$ greater than $C\sqrt{\|a_n\|}$ for some large $C$, so that the infimum will be taken on $\|(s, s+t)\| < C\sqrt{\|a_n\|}$. Thus, we have

$$\begin{aligned}
&\sqrt{n} \inf_{(s,s+t)\in B(x_k)} E_n m_j(W_i, \theta_0 + a_n) I(s < X_i < s+t) \\
&= \sqrt{n} \inf_{\rho(0,(s,t))<\varepsilon_n, \|(s-x_k,t)\|<C\sqrt{\|a_n\|}} \int_{s<x<s+t} \bar{m}_j(\theta_0 + a_n, x) f_X(x)\, dx + o_p(1).
\end{aligned}$$

This will be equal up to $o(1)$ to the infimum of

$$\sqrt{n} \int_{s<x<s+t} \left[\frac{1}{2}(x - x_k)V_j(x_k)(x - x_k)' + \bar{m}_{\theta,j}(\theta_0, x_k)a_n\right] f_X(x_k)\, dx$$

once we show that the difference between this expression and $\sqrt{n}\int_{s<x<s+t} \bar{m}_j(\theta_0 + a_n, x) f_X(x)\, dx$ goes to zero uniformly over $\|(s - x_k, t)\| \le C\sqrt{\|a_n\|}$ (the infimum of this last display will be taken at a sequence where $\|(s - x_k, t)\| \le C\sqrt{\|a_n\|}$ anyway, so that the infimum can be taken over all of $\mathbb{R}^{2d}$).

The difference between these terms is

$$\begin{aligned}
&\sqrt{n} \int_{s<x<s+t} \left[\frac{1}{2}(x - x_k)V_j(x_k)(x - x_k)' + \bar{m}_{\theta,j}(\theta_0, x_k)a_n\right] [f_X(x) - f_X(x_k)]\, dx \\
&+ \sqrt{n} \int_{s<x<s+t} \frac{1}{2}\left[(x - x_k)V_j(x^*(x))(x - x_k)' - (x - x_k)V_j(x_k)(x - x_k)'\right] f_X(x)\, dx \\
&+ \sqrt{n} \int_{s<x<s+t} \left[\bar{m}_{\theta,j}(\theta^*(a_n), x) - \bar{m}_{\theta,j}(\theta_0, x_k)\right] a_n f_X(x)\, dx.
\end{aligned}$$

These can all be bounded using the change of variables $u = (x - x_k)n^{1/(2(d+2))}$ and the continuity of densities, conditional means, and their derivatives. The first term is

$$\begin{aligned}
&\sqrt{n} \int_{n^{1/(2(d+2))}(s-x_k)<u<(s+t-x_k)n^{1/(2(d+2))}} \left[\frac{1}{2}uV_j(x_k)u'n^{-1/(d+2)} + \bar{m}_{\theta,j}(\theta_0, x_k)an^{-1/(d+2)}\right] \\
&\times [f_X(n^{-1/(2(d+2))}u + x_k) - f_X(x_k)]n^{-d/(2(d+2))}\, du \\
&= \int_{n^{1/(2(d+2))}(s-x_k)<u<(s+t-x_k)n^{1/(2(d+2))}} \left[\frac{1}{2}uV_j(x_k)u' + \bar{m}_{\theta,j}(\theta_0, x_k)a\right] \\
&\times [f_X(n^{-1/(2(d+2))}u + x_k) - f_X(x_k)]\, du.
\end{aligned}$$

The integrand converges to zero uniformly over $u$ in any bounded set by the continuity of $f_X$ at $x_k$, and the area of integration is bounded by $\|u\| \leq 2n^{1/(2(d+2))}\|(s-x_k,t)\| \leq 2Cn^{1/(2(d+2))}\sqrt{\|a\|}n^{-1/(2(d+2))} = 2C\sqrt{\|a\|}$ on $\|(s-x_k,t)\| \leq C\sqrt{\|a_n\|}$. Using the same change of variables, the second term is bounded by the integral of

$$\frac{1}{2}\left[u'V_j(x^*(n^{-1/(2(d+2))}u+x_k))u' - uV_j(x_k)u'\right] f_X(n^{-1/(2(d+2))}u+x_k)$$

over a bounded region, and this converges to zero uniformly in any bounded region by continuity of the second derivative matrix. The last term is, by the same change of variables, bounded by the integral of

$$\left[\bar{m}_{\theta,j}(\theta^*(a_n), n^{-1/(2(d+2))}u+x_k) - \bar{m}_{\theta,j}(\theta_0,x_k)\right] af_X(n^{-1/(2(d+2))}u+x_k)$$

over a bounded region, and this converges to zero by continuity of $m_{\theta,j}(\theta,x)$ at $(\theta_0,x_k)$.

Thus,

$$\begin{aligned}
&\sqrt{n} \inf_{(s,s+t)\in B(x_k)} E_n m_j(W_i,\theta_0+a_n)I(s<X_i<s+t) \\
&= \inf_{\|(s-k,t)\|\leq C\sqrt{\|a_n\|}} \sqrt{n}\int_{s<x<s+t}\left[\frac{1}{2}(x-x_k)V_j(x_k)(x-x_k)' + \bar{m}_{\theta,j}(\theta_0,x_k)a_n\right] f_X(x_k)\,dx + o_p(1) \\
&= \inf_{\|(s-x_k,t)\|\leq C\sqrt{\|a\|}} \int_{(s-x_k)<u<(s-x_k+t)}\left[\frac{1}{2}uV_j(x_k)u' + \bar{m}_{\theta,j}(\theta_0,x_k)a_n\right] f_X(x_k)\,du + o_p(1)
\end{aligned}$$

where the last equality follows from the same change of variables and a change of coordinates in $(s,t)$. The result follows since, for large enough $C$, the unconstrained infimum is taken on $\|(s-x_k,t)\| \leq C\sqrt{\|a\|}$, and $C$ can be chosen arbitrarily large. □

# References

Adler, R. J. (1990): "An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes," *Lecture Notes-Monograph Series*, 12, i–155.

Andrews, D. W., S. Berry, and P. Jia (2004): "Confidence regions for parameters in discrete games with multiple equilibria, with an application to discount chain store location," .

Andrews, D. W., and P. Guggenberger (2009): "Validity of Subsampling and ?plugin Asymptotic? Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25(03), 669–709.

Andrews, D. W., and X. Shi (2009): "Inference Based on Conditional Moment Inequalities," *Unpublished Manuscript, Yale University, New Haven, CT.*

Andrews, D. W. K., and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *SSRN eLibrary.*

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78(1), 119–157.

Armstrong, T. (2011): "Weighted KS Statistics for Inference on Conditional Moment Inequalities," *Unpublished Manuscript.*

Beresteanu, A., and F. Molinari (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76(4), 763–814.

Bierens, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20(1), 105–134.

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

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2009): "Intersection bounds: estimation and inference," *Arxiv preprint arXiv:0907.3503.*

Chetty, R. (2010): "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply," *NBER Working Paper.*

Ciliberto, F., and E. Tamer (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77(6), 17911828.

Davydov, Y. A., M. A. Lifshits, and N. V. Smorodina (1998): *Local Properties of Distributions of Stochastic Functionals.* American Mathematical Society.

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.

Ichimura, H., and P. E. Todd (2007): "Chapter 74 Implementing Nonparametric and Semiparametric Estimators," vol. Volume 6, Part 2, pp. 5369–5468. Elsevier.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

Khan, S., and E. Tamer (2009): "Inference on endogenously censored regression models using conditional moment inequalities," *Journal of Econometrics*, 152(2), 104–119.

Kim, J., and D. Pollard (1990): "Cube Root Asymptotics," *The Annals of Statistics*, 18(1), 191219.

Kim, K. i. (2008): "Set estimation and inference with models characterized by conditional moment inequalities," .

Koenker, R., and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46(1), 33–50, ArticleType: research-article / Full publication date: Jan., 1978 / Copyright  1978 The Econometric Society.

——— (1982): "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50(1), 43–61, ArticleType: research-article / Full publication date: Jan., 1982 / Copyright  1982 The Econometric Society.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference.*

Lee, S., K. Song, and Y. Whang (2011): "Testing functional inequalities," .

Lehmann, E. L., and J. P. Romano (2005): *Testing statistical hypotheses*. Springer.

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80(2), 319–323.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

Menzel, K. (2008): "Estimation and Inference with Many Moment Inequalities," *Preprint, Massachussetts Institute of Technology.*

Moon, H. R., and F. Schorfheide (2009): "Bayesian and Frequentist Inference in Partially Identified Models," *National Bureau of Economic Research Working Paper Series*, No. 14882.

Pagan, A., and A. Ullah (1999): *Nonparametric econometrics.* Cambridge University Press.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2006): "Moment Inequalities and Their Application," .

Pitt, L. D., and L. T. Tran (1979): "Local Sample Path Properties of Gaussian Fields," *The Annals of Probability*, 7(3), 477–493.

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

Pollard, D. (1984): *Convergence of stochastic processes.* David Pollard.

Ponomareva, M. (2010): "Inference in Models Defined by Conditional Moment Inequalities with Continuous Covariates," .

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.

Romano, J. P., and A. M. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169–211.

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

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

Wright, J. H. (2003): "Detecting Lack of Identification in Gmm," *Econometric Theory*, 19(02), 322–330.

Figure 1: Case with faster than root-$n$ convergence of KS statistic

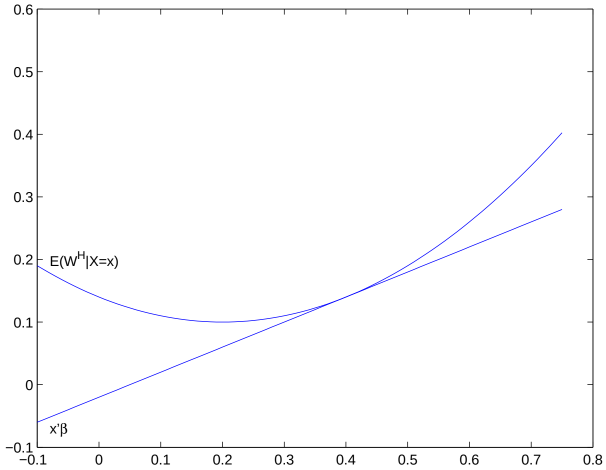


Figure 2: Cases with root-$n$ convergence of KS statistic ($\beta_1$) and faster rates ($\beta_2$)

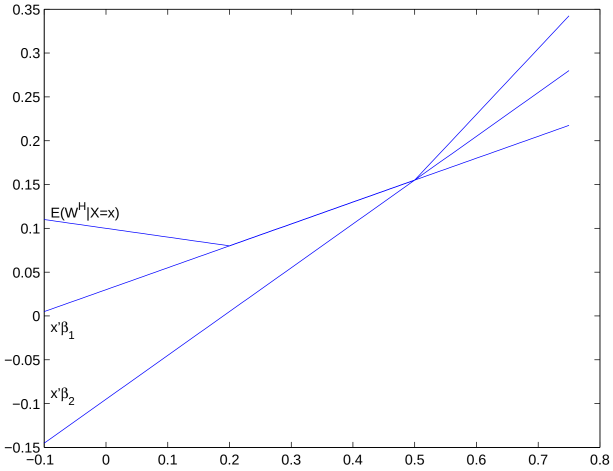


Figure 3: Conditional Means of $W_i^H$ and $W_i^L$ for Design 1

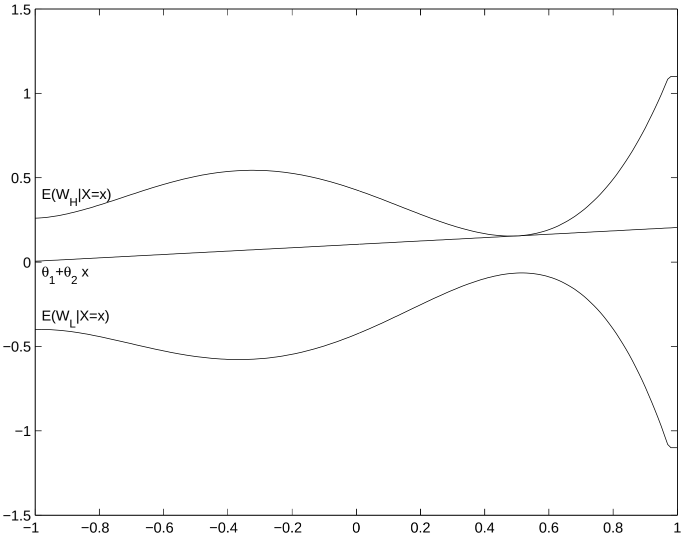


Figure 4: Conditional Means of $W_i^H$ and $W_i^L$ for Design 2

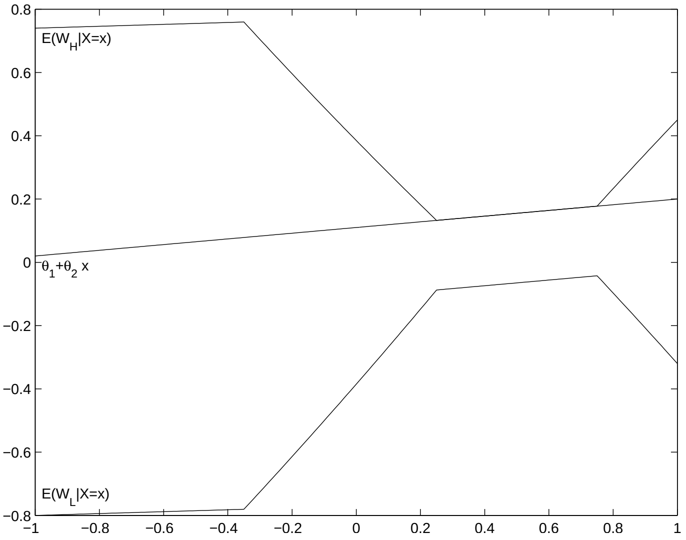


Figure 5: Histograms for $n^{3/5}S(T_n(\theta))$ for Design 1 ($n^{3/5}$ Convergence)

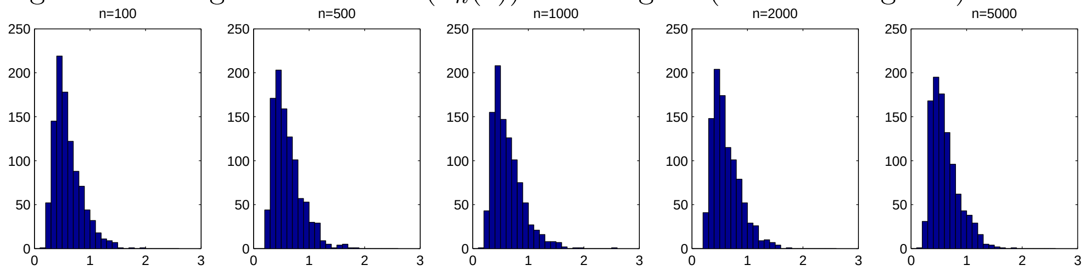


Figure 6: Histograms for $n^{1/2}S(T_n(\theta))$ for Design 2 ($n^{1/2}$ Convergence)

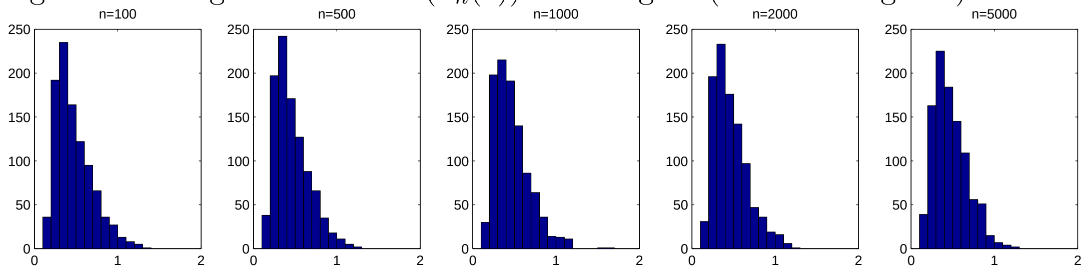


Figure 7: Data for Empirical Illustration

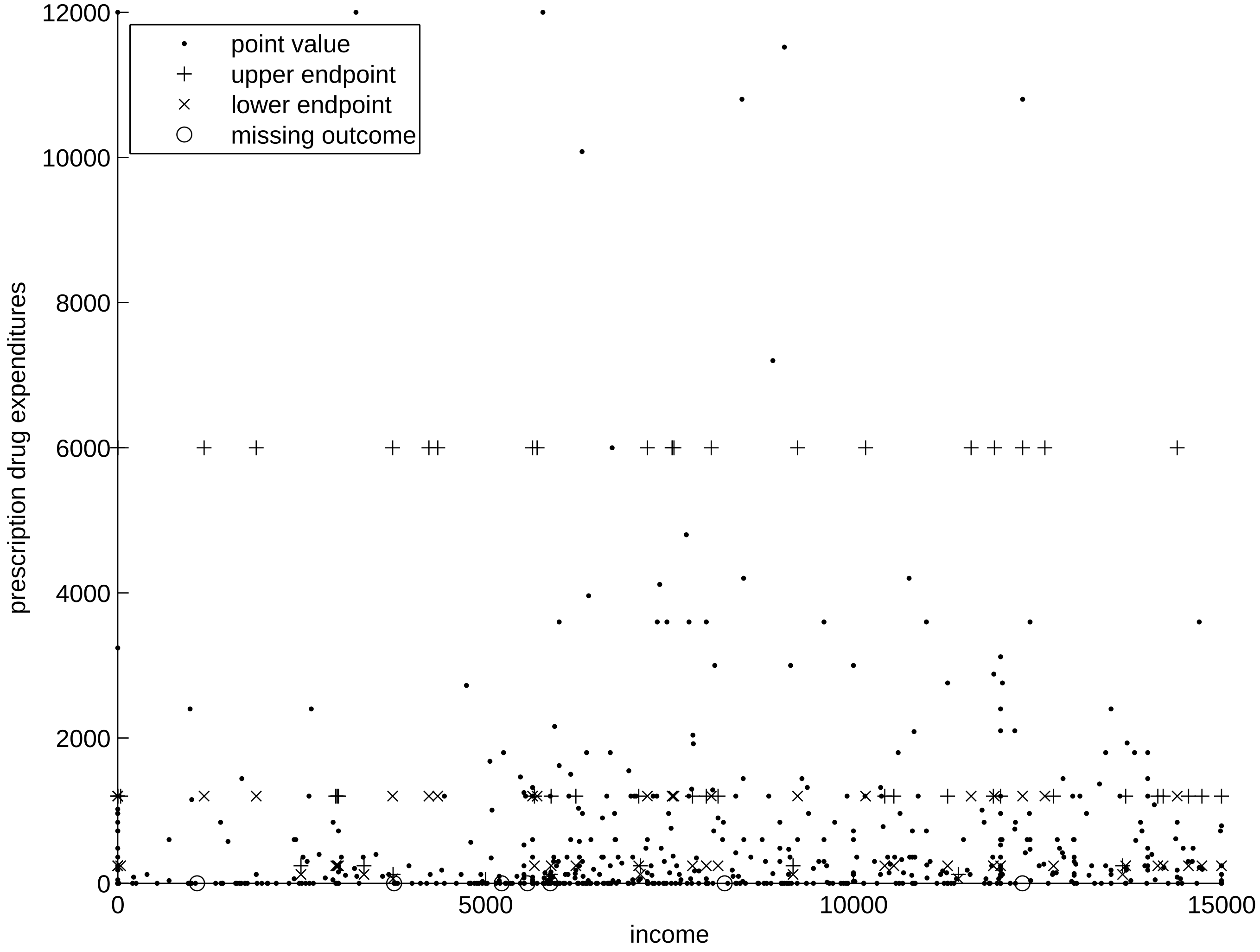


Figure 8: 95% Confidence Region Using Estimated Rate

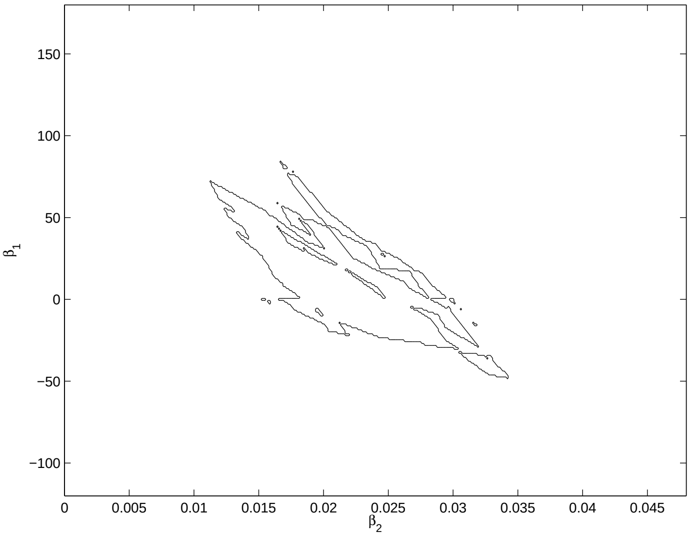


Figure 9: 95% Confidence Region Using Conservative Rate

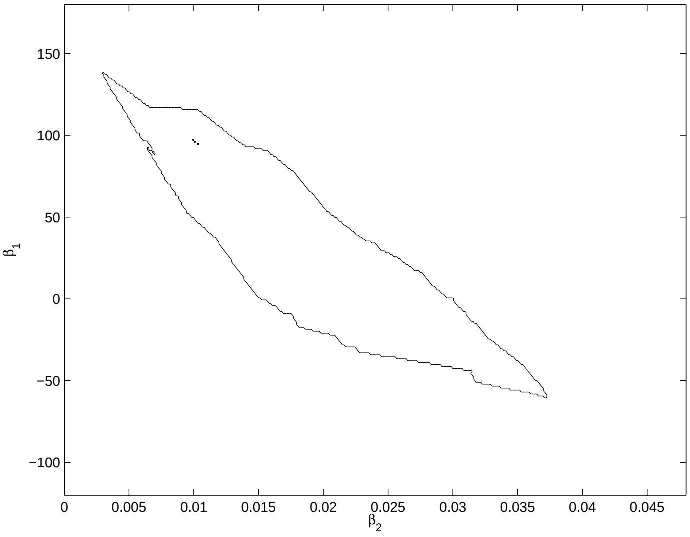
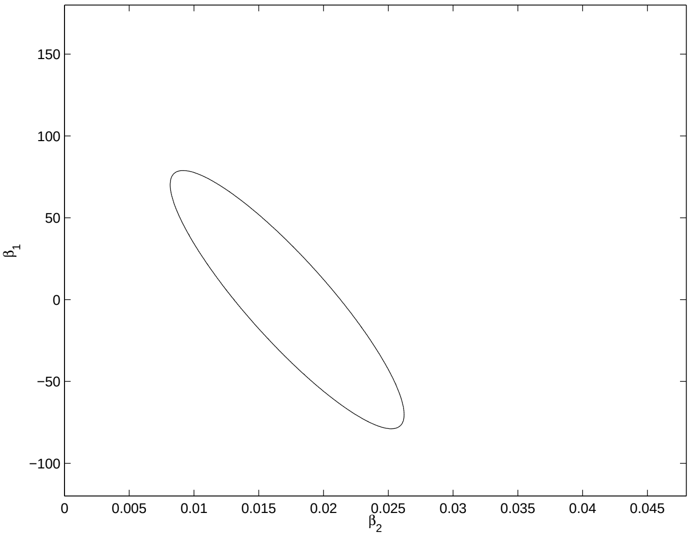


Figure 10: 95% Confidence Region Using LAD with Points



|  | $n = 100$ | $n = 500$ | $n = 1000$ | $n = 2000$ | $n = 5000$ |
|---|---|---|---|---|---|
|  | nominal 90% coverage |  |  |  |  |
| estimated rate | 0.873 | 0.890 | 0.897 | 0.889 | 0.879 |
| conservative rate ($n^{1/2}$) | 0.991 | 0.987 | 0.987 | 0.995 | 0.996 |
| (infeasible) exact rate ($n^{3/5}$) | 0.921 | 0.909 | 0.905 | 0.903 | 0.890 |
|  | nominal 95% coverage |  |  |  |  |
| estimated rate | 0.940 | 0.943 | 0.954 | 0.947 | 0.934 |
| conservative rate ($n^{1/2}$) | 0.998 | 1.000 | 0.998 | 1.000 | 0.999 |
| (infeasible) exact rate ($n^{3/5}$) | 0.976 | 0.965 | 0.949 | 0.956 | 0.953 |

Table 1: Coverage Probabilities for Design 1

|  | $n = 100$ | $n = 500$ | $n = 1000$ | $n = 2000$ | $n = 5000$ |
|---|---|---|---|---|---|
|  | nominal 90% coverage |  |  |  |  |
| estimated rate | 0.780 | 0.910 | 0.928 | 0.925 | 0.924 |
| conservative rate ($n^{1/2}$) | 0.949 | 0.947 | 0.938 | 0.932 | 0.924 |
| (infeasible) exact rate ($n^{1/2}$) | 0.949 | 0.947 | 0.938 | 0.932 | 0.924 |
|  | nominal 95% coverage |  |  |  |  |
| estimated rate | 0.885 | 0.945 | 0.966 | 0.971 | 0.979 |
| conservative rate ($n^{1/2}$) | 0.991 | 0.982 | 0.975 | 0.974 | 0.979 |
| (infeasible) exact rate ($n^{1/2}$) | 0.991 | 0.982 | 0.975 | 0.974 | 0.979 |

Table 2: Coverage Probabilities for Design 2

| | $n = 100$ | $n = 500$ | $n = 1000$ | $n = 2000$ | $n = 5000$ |
|---|---|---|---|---|---|
| | nominal 90% coverage | | | | |
| estimated rate | 0.26 | 0.13 | 0.08 | 0.06 | 0.03 |
| conservative rate ($n^{1/2}$) | 0.33 | 0.17 | 0.12 | 0.09 | 0.06 |
| (infeasible) exact rate ($n^{3/5}$) | 0.21 | 0.10 | 0.07 | 0.05 | 0.03 |
| | nominal 95% coverage | | | | |
| estimated rate | 0.35 | 0.17 | 0.11 | 0.07 | 0.05 |
| conservative rate ($n^{1/2}$) | 0.39 | 0.22 | 0.15 | 0.11 | 0.07 |
| (infeasible) exact rate ($n^{3/5}$) | 0.29 | 0.13 | 0.09 | 0.06 | 0.04 |

Table 3: Mean of $\hat{u}_{1-\alpha} - \theta_{1,D1}$ for Design 1

| | $n = 100$ | $n = 500$ | $n = 1000$ | $n = 2000$ | $n = 5000$ |
|---|---|---|---|---|---|
| | nominal 90% coverage | | | | |
| estimated rate | 0.11 | 0.08 | 0.06 | 0.04 | 0.02 |
| conservative rate ($n^{1/2}$) | 0.20 | 0.09 | 0.06 | 0.04 | 0.02 |
| (infeasible) exact rate ($n^{1/2}$) | 0.20 | 0.09 | 0.06 | 0.04 | 0.02 |
| | nominal 95% coverage | | | | |
| estimated rate | 0.18 | 0.10 | 0.07 | 0.05 | 0.03 |
| conservative rate ($n^{1/2}$) | 0.27 | 0.11 | 0.08 | 0.05 | 0.03 |
| (infeasible) exact rate ($n^{1/2}$) | 0.27 | 0.11 | 0.08 | 0.05 | 0.03 |

Table 4: Mean of $\hat{u}_{1-\alpha} - \theta_{2,D2}$ for Design 2

| | $\theta_1$ | $\theta_2$ |
|---|---|---|
| Estimated Rate | $[-48, 84]$ | $[0.0113, 0.0342]$ |
| Conservative Rate | $[-60, 138]$ | $[0.0030, 0.0372]$ |
| LAD with Points | $[-63, 63]$ | $[0.0100, 0.0244]$ |

Table 5: 95% Confidence Intervals for Components of $\theta$