Non-Standard Rates of Convergence of Criterion-Function-Based Set Estimators for Binary Response Models[*]

JASON R. BLEVINS

The Ohio State University, Department of Economics

February 18, 2015[†]
Forthcoming at the *Econometrics Journal*

Abstract. This paper establishes consistency and non-standard rates of convergence for set estimators based on contour sets of criterion functions for a semiparametric binary response model under a conditional median restriction. The model may be partially identified due to potentially limited-support regressors. A set estimator analogous to the maximum score estimator is essentially cube-root consistent for the identified set when a continuous but possibly bounded regressor is present. Arbitrarily fast convergence occurs when all regressors are discrete. We also establish the validity of a subsampling procedure for constructing confidence sets for the identified set. As a technical contribution, we provide more convenient sufficient conditions on the underlying empirical processes for cube root convergence and a sufficient condition for arbitrarily fast convergence, both of which can be applied to other models. Finally, we carry out a series of Monte Carlo experiments which verify our theoretical findings and shed light on the finite sample performance of the proposed procedures.

Keywords: partial identification, cube-root asymptotics, semiparametric models, limited support regressors, transformation model, binary response model, maximum score estimator.

JEL Classification: C13, C14, C25.

[*] This paper is based in part on Chapter 2 of my Duke University dissertation. I am grateful to the members of my committee, Han Hong, Shakeeb Khan, Paul Ellickson, and Andrew Sweeting, for their guidance and support. I also thank Arie Beresteanu, Federico Bugni, Stephen Cosslett, Yanqin Fan, Joseph Hotz, Robert de Jong, Joris Pinkse, Jean-François Richard, Azeem Shaikh, Xiaoxia Shi, Elie Tamer, Jörg Stoye, and Randy Walsh for helpful discussions and seminar participants at Duke University, Ohio State University, the University of Illinois at Urbana-Champaign, the University of Pittsburgh, the University of Texas at Austin, Vanderbilt University, the 2009 Triangle Econometrics Conference and the 2010 Annual Meeting of the Midwest Econometrics Group for useful comments.

[†] This paper is a revision of a working paper whose first draft appeared on October 31, 2011.

1. Introduction

This paper considers a semiparametric binary response model and develops several asymptotic results for criterion-function-based set estimators of the kind considered by Chernozhukov, Hong, and Tamer (2007) (henceforth CHT). First, we verify the conditions of CHT for the semiparametric binary choice model under a conditional median restriction to establish cube-root consistency of a set estimator for the identified set when a continuous regressor is present. A second, technical contribution is to provide new sufficient conditions (in Appendix B) for cube-root consistency of set estimators based on contour sets of criterion functions that can be used to analyze other models and which may be easier to verify than the more general ones currently available in the literature. Third, we verify the conditions of Romano and Shaikh (2010) for subsampling-based inference for the binary choice model and determine the appropriate scaling sequence for the inferential statistic used in their procedure for this case. Fourth, for the binary response model with discrete regressors we show that the rate of convergence is arbitrarily fast, which agrees with previous findings of Komarova (2013) for a related estimator based on a recursive linear programming algorithm. Fifth, we show that the source of this property is a discontinuity in the limiting objective function and give a general condition under which the identified set can be estimated at an arbitrarily fast rate in other models with this feature. Finally, we carry out a series of Monte Carlo experiments to verify our theoretical findings and explore the small sample behavior of the proposed estimators.

This paper builds on a large literature on partially identified models. We consider criterion-function-based estimation and inference which started with Manski and Tamer (2002), who analyzed a semiparametric binary response model with interval-valued data under a conditional quantile restriction. They derived the sharp identified set for the model, proposed a set estimator, defined as an appropriately-chosen contour set of a modified maximum score objective function, and showed that it was consistent. Chernozhukov, Hong, and Tamer (2007) developed a broad framework for criterion-function-based estimation and established general conditions for consistency and rates of convergence of estimators in this class. They also proposed a subsampling-based procedure for obtaining confidence sets with some pre-specified coverage probability. Subsampling-based inference was further explored by Romano and Shaikh (2008, 2010) and Andrews and Guggenberger (2009) while Bugni (2010) and Canay (2010) have proposed bootstrap procedures. These and many other authors have studied inference in moment equality and inequality models, including but certainly not limited to Andrews and Barwick (2012), Pakes, Porter, Ho, and Ishii (2011), Beresteanu and Molinari (2008), Beresteanu, Molchanov, and Molinari (2011), Imbens and Manski (2004), Stoye (2009), Kim (2008), Khan and Tamer (2009), Menzel (2014), Andrews and Shi (2013), Andrews and Soares (2010), and Yildiz (2012).

Our results also follow a long line of work on identification and estimation of binary response models under various conditions. Once parametric binary response models were well-understood (McFadden, 1974; Maddala, 1983; Amemiya, 1985), semiparametric methods emerged to estimate models without making parametric assumptions about the error distribution. Such methods include maximum score (Manski, 1975, 1985; Kim and Pollard, 1990; Horowitz, 1992), distribution-free maximum likelihood (Cosslett, 1983), average derivative estimation (Stoker, 1986), maximum rank correlation (Han, 1987), kernel estimators (Ichimura, 1993; Klein and Spady, 1993), and instrumental variables (Lewbel, 2000). Matzkin (1992) studied nonparametric identification and estimation of binary response models.

Finally, this paper is also related to a growing literature concerned with semiparametric estimation of models with limited support regressors, typically involving either discrete or interval-valued regressors. Bierens and Hartog (1988) showed that there are infinitely many single-index representations of the mean regression of a dependent variable when all covariates are discrete. Manski and Tamer (2002) considered partial identification and estimation of binary response models with an interval-valued regressor. Honoré and Tamer (2006) discussed partial identification due to the initial conditions problem in dynamic random effects discrete choice models with discrete regressors. Magnac and Maurin (2008) considered a similar model, but in the cases of discrete or interval-valued regressors and in the presence of a special regressor which satisfies both a partial independence and a large support condition. Honoré and Lleras-Muney (2006) estimated a partially identified competing risks model with interval outcome data and discrete explanatory variables. Horowitz (2009) discussed the generic non-identification of single-index and binary response models with only discrete regressors, a result which serves to motivate our analysis. Komarova (2013) proposed consistent estimators, based on a linear programming procedure, of the identified set in a binary response model with discrete regressors. Wan and Xu (2015) study the asymptotic properties of set estimators for semiparametric binary response models with interval valued regressors. Finally, the importance of the theoretical topics addressed in this paper are highlighted by empirical work using maximum score methods, such as that of Bajari, Fox, and Ryan (2008).

2. Semiparametric Binary Response Model

Our leading example throughout the paper is the semiparametric binary response model. Manski (1988) studied identification of additively separable binary response models in the presence of a continuous regressor and compared the identification power of several assumptions, showing that mean independence has no identifying power but that quantile independence can be sufficient for point identification. As such, we focus on the binary response model under a conditional median restriction. Estimators for this model are based on a simple rank condition of the form $y = 1 \iff x'\theta \geq 0$ where y is the binary outcome, x is a vector of regressors, and θ is a vector of parameters. In the point identified case, this includes the maximum score estimator of Manski (1975, 1985) and smoothed maximum score estimator of Horowitz (1992).

Both these and the related semiparametric methods mentioned above typically assume the existence of an exogenous explanatory variable with rich support. Rank conditions have been successful in estimating more general regression models, but the known conditions for point identification still include a rich support condition (Han, 1987; Abrevaya, 2000). In practice, however, it is not uncommon to encounter datasets with genuinely discrete or bounded variables. Without a regressor with full support on the real line, under semiparametric assumptions, the models we consider are only partially identified in general (Horowitz, 2009).

We now formalize the basic linear-index binary response model of interest.

Model 1 (Semiparametric Binary Response Model)**.** *Let the outcome $y \in \{0,1\}$ be determined as*

$$y = 1\{x'\theta + u \geq 0\}$$

where x is a random vector with support $\mathscr{X} \subseteq \mathbb{R}^K$, and θ is the parameter of interest, a member of some parameter space $\Theta \subset \mathbb{R}^K$. The distribution of u satisfies $\operatorname{Med}(u \mid x) = 0$ *F_x-a.s.*

The model and assumptions are the same as in the maximum score model of Manski (1975, 1985), but without the support and rank assumptions on x. Although we assume that the conditional median of the error term is zero, this is for simplicity. A similar assumption could be made on any other quantile instead (Manski, 1988). In contrast to Magnac and Maurin (2008), we make no additional assumptions on the support of u.

We note that even in the point identified case, the maximum score estimator is essentially a set estimator. Because the sample objective function is a step function, there is no guidance about which point from the set of maximizers to choose as a point estimate. Asymptotically, the estimator is consistent for any such selection. In practice the selection rule is typically implicit or ad hoc, being determined by the stopping criteria of an optimization routine. In the absence of a suitable selection rule, one may as well regard the estimator as a set estimator consisting of all values of θ that maximize the objective function. In this sense, the estimators we present below can be used without regard to whether the model is point identified or partially identified. Consistency guarantees that the limit is the respective population point or set of interest.

Modulo assumptions on the errors, point identification of θ hinges on what one knows about the distribution of x. The validity of a full support assumption is application-specific. Many variables such as age, number of children, years of education, and gender are inherently discrete and so a full support assumption is clearly inappropriate in these cases. Similarly, even variables such as income have only partial support on the real line. One advantage of the estimators we propose is that one need not distinguish between the point and partial identification. That is, they do not require a regressor with full support but exploit the additional information provided by one when available. We work under the following alternative assumptions.

Assumption C1 (Continuous Regressor)**.** The K-th component of the random vector x, denoted x_K, has positive density everywhere on a set $\mathcal{X}_K \subseteq \mathbb{R}$ conditional on almost every value of the remaining components and $\theta_K \neq 0$.

Assumption C2 (Discrete Regressors)**.** The vector-valued random variable x has finite support $\mathcal{X} = \{x^1, x^2, \dots, x^L\} \subset \mathbb{R}^K$ for $L < \infty$ and $P(y = 1 \mid x^l) \neq \frac{1}{2}$ for all $l = 1, \dots, L$.

Manski (1975, 1985) showed that when one component of x is continuously distributed and has support equal to $\mathbb{R}$, conditional on almost every value of the remaining components, then θ is point identified (provided that the components of x are also linearly independent). Assumption C1 is weaker than this because the support of the continuous component may be bounded. Note that Assumption C1, which is a restriction on the conditional density, does not rule out the possibility that $\mathcal{X}_K = \mathbb{R}$, but it also includes cases where the support of x is bounded. In Assumption C2, the requirement that the population response probabilities are not exactly $\frac{1}{2}$ serves to avoid problems that prevent consistent estimation of the identified set in this case, as discussed in detail by Komarova (2013).

To provide another example, consider the following semiparametric transformation model which generalizes Model 1 by substituting an unknown, weakly monotonic function Λ in place of the indicator function and by allowing the outcome to be continuous.

Model 2 (Semiparametric Transformation Model)**.** *Let $y \in \mathbb{R}$ be determined by $y = \Lambda(x'\theta + u)$ where x is a random vector with support $\mathcal{X} \subseteq \mathbb{R}^K$, θ is the parameter of interest, a member of some parameter space $\Theta \subset \mathbb{R}^K$, and the function $\Lambda : \mathbb{R} \to \mathbb{R}$ is (weakly) monotonic. The distribution of u satisfies* $\operatorname{Med}(u \mid x) = 0$ *F_x-a.s.*

Even with continuous variation in y, θ may only be partially identified if no component of x has full support on $\mathbb{R}$, conditional on the remaining components, when u is heteroskedastic. In contrast, Han (1987) achieved point identification by assuming one component of x has full support, conditional on the remaining components, and that u and x are independent. Without independence, Model 2 is very similar to Model 1 in terms of identification and estimation (the latter arises with $\Lambda(\cdot) = 1\{\cdot \geq 0\}$), with both being characterized in terms of a rank condition that can be used to construct an objective function. We will thus focus on Model 1 in the remainder.

2.1. Partial identification

Relaxing assumptions in econometric models is often desirable but can lead to a failure of point identification. For example, in Model 1 we avoid both a parametric distributional assumption on u and a support condition on x, either of which would suffice for point identification. Fortunately, a recent and growing literature on partially identified models has shown that in many cases we can still carry out inference about the parameters of interest even under assumptions weaker than those known to provide point identification.[1] In particular, a criterion-function-based approach to set estimation, motivated by classical extremum estimation of point-identified models, has proven useful for analyzing partially identified models such as those based on moment inequalities. Set estimators for this class of models, under certain regularity conditions, are essentially $\sqrt{n}$-consistent (Chernozhukov et al., 2007).

Model 1 does not satisfy the same regularity conditions and so the set estimator based on the maximum score objective function is not $\sqrt{n}$-consistent. Kim and Pollard (1990) showed that the point estimator (under a full support condition) is cube-root-consistent and has a non-standard limiting distribution. These properties are often perceived as disadvantages, but the maximum score estimator has proven to be important and useful due to its robustness, both to unknown error distributions and to heteroskedasticity of unknown form. Furthermore, recent work on classical MCMC-based inference by Jun, Pinkse, and Wan (2011) makes point estimation and inference for this model much more accessible.

The primitives of Model 1 are θ and $F_{u|x}$, but θ is the only finite-dimensional parameter of interest. The identified set is the collection of parameters θ that are consistent with the data generating process P for *some* distribution $F_{u|x}$. The following lemma provides a tractable representation of the identified set for θ in Model 1 in terms of observables.

Lemma 1. *In Model 1, the identified set is*

$$\Theta_0 = \left\{\theta \in \Theta : \operatorname{sgn}\left(P(y=1 \mid x) - 1/2\right) = \operatorname{sgn}(x'\theta) \quad F_x - a.s.\right\} \tag{1}$$

where $\operatorname{sgn}(\cdot)$ *is defined as* $\operatorname{sgn}(z) = 1\{z \geq 0\} - 1\{z < 0\}$. *Furthermore,* Θ_0 *is nonempty and convex.*

Proof of Lemma 1. That the identified set, Θ_0, equals the set on the right hand side of (1) follows from Proposition 2 of Manski (1988). The set Θ_0 is nonempty because $\theta_0 \in \Theta_0$. To see that Θ_0 is convex, let $\theta^1, \theta^2 \in \Theta_0$, let $\alpha \in (0,1)$, and define $\tilde{\theta} \equiv \alpha\theta^1 + (1-\alpha)\theta^2$. Since $\theta^1, \theta^2 \in \Theta_0$, from (1) it must be the case that $\operatorname{sgn}(x'\theta^1) = \operatorname{sgn}(x'\theta^2)$ F_x-a.s. Furthermore, $0 < \alpha < 1$ and (1) imply

$$\operatorname{sgn}\left(x'\tilde{\theta}\right) = \operatorname{sgn}\left(\alpha x'\theta^1 + (1-\alpha)x'\theta^2\right) = \operatorname{sgn}\left(x'\theta^1\right) = \operatorname{sgn}\left(P(y=1 \mid x) - 1/2\right) \quad F_x\text{-a.s.}$$

Therefore, $\tilde{\theta} \in \Theta_0$. ■

[1] See Manski (2003) and Tamer (2010) and the references therein for broad surveys of this literature.

Although it is possible for point identification to obtain under Assumption C1 with additional restrictions on the conditional distribution of one component of x (Horowitz, 2009, Corollary 4.1), we note that Assumption C1 is not on its own sufficient for point identification. Similarly, it may also be possible to identify the signs of individual components of θ with only bounded, but continuous support of one component of x (Manski, 1988). However, without additional assumptions the present assumptions are not sufficient for point identification. Similarly, there are special cases under Assumption C2 in which θ is point identified but additional assumptions are needed to guarantee it (Horowitz, 1992, Section 4.2.2). Finally, we note that although the focus of this paper is inference on the identified set Θ_0, another widely-adopted approach to inference in the literature is to focus on the true parameter $\theta_0 \in \Theta_0$.

2.2. Criterion-Function-Based Set Estimation

Following Manski and Tamer (2002) and subsequent work by Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2008, 2010), Bugni (2010), Kim (2008), Yildiz (2012), and many others, we consider set inference in models where the identified set is characterized by some criterion function Q. The analogy principle suggests defining an estimator $\hat{\Theta}_n$ for Θ_0 based on the set of maximizers of the sample criterion function Q_n, which is the finite sample analog of Q. In particular, estimators are defined in terms of upper contour sets of Q_n. Let $C_n(\tau_n)$ denote the upper contour set of level τ_n, defined as

$$C_n(\tau_n) \equiv \left\{\theta \in \Theta : Q_n(\theta) \geq \sup_{\Theta} Q_n - \tau_n\right\}, \tag{2}$$

where τ_n is a non-negative "slackness" sequence which converges zero in probability.

Taking only the set of maximizers (by setting $\tau_n = 0$) may result in an inconsistent estimator (Manski and Tamer, 2002). This is a problem with estimation in partially identified models: the literature has provided guidance about the rate at which τ_n must converge to zero, but not about choosing the constant of proportionality. This leaves a degree of freedom in choosing τ_n and it is not clear how the choice affects finite sample performance. As such, in the Monte Carlo section we compare the performance of four data-driven choices of the constant of proportionality for the sequence τ_n. Fortunately, this is not a problem for doing inference and reporting confidence sets, which does not require a slackness sequence (Romano and Shaikh, 2010).

To discuss set convergence, we must first specify a metric space. Again following the literature, we consider convergence in terms of the *Hausdorff distance*, a generalization of Euclidean distance to spaces of sets. Let (Θ, d) be a metric space where d is the standard Euclidean distance. For a pair of subsets $A, B \subset \Theta$, the Hausdorff distance between A and B is

$$d_{\mathrm{H}}(A, B) \equiv \max\left\{\sup_{\theta\in B} \rho(\theta, A), \sup_{\theta\in A} \rho(\theta, B)\right\}, \tag{3}$$

where $\rho(\theta, A) \equiv \inf_{\tilde{\theta} \in A} d(\theta, \tilde{\theta})$ is the shortest distance from the point θ to the set A. Intuitively, the Hausdorff distance between A and B is the farthest distance between an arbitrary point in one of the sets to the nearest neighbor in the other set.

Conveniently, we can characterize the identified set in Model 1 using the usual maximum score objective function. The population and sample analog objective functions are

$$Q(\theta) = \mathrm{E}\left[(2y-1)\,\mathrm{sgn}(x'\theta)\right] \quad \text{and} \quad Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(2y_i - 1)\,\mathrm{sgn}(x_i'\theta).$$

The following lemma establishes that the set of maximizers of Q provides a sharp characterization of the identified set, justifying the use of the proposed criterion-function-based set estimators.

Lemma 2. *For Model 1 with either Assumption C1 or C2,* $\arg\max_{\Theta} Q = \Theta_0$.

Proof of Lemma 2. It follows from the structure of Q and the proof of Lemma 1 that Q is maximized whenever the signs of $2y-1$ and $x'\theta$ agree F_x-a.s., so that their product equals 1 and not -1. Importantly, under both assumptions C1 and C2 the event $2y-1=0$ occurs with probability zero so the potential ambiguity over the sign of $x'\theta$ in that case is inconsequential. ■

In each case—continuous and discrete regressors—this function has features that differ from the objective functions used for moment inequality models. This will lead us to introduce slight modifications of the conditions of CHT for showing consistency and deriving the rates of convergence of the set estimators. First, under Assumption C1 the rate of uniform convergence in probability of Q_n to Q is $n^{-1/2}$ over Θ while for the moment inequality models considered by CHT the rate is faster over Θ_0 than Θ. We will return to this point when considering the rate of convergence in the next section, but it will also allow us to simplify the conditions for consistency. Second, under Assumption C2 the population objective function is a step function and so we must allow for models where Q is discontinuous. This will also be an important determinant of the rate of convergence which we will return to after establishing consistency.

2.3. Consistency

Both Manski and Tamer (2002) and CHT give consistency results for criterion function based set estimators. Manski and Tamer (2002) originally considered a case where the limiting objective function is continuous, so their set consistency result (Proposition 3) will not apply in the case of Model 1 with discrete regressors. However, Theorem 3.1 of CHT requires only semi-continuity, which will suffice for our purposes. Here, we present a specialized version of their consistency theorem which will suffice for our needs and which is stated in terms of maximization, rather than minimization, for clarity. [2] In particular, Theorem 1 below provides conditions on Q, Q_n, and the sequence τ_n to ensure that $\hat{\Theta}_n \equiv C_n(\tau_n)$ is consistent for Θ_0.

Assumption A1 (Compactness)**.** Θ is a nonempty, compact subset of $\mathbb{R}^K$.

Assumption A2 (Well-Separated Maximum)**.** There exists a population criterion function Q such that for all $\eta > 0$, there exists a $\delta_\eta > 0$ such that $\sup_{\Theta \backslash \Theta_0^\eta} Q \leq \sup_\Theta Q - \delta_\eta$.

Assumption A3 (Uniform Convergence)**.** There exists a sample criterion function Q_n and a known sequence of constants $a_n \to \infty$ such that $\sup_\Theta |Q_n - Q| = O_p(1/a_n)$.

Assumptions A1 and A3 are analogous to the standard compactness and uniform convergence conditions for consistency of M-estimators for singletons (cf. Amemiya, 1985; Newey and McFadden, 1994). Assumption A2 requires the population objective function to have a well-separated maximum. This serves to rule out pathological cases that can arise in the absence of continuity. It is satisfied, for example, when Q is a continuous function or a step function (which may have only a finite number of steps) or when Q is upper semicontinuous in a neighborhood of the identified set (cf. condition C.1(b) of CHT). Assumption A3 requires that Q_n converge uniformly in probability to Q and is similar to others in the literature on set estimation in that it also requires that the rate of uniform convergence is known (cf. conditions C.1(d) and C.1(e) of CHT). In this sense it is slightly stronger than the usual assumption for M-estimators, but this rate is easy to determine in applications. For Model 1 and other models that satisfy the conditions in Appendix B, we show that $a_n = n^{1/2}$.

Theorem 1 (Consistency)**.** *Suppose that Assumptions A1–A3 hold and let τ_n be a nonnegative sequence of random variables such that $\tau_n \xrightarrow{p} 0$. Then, $\sup_{\theta \in \hat{\Theta}_n} \rho(\theta, \Theta_0) \xrightarrow{p} 0$. Furthermore, if $a_n \tau_n \xrightarrow{p} \infty$, then $\lim_{n \to \infty} P(\Theta_0 \subseteq \hat{\Theta}_n) = 1$ and $d_{\mathrm{H}}(\hat{\Theta}_n, \Theta_0) \xrightarrow{p} 0$.*

The proof, along with other longer proofs and auxiliary results, is given in the appendix. Note that the first conclusion of Theorem 1 actually holds without the slackness sequence: $\hat{\Theta}_n$ becomes arbitrarily close to being a subset of Θ_0 in probability for any $\tau_n = o_p(1)$ including $\tau_n = 0$. The

[2] First, we note that Assumption A2 here plays a similar role as the upper (lower) semi-continuity of condition C.1(b) of CHT but does not restrict the objective function outside of a neighborhood of the identified set. Second, our assumption on the rate of uniform convergence is a simplified version of the CHT assumptions. We assume that the rate of uniform convergence in probability of the objective function is known over the entire parameter space (Assumption A3). On the other hand, CHT assume that the rate of (one-sided) uniform convergence in probability on the entire parameter space is known (condition C.1(d)) and that the rate of uniform convergence in probability over the identified set is also known but may be different (conditions C.1(e)). In CHT, these rates differ in the moment inequality examples they consider. These rates are equal in the models we consider, allowing us the small simplification of needing to introduce notation for only one rate.

slackness sequence ensures that the other inclusion holds—that $\hat{\Theta}_n$ covers Θ_0 in probability—by expanding the contour sets by an amount which becomes negligible as $n \to \infty$. By expanding it at the right rate—with τ_n converging to zero in probability, but not faster than $1/a_n$—we ensure that $\hat{\Theta}_n$ is large enough to cover Θ_0 with probability approaching one. Combining these two results yields consistency in the Hausdorff metric.

In the binary choice model with iid data, $a_n = n^{1/2}$ and we can obtain a consistent estimator by choosing τ_n to be a sequence which converges to zero slower than $n^{-1/2}$. Permissible choices are, for example, $\tau_n = \sqrt{\ln n / n}$ and $\tau_n = n^{-0.49}$. As we discuss in the next section, the rate of convergence will be faster when τ_n is closer to being proportional to $n^{-1/2}$. Hence, in our Monte Carlo experiments we choose τ_n to be proportional to $n^{-0.49}$ and consider different choices for the constant of proportionality.[3]

Assumption C3 (Random Sampling)**.** The sample consists of n independent observations from the population distribution of observables.

Lemma 3. *In Model 1 with either Assumption C1 or C2 and Assumption C3 for any sequence* $\tau_n \xrightarrow{\text{p}} 0$ *with* $n^{1/2}\tau_n \xrightarrow{\text{p}} \infty$, $d_{\text{H}}(\hat{\Theta}_n, \Theta_0) \xrightarrow{\text{p}} 0$.

3. Non-Standard Rates of Convergence

The rate of convergence of the Hausdorff distance $d_{\text{H}}(\hat{\Theta}_n, \Theta_0)$ is the slowest rate at which the component distances in (3), $\sup_{\theta \in \Theta_0} \rho(\theta, \hat{\Theta}_n)$ and $\sup_{\theta \in \hat{\Theta}_n} \rho(\theta, \Theta_0)$, converge to zero. The second part of Theorem 1 establishes that with only Assumptions A1–A3, the first distance equals zero with probability approaching one. Thus, the rate of convergence of the second component distance determines the overall rate.

For the binary choice model, the rate of convergence depends on whether Assumption C1 or C2 is satisfied. In the case of a continuous regressor, we use results of CHT to establish cube-root consistency and then verify the conditions of Romano and Shaikh (2010) for constructing confidence sets. In the case of only discrete regressors, our results for the binary choice model are closely related to those of Komarova (2013). We verify that the criterion-function-based estimator using the maximum score objective function has an arbitrarily fast rate of convergence like the estimator based on the recursive linear programming procedure she developed. Our contribution relative to her work is to show that arbitrarily fast convergence will arise more generally in any model for which Q exhibits a discontinuity at the boundary of the identified set.

[3] Sequences such as $n^{-0.49999}$ will yield a faster rate but the differences are small even for very large sample sizes.

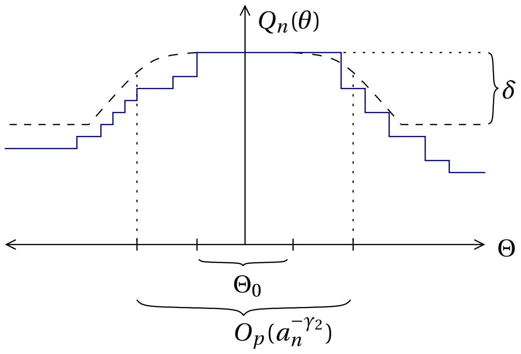


FIGURE 1. Polynomial majorant condition (Assumption A4) with $\gamma_1 = 2$.

Note: Here Q_n is bounded above in probability by a polynomial in $\rho(\theta, \Theta_0)$ outside of an $O_p(a_n^{-\gamma_2})$ neighborhood of Θ_0 and $\delta > 0$ is a threshold below which the bound is relaxed.

3.1. Polynomial Rates of Convergence

In this section we consider models for which Q_n satisfies a polynomial curvature condition based on Condition C.2 of CHT. In particular, when $Q_n(\theta)$ is stochastically bounded from above by a polynomial in $\rho(\theta, \Theta_0)$ outside of a shrinking neighborhood of Θ_0, we show that the rate of convergence depends on the curvature of the bounding polynomial, the rate at which the neighborhood shrinks, and the rate at which τ_n converges to zero.

Assumption A4 (Polynomial Majorant)**.** There exist positive constants δ, c, γ_1, and γ_2 with $\gamma_1 > 1$ and $\gamma_1\gamma_2 \geq 1$ such that for $\varepsilon \in (0,1)$ there are positive constants c_ε and n_ε such that for all $n \geq n_\varepsilon$,

$$Q_n(\theta) \leq \sup_{\Theta} Q - c \cdot (\rho(\theta, \Theta_0) \wedge \delta)^{\gamma_1}$$

uniformly on the set $\{\theta \in \Theta : \rho(\theta, \Theta_0) \geq (c_\varepsilon / a_n)^{\gamma_2}\}$ with probability at least $1 - \varepsilon$.

Assumption A4 is a marginal relaxation of Condition C.2 of CHT needed to cover cases of interest in this paper. In particular, their assumption is a special case of the above where $\gamma_1 = 1/\gamma_2$. In other words, to analyze Model 1 we must allow the degree of the bounding polynomial to differ from the parameter determining the rate at which the sequence neighborhoods of Θ_0 shrinks. As in the case of M-estimators for point identified models, this generalization allows for models with non-standard rates of convergence (cf. van der Vaart, 1998, Theorem 5.52). As an example, Figure 1 illustrates a possible quadratic bounding polynomial ($\gamma_1 = 2$). As depicted in the figure, δ is a threshold value below which the bound is relaxed.

Theorem 2 (Rate of Convergence with a Polynomial Majorant)**.** *Suppose that Assumptions A1–A4 hold. If $\tau_n \xrightarrow{p} 0$ and $a_n \tau_n \xrightarrow{p} \infty$, then $d_H(\hat{\Theta}_n, \Theta_0) = O_p(\tau_n^{\gamma_2})$.*

Although the rate a_n does not appear explicitly in the conclusion of Theorem 2, the rate of convergence of $\hat{\Theta}_n$ depends implicitly on a_n because the curvature and rate constants γ_1 and γ_2 must be chosen relative to a_n in order to satisfy Assumption A4. To see this, note that if we choose $\tau_n \approx a_n$ then the rate of convergence is $\tau_n^{\gamma_2} \approx a_n^{-\gamma_2}$.

To see why we need to allow $\gamma_1 \neq 1/\gamma_2$, suppose for a moment that they are equal. Since the objective function for the binary response model is approximately quadratic near the identified set we have $\gamma_1 = 2$ and so the equality requires $\gamma_2 = 1/2$. Furthermore, for the binary choice model we have $a_n = n^{1/2}$. So, by Theorem 3.1 of CHT the rate of convergence of the set estimator would be $\tau_n^{-\gamma_2} = \tau_n^{-1/2}$, which can be nearly as fast as, but no faster than $n^{-1/4}$ if we choose τ_n appropriately. However, we can improve on this because the quadratic bound actually holds outside of a sequence of neighborhoods that shrinks at the faster rate $n^{-1/3}$ (i.e., $a_n^{-\gamma_2}$ with $\gamma_2 = 2/3$) rather than the slower rate $n^{-1/4}$ (i.e., $a_n^{-\gamma_2}$ with $\gamma_2 = 1/2$) at which we are restricted to use when $\gamma_1 = 1/\gamma_2$. In other words, by allowing $\gamma_1 \neq 1/\gamma_2$, we can show that the rate of convergence is $\tau_n^{-2/3}$ which can be made arbitrarily close to $n^{-1/3}$.

3.2. Cube Root Consistency in the Semiparametric Binary Response Model

The properties of the maximum score objective function in the continuous covariate case have been studied by Kim and Pollard (1990), Abrevaya and Huang (2005), and others. The following lemma formalizes the cube-root consistency result for the set estimator for our model. Although our assumptions are weaker overall, we still need to make additional assumptions on the distribution of x, which, for comparison, are intentionally close to the assumptions of Abrevaya and Huang (2005) and Horowitz (1992) in analyzing the model in the point identified case.

It is well known that θ is only identified up to scale, so we normalize the coefficient on the last component of x, denoted θ_K, to be either 1 or -1 and consider estimation of $\beta \in \mathbb{R}^{k-1}$ where $\theta = (\beta', \theta_K)'$. Let $\tilde{x}$ denote the first $K-1$ components of x. Then $x'\theta = \tilde{x}'\beta + x_K$. Being limited to a finite set, we can estimate θ_K at a faster rate than the remaining components, so without loss of generality we only consider the case where $\theta_K = 1$. Therefore, for stating the following lemma we abuse notation slightly by writing $Q(\beta) = Q((\beta', 1)')$. Accordingly, let $B \subset \mathbb{R}^{k-1}$, $B_0 \subset B$, and $\hat{B}_n$ denote, respectively, the parameter space for β, the identified set, and the set estimator.

Lemma 4. *Suppose that Assumptions C1 and C3 hold in Model 1. In addition, suppose:*

a. The components of $\tilde{x}$ and $\tilde{x}\tilde{x}'$ have finite first absolute moments.

b. The function $\partial f_{x_K|\tilde{x}}(x_K \mid \tilde{x})/\partial x_K$ exists and for some $M > 0$, $\left| f_{x_K|\tilde{x}}(x_K \mid \tilde{x})/\partial x_K \right| < M$ and $f_{x_K|\tilde{x}}(x_K \mid \tilde{x}) < M$ for all x_K and almost every $\tilde{x}$.

c. For all u in a neighborhood of 0, all x_K in a neighborhood of $-\tilde{x}'\beta_0$, almost every $\tilde{x}$, and some $M > 0$, the function $f_{u|x}(u \mid \tilde{x}, x_K)$ exists and $f_{u|x}(u \mid \tilde{x}, x_K) < M$.

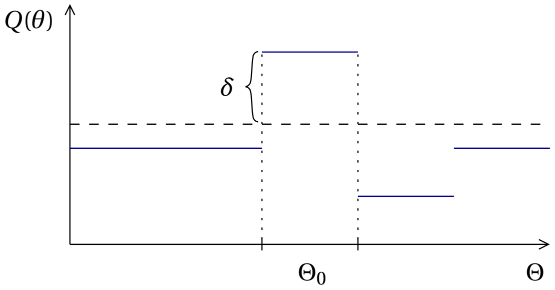


FIGURE 2. Constant majorant condition (Assumption A4').

Note: Here Q has a discontinuity at the boundary of Θ_0 such that $Q(\theta) \le \sup_\Theta Q - \delta$ for some $\delta > 0$.

d. *For all u in a neighborhood of 0, all x_K in a neighborhood of $-\tilde{x}'\beta_0$, almost every $\tilde{x}$, and some $M > 0$, the function $\partial F_{u|x}(u \mid \tilde{x}, x_K)/\partial x_K$ exists and $\left|\partial F_{u|x}(u \mid \tilde{x}, x_K)/\partial x_K\right| < M$.*

e. *B_0 is compact and contained in the interior of B.*

f. *$V(\beta) \equiv \mathrm{E}\left[2 f_{u|x}(0 \mid \tilde{x}, -\tilde{x}'\beta) f_{x_K|\tilde{x}}(-\tilde{x}'\beta \mid \tilde{x})\tilde{x}\tilde{x}'\right]$ is positive definite for all $\beta \in \mathrm{bd}(B_0)$.*

Then for any sequence τ_n such that $\tau_n \xrightarrow{\mathrm{p}} 0$ and $n^{1/2}\tau_n \xrightarrow{\mathrm{p}} \infty$, $d_{\mathrm{H}}(\hat{\Theta}_n, \Theta_0) = O_p(\tau_n^{2/3})$.

3.3. Arbitrarily Fast Convergence

This section addresses a special case in which the limiting objective function Q is not continuous, but has a discontinuity at the boundary of the set Θ_0 as illustrated by Figure 2. This occurs, for example, in Model 1 under Assumption C2 (discrete regressors), where Q is a step function. We show that in such cases $\hat{\Theta}_n$ converges arbitrarily fast in probability to Θ_0, as opposed to the (potentially) non-standard but polynomial rates we found above.

Assumption A4' (Constant Majorant)**.** There exists a $\delta > 0$ such that $Q(\theta) \le \sup_\Theta Q - \delta$ for all $\theta \in \Theta \setminus \Theta_0$.

Theorem 3 (Rate of Convergence with a Constant Majorant)**.** *Suppose that Assumptions A1–A3 and A4' hold. If $\tau_n \xrightarrow{\mathrm{p}} 0$ and $a_n\tau_n \xrightarrow{\mathrm{p}} \infty$, then $\hat{\Theta}_n = \Theta_0$ with probability approaching one.*

Thus, when Q exhibits a jump at the boundary of the identified set, the probability that the estimate actually *equals* the identified set can be made arbitrarily close to one by choosing n large enough. This is equivalent to saying that $\hat{\Theta}_n$ converges arbitrarily fast to Θ_0. That is, for *any* sequence r_n, including powers of n and exponential forms, $r_n d_{\mathrm{H}}(\hat{\Theta}_n, \Theta_0) \xrightarrow{\mathrm{p}} 0$.

Limiting objective functions satisfying Assumption A4' arise, for example, when the regressors are all discrete. Suppose, as in the binary choice model, that the objective function can be expressed in terms of a class of real-valued functions $\mathscr{F}$, where for each $\theta \in \Theta$, $Q(\theta) = Pf(\cdot, \theta)$

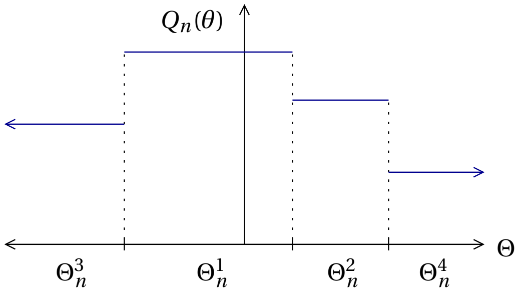


FIGURE 3. A realization of Q_n and the partition of Θ it generates.

Note: A step function Q_n generates a finite number of contour set estimates $\hat{\Theta}_n$ given by $\cup_{j=1}^{k}\Theta_n^j$ for $k = 1,\ldots,4$.

and $Q_n(\theta) = P_n f(\cdot,\theta)$ for $f(\cdot,\theta) \in \mathscr{F}$ and where P and P_n denote, respectively, the population and empirical measures of the observables. When the explanatory variables are discrete and the functions in $\mathscr{F}$ are discontinuous, then Q may be discontinuous at the boundary of the identified set. With a continuous regressor Q may be smooth even if the functions in $\mathscr{F}$ are discontinuous.

Intuitively, Theorem 3 states that with probability approaching one we are able to perfectly distinguish values of θ that belong to Θ_0 from those that do not. This happens because Q_n is converging uniformly to Q at a rate faster than the rate at which τ_n approaches zero, while at the same time τ_n will eventually become smaller than δ, the size of the discrete jump. The result is that the contour sets $C_n(\tau_n)$ become identically equal to Θ_0 with probability approaching one.

Figure 3 illustrates the notion that, due to the discrete nature of Q_n, there are only a finite (though potentially very large) number of possible estimates $\hat{\Theta}_n$. For the realization of Q_n in the figure, the contour sets determine a partition of Θ into four disjoint sets: $\Theta = \Theta_n^1 \cup \Theta_n^2 \cup \Theta_n^3 \cup \Theta_n^4$. Our set estimates are upper contour sets of Q_n so there are four possible estimates: Θ_n^1, $\Theta_n^1 \cup \Theta_n^2$, $\Theta_n^1 \cup \Theta_n^2 \cup \Theta_n^3$, and $\Theta_n^1 \cup \Theta_n^2 \cup \Theta_n^3 \cup \Theta_n^4$. In higher dimensions, and for large sample sizes, the combinatorics of the problem dictate that the number of possibilities becomes large very quickly. On the other hand, as $n \to \infty$, the contour sets of Q_n approach those of Q, and the collection of possible estimates contains a set equal to Θ_0 with probability approaching one.

For a more concrete example, consider the population objective function for the maximum score estimator with discrete regressors:

$$Q(\theta) = \mathrm{E}_x \mathrm{E}_{y|x}\left[(2y-1)\operatorname{sgn}\left(x'\theta\right)\right] = \sum_{x\in\mathscr{X}} P(x)\left[2P(y=1 \mid x) - 1\right]\operatorname{sgn}\left(x'\theta\right). \tag{4}$$

The expectation becomes a sum of discontinuous functions of θ, so the population objective function is a step function in this setting. The size of the jump near the identified set—the value δ in Assumption A4'—is bounded below by the smallest nonzero value of $\left|P(x)\left[2P(y=1 \mid x)-1\right]\right|$ for some $x \in \mathscr{X}$. Hence, the constant majorant condition holds for Model 1 under Assumption C2 and we can apply Theorem 3 to show that $\hat{\Theta}_n$ converges arbitrarily fast to Θ_0.

Lemma 5. *Suppose that Assumptions C2 and C3 hold in Model 1. For any sequence τ_n such that $\tau_n \xrightarrow{\text{p}} 0$ and $n^{1/2}\tau_n \xrightarrow{\text{p}} \infty$, then for all positive sequences $r_n \to \infty$, $r_n d_{\text{H}}(\hat{\Theta}_n, \Theta_0) \xrightarrow{\text{p}} 0$.*

This is similar to Corollary 6.3 of Komarova (2013), who showed that in binary response models with discrete regressors a linear-programming-based estimator converges in probability in the Hausdorff metric at an arbitrarily fast rate. Hence, we have verified that the criterion-function-based estimator using the maximum score objective function also converges arbitrarily fast. Additionally, Theorem 3 isolates the source of this behavior and provides a condition (Assumption A4') under which other criterion-function-based estimators will have the same property. Komarova (2013) also showed that the maximum score objective functions provides a sharp characterization of the identified set when all regressors are discrete, considered inference on functions of the parameters (including individual components), and proposed several solutions to deal with model misspecification issues.

Arbitrarily fast rates of convergence arise in other areas of econometrics. The situation is similar to that of M-estimation with a finite parameter space, such as when estimating the sign of a parameter, where $\Theta = \{-1, 1\}$. For example, Andrews and Guggenberger (2008) illustrate a case where the rate of convergence of the least squares estimator in a nearly-unit-root AR(1) model is arbitrarily fast. Bhattacharya (2009) found an arbitrarily fast rate of convergence for a set estimator in the context of treatment assignment problems where a large number of individuals are optimally assigned to a finite number of treatments. As in other models, an alternative asymptotic framework could yield a more typical, polynomial rate of convergence.[4]

3.4. Confidence Sets

We now consider the problem of constructing a sequence of confidence sets B_n for which

$$\liminf_{n\to\infty} P(\Theta_0 \subseteq B_n) \geq 1 - \alpha \tag{5}$$

for a given value of α. This is a complex problem in general, if the sets B_n are not restricted to be members of a more tractable family of sets. As such, we focus on the case where B_n is a sequence of upper contour sets of Q_n. Under this restriction, the problem of choosing a sequence of arbitrary sets is reduced to that of choosing a sequence of levels κ_n.

Now, the coverage of a particular contour set can be inferred using the following statistic, for

[4] For example, one might consider a setting where the size of the jump at the boundary of the identified set in Assumption A4', say δ_n, is shrinking with the sample size at some rate. To achieve both consistency and a non-degenerate rate of convergence, it might also be necessary to let the support of the jump shrink towards the identified set at some rate, much like the majorant condition in Assumption A4.

some scaling sequence $b_n > 0$ (discussed below):

$$R_n \equiv b_n \left(\sup_{\Theta} Q_n - \inf_{\Theta_0} Q_n \right). \tag{6}$$

This statistic is of interest because it can be related directly to the coverage probability. For any sequence of levels κ_n we have

$$P\left(\Theta_0 \subseteq C_n(\kappa_n/b_n)\right) = P\left(\inf_{\Theta_0} Q_n \geq \sup_{\Theta} Q_n - \kappa_n/b_n \right) = P\left(R_n \leq \kappa_n\right).$$

Thus, coverage probabilities of contour sets are related to quantiles of the distribution of R_n. However, calculating these quantiles is problematic because Θ_0 is unknown.

To obtain confidence sets, we will use subsampling to approximate quantiles of the limiting distribution of R_n using the step-down procedure of Romano and Shaikh (2010, Algorithm 2.1). Their procedure requires neither an initial estimate of Θ_0 nor a slackness sequence. Our contribution is to establish that the appropriate scaling sequence is $b_n = n^{2/3}$ for the binary choice model under Assumption C1. Under Assumption C2, in light of the arbitrarily fast rate of convergence we recommend simply reporting the consistent set estimate itself.

Lemma 6. *In Model 1 under the assumptions of Lemma 4 and with $b_n = n^{2/3}$:*

i. *R_n converges in distribution to $R \equiv \sup_{\beta \in \mathrm{bd}(B_0), t \in \mathbb{R}^{K-1}} \left[W(\beta, t) - t' V(\beta) t \right]$ where for each $\beta \in \mathrm{bd}(B_0)$, $W(\beta, \cdot)$ is a mean zero Gaussian process with almost surely continuous sample paths and $V(\beta)$ is a positive definite matrix (defined in condition f of Lemma 4).*

ii. *Algorithm 2.1 of Romano and Shaikh (2010) yields confidence sets B_n satisfying (5) for any $\alpha < 1/2$.*

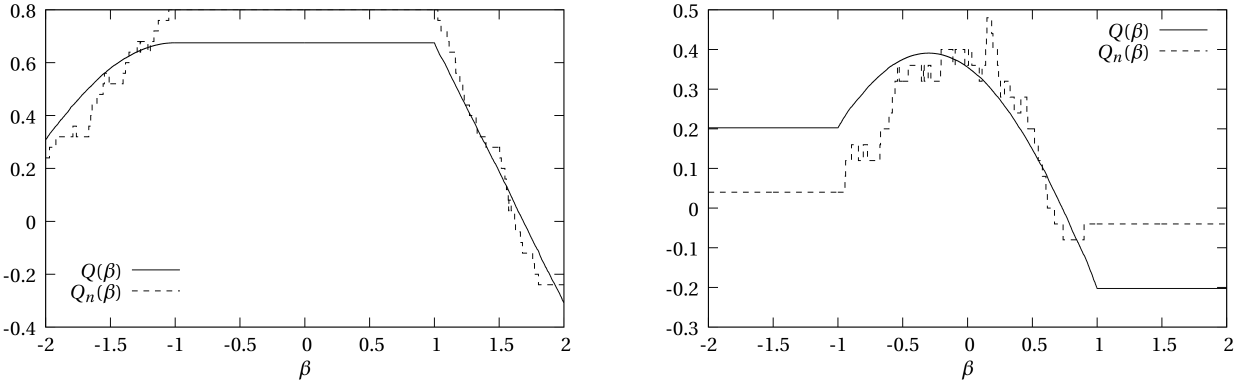
4. Monte Carlo Experiments

In this section, describe a series of Monte Carlo experiments[5] for three different specifications of the semiparametric binary choice model designed to both illustrate the asymptotic properties derived above and to shed light on the finite sample properties. All specifications have two regressors, the second of which is discrete in all cases. Specifications C1 and C2 have a continuous first regressor while Specification D1 has a discrete first regressor. The first coefficient is normalized to one in all cases, so we estimate β where $\theta = (1, \beta)$, $\beta \in B$ is a scalar, and $B \subset \mathbb{R}$ denotes the parameter space. Although β is a scalar here it would typically be a vector in practice.

Observations in our simulated samples are generated according to the model

$$y_i = 1\{x_{1i} + \beta_0 x_{2i} + u_i \geq 0\}$$

[5] Fortran programs to reproduce our results are available at `http://jblevins.org/research/cuberoot`.



(a) Specification C1

(b) Specification C2

FIGURE 4. Objective functions: Q and one realization of Q_n for $n = 50$.

where $u_i \sim \mathrm{N}(0,1)$. The parameter space is $B = [-2,2]$. We vary the distributions of x_1 and x_2 and the population parameter β_0 across the three specifications.

Specification C1, the first continuous-regressor specification, is a partially identified model where $x_1 \sim \mathrm{U}(1,2)$, $x_2 \sim \mathrm{U}(\{-1,1\})$, and $\beta_0 = -1$. That is, x_1 has a continuous uniform distribution on an interval and x_2 has a discrete uniform distribution on a finite set. The identified set for this specification is $B_0 = [-1,1]$. The population objective function for this model, Q, is plotted in Figure 4(a), along with one realization of the finite-sample objective function Q_n for $n = 50$.

Specification C2 is identical to Specification C1 with the exception that the population parameter is $\beta_0 = -0.3$. This results in the identified set $B_0 = \{-0.3\}$, a singleton. The objective function for this model is plotted in Figure 4(b). Clearly the population objective function is maximized at a single point while the finite-sample objective function is a step function as before.

For each specification, we generated $R = 1000$ simulated data sets for each sample size $n \in \{250, 2000, 16000, 128000, 1024000\}$. We use these datasets to produce set estimates and confidence sets. We consider both small and very large sample sizes as a means of providing simulation evidence for both the finite sample properties of the estimator as well as the asymptotic properties established above. We considered four choices for the slackness sequence τ_n. The conditions for consistency require that τ_n tends to zero no faster than $n^{-1/2}$. To achieve the fastest rate of convergence, we should choose a sequence τ_n with a rate of convergence close to $n^{-1/2}$, but otherwise the choice of τ_n is arbitrary (i.e., we are free to vary the constant of proportionality). For comparison, we choose three sequences that are proportional to $n^{-0.49}$. The fourth sequence is $\tau_n = 0$, which corresponds to set estimators obtained by simply maximizing the function (i.e., using a degenerate slackness sequence) that are not consistent in general.

We report estimates for Specifications C1 and C2 in Tables 1 and 2. The first column gives the slackness sequences used, which we describe in more detail below. The second column reports

TABLE 1. Estimates for Specification C1 for $R = 1000$ replications.

τ_n	n	Mean $\hat{B}_n$	St. Dev.	d_{H}
$n^{-0.49}$	250	[-1.372, 0.914]	[0.134, 0.523]	0.478
	2000	[-1.231, 0.999]	[0.064, 0.203]	0.248
	16000	[-1.136, 1.007]	[0.031, 0.002]	0.136
	128000	[-1.084, 1.002]	[0.015, 0.001]	0.084
	1024000	[-1.051, 1.001]	[0.007, 0.000]	0.051
$Sn^{-0.49}$	250	[-1.304, 0.673]	[0.130, 0.795]	0.597
	2000	[-1.187, 0.857]	[0.066, 0.540]	0.323
	16000	[-1.110, 0.988]	[0.032, 0.179]	0.125
	128000	[-1.067, 0.997]	[0.016, 0.090]	0.071
	1024000	[-1.041, 1.000]	[0.007, 0.000]	0.041
$Mn^{-0.49}$	250	[-1.274, 0.602]	[0.141, 0.843]	0.623
	2000	[-1.176, 0.844]	[0.071, 0.558]	0.323
	16000	[-1.106, 0.986]	[0.035, 0.190]	0.123
	128000	[-1.064, 0.997]	[0.018, 0.090]	0.068
	1024000	[-1.039, 1.000]	[0.009, 0.000]	0.039
0	250	[-1.130, -0.688]	[0.135, 0.859]	1.701
	2000	[-1.056, -0.808]	[0.108, 0.673]	1.814
	16000	[-1.030, -0.819]	[0.027, 0.621]	1.819
	128000	[-1.014, -0.726]	[0.014, 0.708]	1.726
	1024000	[-1.007, -0.609]	[0.007, 0.802]	1.609

Note: For $\beta_0 = -1$ the identified set is $B_0 = [-1, 1]$. Mean $\hat{B}_n$ denotes the average of the endpoints of $\hat{B}_n$, St. Dev. denotes the standard deviations of the endpoints, and d_{H} denotes the average Hausdorff distance $d_{\mathrm{H}}(\hat{B}_n, B_0)$.

the sample size, which is increased by factors of eight. The third column reports the means of the endpoints of the estimated intervals across all $R = 1000$ replications while the fourth column reports the standard deviations of those endpoints. The last column gives the average Hausdorff distance between the estimated sets and the identified set, $d_{\mathrm{H}}(\hat{B}_n, B_0)$.

The first slackness sequence used is $\tau_n = n^{-0.49}$, where the constant of proportionality is one. The second and third slackness sequences, which we refer to as $Sn^{-0.49}$ and $Mn^{-0.49}$ are chosen by selecting the constants of proportionality to be equal to, respectively, the supremum of the functional values over the parameter space and the median of the differences relative to the maximum value. In practice, we use $S \equiv \sup_{\beta \in B} Q_n(\beta)$ and $M \equiv \sup_{\beta \in B} Q_n(\beta) - \min\{Q_n(\beta^j)\}_{j=1}^{J}$, where the J grid points β^j are uniformly spaced on B. These choices adapt the scale of the slackness sequence to the scale of the functional values.[6]

[6] To give a better sense of the relative magnitudes of these sequences, here we report the average values of each sequence across the $R = 1000$ experiments for $n = 250$. For this sample size, $n^{-0.49} = 0.067$ is the same for all three specifications. The average values of the other sequences are, for C1, $\overline{S}n^{-0.49} = 0.047$ and $\overline{M}n^{-0.49} = 0.067$, for C2, $\overline{S}n^{-0.49} = 0.029$ and $\overline{M}n^{-0.49} = 0.042$, and for D1, $\overline{S}n^{-0.49} = 0.046$ and $\overline{M}n^{-0.49} = 0.036$.

TABLE 2. Estimates for Specification C2 for $R = 1000$ replications.

τ_n	n	Mean $\hat{B}_n$	St. Dev.	d_{H}
$n^{-0.49}$	250	[-0.636, 0.031]	[0.198, 0.171]	0.441
	2000	[-0.513, -0.091]	[0.073, 0.080]	0.258
	16000	[-0.430, -0.172]	[0.036, 0.035]	0.151
	128000	[-0.381, -0.220]	[0.017, 0.017]	0.091
	1024000	[-0.350, -0.251]	[0.008, 0.007]	0.054
$Sn^{-0.49}$	250	[-0.466, -0.131]	[0.178, 0.183]	0.293
	2000	[-0.412, -0.195]	[0.086, 0.091]	0.168
	16000	[-0.368, -0.232]	[0.044, 0.040]	0.095
	128000	[-0.344, -0.256]	[0.020, 0.020]	0.057
	1024000	[-0.328, -0.272]	[0.009, 0.009]	0.034
$Mn^{-0.49}$	250	[-0.384, -0.208]	[0.196, 0.177]	0.223
	2000	[-0.366, -0.243]	[0.093, 0.093]	0.128
	16000	[-0.340, -0.259]	[0.045, 0.044]	0.071
	128000	[-0.327, -0.273]	[0.022, 0.022]	0.042
	1024000	[-0.318, -0.282]	[0.010, 0.010]	0.025
0	250	[-0.325, -0.269]	[0.189, 0.188]	0.173
	2000	[-0.313, -0.297]	[0.094, 0.095]	0.083
	16000	[-0.301, -0.297]	[0.047, 0.047]	0.039
	128000	[-0.300, -0.299]	[0.024, 0.024]	0.020
	1024000	[-0.300, -0.300]	[0.011, 0.011]	0.009

Note: For $\beta_0 = -0.3$ the identified set is $B_0 = \{-0.3\}$. Mean $\hat{B}_n$ denotes the average of the endpoints of $\hat{B}_n$, St. Dev. denotes the standard deviations of the endpoints, and d_{H} denotes the average Hausdorff distance $d_{\mathrm{H}}(\hat{B}_n, B_0)$.

If the estimator is consistent for a particular choice of τ_n, then the Hausdorff distance should converge to zero and the sequence of set estimates should converge to the identified set. All three of the nonzero sequences are guaranteed to produce consistent estimates, but their finite sample properties differ. Neither sequence appears to be uniformly better than the others. In small samples, the sequence $\tau_n = n^{-0.49}$ performs worst for Specifications C2 and D1. The sequence $\tau_n = Mn^{-0.49}$ performs worst for Specification C1 in small samples but best for large samples. As expected, although the zero sequence yields consistent estimates for the point-identified Specification C2 it results in inconsistent estimates for Specification C1. The sequence $Sn^{-0.49}$ appears to be the most robust across the three specifications we consider, followed closely by $Mn^{-0.49}$. Yet, the fact that the performance is sensitive to the choice of τ_n is further motivation for simply reporting confidence sets which can be calculated without using a slackness sequence.

According to our theoretical results for the continuous-regressor specifications, when the slackness sequence is such that the estimator is consistent, it should converge at essentially

TABLE 3. Coverage of confidence sets for Specifications C1 and C2 for $R = 1000$ replications.

Sample Size	Specification C1				Specification C2			
n	0.500	0.750	0.900	0.990	0.500	0.750	0.900	0.990
250	0.545	0.785	0.913	0.992	0.624	0.846	0.956	0.999
2000	0.576	0.798	0.914	0.989	0.601	0.833	0.953	0.995
16000	0.522	0.744	0.917	0.995	0.568	0.821	0.935	0.995
128000	0.473	0.771	0.887	0.989	0.538	0.774	0.902	0.987
1024000	0.502	0.795	0.894	0.988	0.613	0.794	0.912	0.994

Note: Columns represent different nominal coverage levels $1-\alpha \in \{0.5, 0.75, 0.9, 0.99\}$.

the rate $n^{-1/3}$. As a rule of thumb,[7] with cube-root consistency when increasing the sample size by factors of eight the Hausdorff distance should decrease by approximately half for the sequences proportional to $n^{-0.49}$. For Specification C1, reported in Table 1, we find that indeed, the estimates appear to be consistent approximately at the cube root rate when the slackness sequence is used. The estimator appears to be inconsistent without the slackness sequence.

For Specification C2, which is actually point identified, we can see from Table 2 that the estimator is consistent even for $\tau_n = 0$, which is expected given the consistency of the maximum score point estimator. The benefits of formally treating the maximum score estimator as a set estimator are the additional robustness to cases where the support condition may not be satisfied and the avoidance of an ad hoc selection rule for choosing a point from the set that maximizes the sample criterion function. The entire set is treated as the estimate since there is usually no a priori reason to prefer any particular point. As before, for sequences proportional to $\tau_n = n^{-0.49}$ we can see that the estimator achieves approximate cube-root convergence since increasing the sample size eight-fold roughly halves the average Hausdorff distance.

Next, we evaluate the performance of the step-down procedure of Romano and Shaikh (2010). These results are reported in Table 3 for both Specifications C1 and C2. For each sample size and each simulated sample we approximate the limiting distribution of R_n using Algorithm 2.1 of Romano and Shaikh (2010) with 1000 subsamples. For sample sizes $n = 250, 2000, 16000, 1024000$ we use subsample sizes $m = 10, 20, 30, 40$, respectively.[8] We then choose the appropriate quantile for each nominal level $1-\alpha \in \{0.50, 0.75, 0.90, 0.99\}$. We obtain 1000 confidence sets for each level (one for each simulated sample) and report the coverage frequencies.

In practice, choosing the block size m for subsampling is important to achieve the nominal coverage desired. We experimented with several other choices for the block size. The results

[7] Let $\overline{d}_{\mathrm{H},n}$ denote the average Hausdorff distance between $\hat{\Theta}_n$ and Θ_0. Approximate cube-root consistency implies $n^{1/3}\overline{d}_{\mathrm{H},n} \approx C$ for some constant C. Solving for $\overline{d}_{\mathrm{H},n}$ and comparing for n and $8n$ yields $\overline{d}_{\mathrm{H},8n} = \frac{1}{2}\overline{d}_{\mathrm{H},n}$.

[8] When the sample size n increases by a factor of 8 the subsample size m only increases linearly so the required condition, $m \to \infty$ and $m/n \to 0$, is satisfied for the sequence chosen.

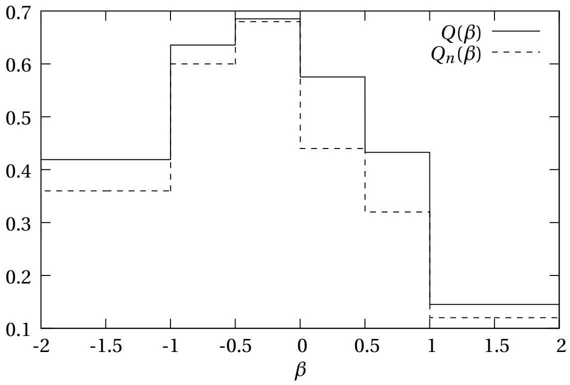


FIGURE 5. Objective functions: Q and one realization of Q_n for Specification D1 for $n = 50$.

for the sequence $m = 5, 10, 15, 20$ were similar to those reported (for $m = 10, 20, 30, 40$). For sequences with larger values (e.g., $m = 50, 200, 800, 3200$) the resulting confidence sets tended to become more conservative, thus still satisfying (5). In practice, one can also allow the block size to be data-dependent as discussed by Politis, Romano, and Wolf (1999, Chapter 9).

Finally, we consider *Specification D1* which has only discrete regressors. The regressors are distributed as $x_1 \sim \mathrm{U}(\{-2, -1, 0, 1, 2\})$, $x_2 \sim \mathrm{U}(\{-2, -1, 0, 1, 2\})$, and $\beta_0 = -0.3$. In this case, the identified set is $B_0 = [-0.5, 0]$. The objective function is plotted in Figure 5 and the estimates are reported in Table 4. As expected, we have arbitrarily fast convergence, with the Hausdorff distance being essentially zero for sample sizes $n \geq 2000$.

5. Conclusion

This paper has developed several asymptotic properties for criterion-function-based set estimators for semiparametric binary choice models without the need to impose support conditions on the regressors. We also provide new sufficient conditions for estimators of this kind in more general models which may exhibit non-standard behavior, such as cube-root consistency or arbitrarily fast consistency. A series of Monte Carlo results illustrates the theoretical results and provides insights into the practical finite sample behavior of the estimator.

Our results provide a basis for deriving the properties of set estimators for other models and suggest several areas for future work in this literature. For example, although we do not consider it explicitly here, the results could be applied to the closely-related multinomial choice model of Manski (1975) and other models based on similar rank conditions. Finally, our Monte Carlo results suggest that further study of data-driven procedures for selecting the constant of proportionality in the slackness sequence is an important topic for future work.

TABLE 4. Estimates for Specification D1 for $R = 1000$ replications.

τ_n	n	Mean $\hat{B}_n$	St. Dev.	d_{H}
$n^{-0.49}$	250	[-0.852, 0.104]	[0.228, 0.203]	0.387
	2000	[-0.506, 0.000]	[0.059, 0.000]	0.007
	16000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	128000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	1024000	[-0.500, 0.000]	[0.000, 0.000]	0.000
$Sn^{-0.49}$	250	[-0.734, 0.046]	[0.251, 0.148]	0.263
	2000	[-0.501, 0.000]	[0.027, 0.000]	0.002
	16000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	128000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	1024000	[-0.500, 0.000]	[0.000, 0.000]	0.000
$Mn^{-0.49}$	250	[-0.625, 0.021]	[0.223, 0.131]	0.149
	2000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	16000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	128000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	1024000	[-0.500, 0.000]	[0.000, 0.000]	0.000
0	250	[-0.543, -0.021]	[0.160, 0.140]	0.059
	2000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	16000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	128000	[-0.500, 0.000]	[0.000, 0.000]	0.000
	1024000	[-0.500, 0.000]	[0.000, 0.000]	0.000

Note: For $\beta_0 = -0.3$ the identified set is $B_0 = [-0.5, 0]$. Mean $\hat{B}_n$ denotes the average of the endpoints of $\hat{B}_n$, St. Dev. denotes the standard deviations of the endpoints, and d_{H} denotes the average Hausdorff distance $d_{\mathrm{H}}(\hat{B}_n, B_0)$.

References

Abrevaya, J. (2000). Rank estimation of a generalized fixed-effects regression model. *Journal of Econometrics 95*, 1–23.

Abrevaya, J. and J. Huang (2005). On the bootstrap of the maximum score estimator. *Econometrica 73*, 1175–1204.

Amemiya, T. (1985). *Advanced Econometrics.* Harvard University Press.

Andrews, D. and P. Guggenberger (2008). Asymptotics for stationary very nearly unit root processes. *Journal of Time Series Analysis 29*, 203–212.

Andrews, D. W. K. and P. J. Barwick (2012). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica 80*, 2805–2826.

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

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

Andrews, D. W. K. and G. Soares (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica 78*, 119–158.

Bajari, P., J. T. Fox, and S. P. Ryan (2008). Evaluating wireless carrier consolidation using semiparametric demand estimation. *Quantitative Marketing and Economics 6*, 299–338.

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

Beresteanu, A. and F. Molinari (2008). Asymptotic properties for a class of partially identified models. *Econometrica 76*, 763–814.

Bhattacharya, D. (2009). Inferring optimal peer assignment from experimental data. *Journal of the American Statistical Association 104*, 486–500.

Bierens, H. J. and J. Hartog (1988). Non-linear regression with discrete explanatory variables, with an application to the earnings function. *Journal of Econometrics 38*, 269–299.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica 78*, 735–753.

Canay, I. A. (2010). EL inference for partially identified models: Large deviations optimality and bootstrap validity. *Journal of Econometrics 156*, 408–425.

Chernozhukov, V., H. Hong, and E. Tamer (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica 75*, 1243–1284.

Cosslett, S. R. (1983). Distribution-free maximum likelihood estimation of the binary choice model. *Econometrica 51*, 765–782.

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

Dudley, R. M. (1987). Universal Donsker classes and metric entropy. *Annals of Probability 15*, 1306–1326.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics 35*, 303–316.

Honoré, B. E. and A. Lleras-Muney (2006). Bounds in competing risks models and the war on cancer. *Econometrica 74*, 1675–1698.

Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica 74*, 611–629.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica 60*, 505–531.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics.* New York: Springer.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics 58*, 71–120.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*, 1845–1857.

Jun, S., J. Pinkse, and Y. Wan (2011). Classical Laplace estimation for $\sqrt[3]{n}$-consistent estimators: Improved convergence rates and rate-adaptive inference. Working paper, Pennsylvania State University.

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

Kim, J. and D. Pollard (1990). Cube root asymptotics. *The Annals of Statistics 18*, 191–219.

Kim, K. (2008). Set estimation and inference with models characterized by conditional moment inequalities. Working paper, University of Minnesota.

Klein, R. W. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica 61*, 387–421.

Komarova, T. (2013). Binary choice models with discrete regressors: Identification and misspecification. *Journal of Econometrics 177*, 14–33.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroskedasticity or instrumental variables. *Journal of Econometrics 97*, 145–177.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press.

Magnac, T. and E. Maurin (2008). Partial identification in monotone binary models: Discrete regressors and interval data. *Review of Economic Studies 75*, 835–864.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*, 205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics 27*, 313–333.

Manski, C. F. (1988). Identification of binary response models. *Journal of the American Statistical Association 83*, 729–738.

Manski, C. F. (2003). *Partial Identification of Probability Distributions.* Springer-Verlag.

Manski, C. F. and E. Tamer (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica 70*, 519–546.

Matzkin, R. L. (1992). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*, 239–270.

McFadden, D. L. (1974). Conditional logit analysis of qualitative choice analysis. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press.

Menzel, K. (2014). Consistent estimation with many moment inequalities. *Journal of Econometrics 182*, 329–350.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In

R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, Amsterdam. New Holland.

Nolan, D. and D. Pollard (1987). U-Processes: Rates of convergence. *Annals of Statistics 15*, 780–799.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–1057.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2011). Moment inequalities and their application. Working paper, Harvard University.

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

Pollard, D. (1989). Asymptotics via empirical processes. *Statistical Science 4*, 341–366.

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

Romano, J. P. and A. M. Shaikh (2010). Inference for the identified set in partially identified econometric models. *Econometrica 78*, 169–211.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica 54*, 1461–1481.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica 77*, 1299–1315.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics 2*, 167–195.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Wan, Y. and H. Xu (2015). Inference in semiparametric binary response models with interval data. *Journal of Econometrics 184*, 347–360.

Yildiz, N. (2012). Consistency of plug-in estimators of upper contour and level sets. *Econometric Theory 28*, 309–327.

A. Proofs of Results for General Models

We begin with some definitions and a preliminary result regarding absolute values of functions. Let B^{c} denote the complement of a set B in Θ. In a slight abuse of notation, we also write B^{ε} to denote an ε-expansion of a set B in Θ, defined as $B^{\varepsilon} \equiv \{\theta \in \Theta : \rho(\theta, B) \leq \varepsilon\}$. We write $a \vee b$ to denote $\max\{a, b\}$ and $a \wedge b$ to denote $\min\{a, b\}$.

Lemma 7. *Let f and g be bounded real functions on $A \subset \mathbb{R}^n$. Then*

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

Proof of Lemma 7.

$$\begin{aligned}\sup_{x\in A}\left|f(x)-g(x)\right| &= \sup_{x\in A}\left[(f(x)-g(x))\vee(g(x)-f(x))\right]\\ &= \sup_{x\in A}[f(x)-g(x)]\vee\sup_{x\in A}[g(x)-f(x)]\\ &\geq \sup_{x\in A}\left[f(x)-\sup_{y\in A}g(y)\right]\vee\sup_{x\in A}\left[g(x)-\sup_{y\in A}f(y)\right]\\ &= \left[\sup_{x\in A}f(x)-\sup_{x\in A}g(x)\right]\vee\left[\sup_{x\in A}g(x)-\sup_{x\in A}f(x)\right]\\ &= \left|\sup_{x\in A}f(x)-\sup_{x\in A}g(x)\right|.\end{aligned}$$

■

Proof of Theorem 1. The proof proceeds in two steps. We first show that $\sup_{\theta\in\hat{\Theta}_n}\rho(\theta,\Theta_0)\xrightarrow{\text{p}}0$. Then, we show that $\lim_{n\to\infty}P(\Theta_0\subset\hat{\Theta}_n)=1$, which implies that $\sup_{\theta\in\Theta_0}\rho(\theta,\hat{\Theta}_n)\xrightarrow{\text{p}}0$. Combining these steps and using the definition of the Hausdorff distance yields the final result.

Step 1 Let $\eta>0$ and $\varepsilon>0$ be given. Uniform convergence in probability of Q_n to Q over Θ (Assumption A3) also implies uniform convergence in probability over $\Theta\setminus\Theta_0^\eta$, so we have both $\sup_\Theta|Q_n-Q|=O_p(1/a_n)$ and $\sup_{\Theta\setminus\Theta_0^\eta}|Q_n-Q|=O_p(1/a_n)$. Under Assumption A2, there exists a $\delta_\eta>0$ such that $\sup_{\Theta\setminus\Theta_0^\eta}Q\leq\sup_\Theta Q-\delta_\eta$. Combining the above, we have

$$\sup_{\Theta\setminus\Theta_0^\eta}Q_n\leq\sup_{\Theta\setminus\Theta_0^\eta}Q+O_p(1/a_n)\leq\sup_\Theta Q-\delta_\eta+O_p(1/a_n)\leq\sup_\Theta Q_n-\delta_\eta+O_p(1/a_n).$$

Recall that by definition of $\hat{\Theta}_n$, $\inf_{\hat{\Theta}_n}Q_n\geq\sup_\Theta Q_n-\tau_n$. Since $\delta_\eta>0$ is constant and $\tau_n=o_p(1)$, there exists an integer n_ε such that for all $n\geq n_\varepsilon$ with probability at least $1-\varepsilon$ both the $O_p(1/a_n)$ term and τ_n are smaller than $\delta_\eta/2$ and so $-\delta_\eta+O_p(1/a_n)<-\delta_\eta/2<-\tau_n$. Therefore, for $n\geq n_\varepsilon$, we have $\inf_{\hat{\Theta}_n}Q_n>\sup_{\Theta\setminus\Theta_0^\eta}Q_n$ which implies $\hat{\Theta}_n\subseteq\Theta_0^\eta$ which in turn implies $\sup_{\theta\in\hat{\Theta}_n}\rho(\theta,\Theta_0)\leq\eta$, all with probability at least $1-\varepsilon$. Since ε and η were arbitrary, $\sup_{\theta\in\hat{\Theta}_n}\rho(\theta,\Theta_0)\xrightarrow{\text{p}}0$.

Step 2 By definition of $\hat{\Theta}_n$, if $\sup_\Theta Q_n-\inf_{\Theta_0}Q_n<\tau_n$, then $\Theta_0\subseteq\hat{\Theta}_n$. We have

$$\begin{aligned}\sup_\Theta Q_n-\inf_{\Theta_0}Q_n &= \left[\sup_\Theta Q_n-\sup_\Theta Q\right]+\left[\sup_\Theta Q-\inf_{\Theta_0}Q_n\right]\\ &\leq \left|\sup_\Theta Q_n-\sup_\Theta Q\right|+\left|\sup_\Theta Q-\inf_{\Theta_0}Q_n\right|\\ &= \left|\sup_\Theta Q_n-\sup_\Theta Q\right|+\left|\inf_{\Theta_0}Q-\inf_{\Theta_0}Q_n\right|\\ &\leq \sup_\Theta|Q_n-Q|+\sup_{\Theta_0}|Q_n-Q|\\ &\leq 2\sup_\Theta|Q_n-Q|.\end{aligned}$$

These steps follow by (1) adding and subtracting $\sup_\Theta Q$, (2) taking the absolute value, (3) noting that Θ_0 maximizes Q, (4) using the fact that $\inf f = -\sup(-f)$ and applying Lemma 7 twice, and (5) recalling that $\Theta_0 \subseteq \Theta$. By Assumption A3, $2\sup_\Theta |Q_n - Q| = O_p(1/a_n)$. Finally, the condition that τ_n approaches zero in probability slower than $1/a_n$ implies $\sup_\Theta Q_n - \inf_{\Theta_0} Q_n < \tau_n$ with probability approaching one. ■

Proof of Theorem 2. Let $\varepsilon > 0$ be given and let δ, c, γ_1, γ_2, c_ε, and n_ε satisfy Assumption A4. Let a_n satisfy Assumption A3 and define

$$\nu_n \equiv \left(\frac{c_1 c_\varepsilon \vee 2\tau_n a_n}{a_n c_1}\right)^{1/\gamma_1}.$$

There exists an $n'_\varepsilon \geq n_\varepsilon$ such that for all $n \geq n'_\varepsilon$, with probability at least $1-\varepsilon$ each of the following are true: (a) $\nu_n \geq (c_\varepsilon/a_n)^{\gamma_2}$, (b) $\nu_n \leq \delta$, and (c) $\sup_\Theta |Q_n - Q| \leq \tau_n$. Condition (a) holds since $\gamma_1\gamma_2 \geq 1$ and so for sufficiently large n with probability at least $1-\varepsilon$

$$\nu_n^{1/\gamma_2} \geq \left(\frac{c_\varepsilon}{a_n}\right)^{\frac{1}{\gamma_1\gamma_2}} \geq \frac{c_\varepsilon}{a_n}.$$

Condition (b) follows because $\nu_n = o_p(1)$, due to the assumptions on τ_n and a_n, and the fact that δ is a strictly positive constant. Condition (c) follows from Assumption A3, under which $\sup_\Theta |Q_n - Q| = O_p(1/a_n)$, and the condition $a_n \tau_n \xrightarrow{\text{p}} \infty$. Therefore, for all $n \geq n'_\varepsilon$, with probability at least $1-\varepsilon$,

$$\sup_{\Theta\setminus\Theta_0^{\nu_n}} Q_n \underset{(1)}{<} \sup_\Theta Q - c_1(\nu_n \wedge \delta)^{\gamma_1} \underset{(2)}{\leq} \sup_\Theta Q - c_1\nu_n^{\gamma_1} \underset{(3)}{\leq} \sup_\Theta Q - 2\tau_n \underset{(4)}{\leq} \sup_\Theta Q_n - \tau_n \underset{(5)}{\leq} \inf_{\hat\Theta_n} Q_n.$$

Inequality (1) holds by Assumptions A2 and A4 and condition (a) under which $\rho(\theta,\Theta_0) > \nu_n \geq (c_\varepsilon/a_n)^{\gamma_2}$ for $\theta \in \Theta \setminus \Theta_0^{\nu_n}$. Inequality (2) is a direct result of condition (b). Inequality (3) holds by definition of ν_n, since $c_1\nu_n^{\gamma_1} \geq 2\tau_n$. Inequality (4) follows from condition (c). Inequality (5) follows by definition of $\hat\Theta_n$. It follows that for $n \geq n'_\varepsilon$, with probability at least $1-\varepsilon$, the set $\hat\Theta_n \cap (\Theta \setminus \Theta_0^{\nu_n})$ is empty, or equivalently, $\hat\Theta_n \subseteq \Theta_0^{\nu_n}$. Finally, recall that by Theorem 1 we have $\lim_{n\to\infty} P(\Theta_0 \subseteq \hat\Theta_n) = 1$. Therefore, for all $n \geq n'_\varepsilon$, with probability at least $1-\varepsilon$, $d_{\text{H}}(\hat\Theta_n,\Theta_0) \leq \nu_n$ and hence $d_{\text{H}}(\hat\Theta_n,\Theta_0) = O_p(\tau_n^{\gamma_2})$. ■

Proof of Theorem 3. First, note that that Assumption A4' implies Assumption A2. Then, from Theorem 1 we have $\lim_{n\to\infty} P(\Theta_0 \subseteq \hat\Theta_n) = 1$. We will prove the result by showing that $\lim_{n\to\infty} P(\hat\Theta_n \subseteq \Theta_0) = 1$ and therefore the Hausdorff distance $d_{\text{H}}(\hat\Theta_n,\Theta_0)$ equals zero with probability approaching one. The logic is very similar to that used in Step 1 of the proof of Theorem 1, but without expanding the set Θ_0. We have

$$\sup_{\Theta\setminus\Theta_0} Q_n \leq \sup_{\Theta\setminus\Theta_0} Q + O_p(1/a_n) \leq \sup_\Theta Q - \delta + O_p(1/a_n) \leq \sup_\Theta Q_n - \delta + O_p(1/a_n),$$

where the first and last inequalities follow from Assumption A3 and the middle inequality follows from Assumption A4', the constant majorant condition. Since $\tau_n = o_p(1)$ and $\delta > 0$ is constant, with probability approaching one we have $\tau_n < \delta/2$ and $O_p(1/a_n) < \delta/2$, leading to $\sup_{\Theta\setminus\Theta_0} Q_n < \sup_\Theta Q_n - \tau_n \leq \inf_{\hat\Theta_n} Q_n$, and therefore, $\hat\Theta_n \subseteq \Theta_0$. ■

B. Sufficient Conditions

This section derives sufficient conditions which may be easier to verify than the conditions of the theorems in the main text. Many of these conditions are stated in terms of empirical process concepts—they are restrictions on the indexing class of functions which generate the finite sample and population objective functions. We briefly summarize some standard notation and definitions below, but refer the reader to Section 2 of Pakes and Pollard (1989) for details.

Let P be the joint distribution of all observables, denoted Z. We shall maintain Assumption C3, that n iid observations of Z are available for use in estimation. Let P_n denote the associated empirical measure. Let $\ell^\infty(B)$ denote the space of uniformly bounded real-valued functions $f : B \to \mathbb{R}$ on an arbitrary set B endowed with the uniform metric $d_\infty(f,g) = \sup_{b\in B} \left|f(b) - g(b)\right|$ for $f, g \in \ell^\infty(B)$.

We focus here on models for which the objective functions can be expressed in terms of a class of real-valued functions $\mathscr{F}$, where for each $\theta \in \Theta$, $Q(\theta) = Pf(\cdot,\theta)$ and $Q_n(\theta) = P_n f(\cdot,\theta)$ for $f(\cdot,\theta) \in \mathscr{F}$. As such, we work with empirical processes indexed by classes of functions $\mathscr{F} = \{f(\cdot,\theta) : \theta \in \Theta\}$. Alternatively, we use parameter space Θ as the indexing set when convenient. Note that this assumption does not include objective functions of the kind considered by CHT and Bugni (2010), where $Q = \left\|Pf(\cdot,\theta)\right\|_{W(\theta)}$ for some appropriate weighting matrix $W(\theta)$, or any of the functions proposed by Andrews and Soares (2010) and related papers. Additionally, the objective functions we consider are not scale invariant, which can be problematic in moment inequality models (cf. Andrews and Soares, 2010, Assumption 1(b)).

Each such model has a different indexing class $\mathscr{F}$. An *envelope* for $\mathscr{F}$ is a function F such that $\sup_{\mathscr{F}} \left|f\right| \leq F$. Let $\mathbb{G}_n = \sqrt{n}(P_n - P)$ denote the standardized empirical process indexed by $\mathscr{F}$. Note that P, P_n, and $\mathbb{G}_n$ all map classes $\mathscr{F}$ to functions in $\ell^\infty(\Theta)$. We work under conditions below, such as manageability of $\mathscr{F}$, that are sufficient for $\mathscr{F}$ to be P-Donsker, meaning that $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathscr{F})$, where $\rightsquigarrow$ denotes weak convergence and $\mathbb{G}$ is a mean-zero Gaussian process indexed by $\mathscr{F}$ with almost surely continuous sample paths and covariance $\mathrm{E}[f(Z)g(Z)] - \mathrm{E}[f(Z)] \cdot \mathrm{E}[g(Z)]$ for all $f, g \in \mathscr{F}$.

B.1. Sufficient Conditions for Consistency

Assumption B1. Θ is a nonempty, compact subset of $\mathbb{R}^K$ and there exists a class of real-valued functions $\mathscr{F} = \{f(\cdot,\theta) : \theta \in \Theta\}$ such that $Q(\theta) = Pf(\cdot,\theta)$ and $Q_n(\theta) = P_n f(\cdot,\theta)$ for all $\theta \in \Theta$.

Assumption B2. Q is piecewise continuous on Θ.

Assumption B3. $\mathscr{F}$ is manageable for some envelope F such that $PF^2 < \infty$.

Lemma 8. *Suppose that Assumptions B1–B3 hold. Then Assumptions A1–A3 hold with* $a_n = n^{1/2}$.

Proof of Lemma 8. Compactness of Θ implies Assumption A1 and piecewise continuity of Q implies Assumption A2. Since $\mathscr{F}$ is manageable with $PF^2 < \infty$, it follows from Corollary 3.2 of Kim and Pollard (1990) that Assumption A3 holds with $a_n = n^{1/2}$. ■

Hence, under Assumptions B1–B3, $\hat{\Theta}_n$ is consistent in the sense of Theorem 1. Under piecewise continuity (Assumption B2), Q may have only a finite number of pieces. This property is not always immediate, but it can be shown in some cases using the uniform law of large numbers when $f(\cdot,\theta)$ is continuous in θ with probability one and dominated by some bounded function F (Newey and McFadden, 1994, Lemma 2.4). Furthermore, Assumption B3 is satisfied in models where $\mathscr{F}$ is a Vapnik-Chervonenkis (VC) subgraph class, in the sense of Dudley (1987), with constant envelope $F < \infty$. In particular, if $\mathscr{F}$ is a class of functions such that $\{\text{subgraph}(f) : f \in \mathscr{F}\}$ is a VC class of sets and $\sup_{\mathscr{F}} |f| \le F < \infty$, then $\mathscr{F}$ is necessarily manageable and $PF^2 < \infty$. The following lemma formalizes these results.

Lemma 9. *Suppose that Assumption B1 holds. If $\mathscr{F}$ is a VC subgraph class such that $|f(\cdot,\theta)| \le M$ for all $\theta \in \Theta$ for the constant function $M < \infty$, then Assumption B3 holds. In addition, if $f(\cdot,\theta)$ is continuous in θ with probability one, then Assumption B2 holds.*

Proof of Lemma 9. Since $\mathscr{F}$ is a VC subgraph class, Lemma 2.12 of Pakes and Pollard (1989) implies that $\mathscr{F}$ is Euclidean in the sense of Nolan and Pollard (1987, Definition 8) for any valid envelope including $F = M$. Since $\mathscr{F}$ is Euclidean, it is also manageable for $F = M$ (cf. Pakes and Pollard, 1989, p. 1033). Since $PF^2 = M^2 < \infty$, this verifies Assumption B3. Furthermore, if $f(\cdot,\theta)$ is continuous in θ with probability one, since it is dominated by $F = M$ for all θ, continuity of Q follows from Lemma 2.4 of Newey and McFadden (1994), verifying Assumption B2. ■

B.2. Sufficient Conditions for Cube Root Convergence

Assumption B4. There exists a neighborhood $\Theta_0^{\eta_1}$ of Θ_0 with $\eta_1 > 0$ and a positive constant c such that $Q(\theta) \le \sup_\Theta Q - c\rho^2(\theta,\Theta_0)$ for all $\theta \in \Theta_0^{\eta_1}$.

Assumption B5. There exists a positive constant η_2 such that for all $\eta \le \eta_2$, the classes $\mathscr{F}_\eta \equiv \{f(\cdot,\theta) : \rho(\theta,\Theta_0) \le \eta\}$ are uniformly manageable with $PF_\eta^2 = O(\eta)$, where $F_\eta(\cdot) \equiv \sup_{\mathscr{F}_\eta} |f(\cdot,\theta)|$ is the natural envelope of $\mathscr{F}_\eta$.

Theorem 4. *Suppose that Assumptions B1-B5 hold. Let $r_n = o(n^{1/3})$ and choose $\tau_n \propto r_n^{-3/2}$. Then $d_{\mathrm{H}}(\hat{\Theta}_n, \Theta_0) = O_p(r_n^{-1})$.*

Proof of Theorem 4. We will verify the conditions of Theorem 2. By Lemma 8, Assumptions B1–B3 imply Assumptions A1–A3 with $a_n = n^{1/2}$. It remains to verify Assumption A4. We will use the following notational conventions: η's denote distances in Θ, δ's denote distances between functional values, ε's denote arbitrarily small probabilities, and κ's denote various constants.

By definition of $\mathbb{G}_n(\theta)$, we can always write

$$Q_n(\theta) = (P_n - P)f(\cdot,\theta) + Pf(\cdot,\theta) = n^{-1/2}\mathbb{G}_n(\theta) + Q(\theta). \tag{7}$$

Let η_1 and η_2 satisfy Assumptions B4 and B5 and choose η to be smaller than the minimum of η_1 and η_2. Then from Assumption A2 there exists a $\delta_\eta > 0$ such that

$$\sup_{\Theta\setminus\Theta_0^\eta} Q \le \sup_\Theta Q - 2\delta_\eta. \tag{8}$$

Combining (7) and (8) and using Assumption A3 gives, for all $\theta \in \Theta \setminus \Theta_0^\eta$,

$$Q_n(\theta) \leq n^{-1/2}\mathbb{G}_n(\theta) + \sup_\Theta Q - 2\delta_\eta.$$

$\mathscr{F}$ is P-Donsker by Assumption B3 and since $\sup_\Theta |\cdot|$ is continuous in $\ell^\infty(\Theta)$, $\sup_\Theta |\mathbb{G}_n(\theta)| = O_p(1)$ by the continuous mapping theorem. It follows that for any $\varepsilon_1 \in (0,1)$ there exists an n_1 such that for all $n \geq n_1$,

$$Q_n(\theta) \leq \sup_\Theta Q - \delta_\eta \tag{9}$$

uniformly on $\Theta \setminus \Theta_0^\eta$ with probability at least $1-\varepsilon_1$.

Now, by Assumption B4, there is a neighborhood $\Theta_0^{\eta_1}$ of Θ_0 such that Q is approximately quadratic in the directed distance $\rho(\theta,\Theta_0)$. That is, for some $\kappa_1 > 0$, $Q(\theta) \leq \sup_\Theta Q - \kappa_1\rho^2(\theta,\Theta_0)$ for all $\theta \in \Theta_0^{\eta_1}$. Similarly, by Assumption B5 and Lemma 4.1 of Kim and Pollard (1990), for all $\kappa_2 > 0$ there exists a sequence of random variables $M_n = O_p(1)$ such that

$$\left|(P_n - P)f(\cdot,\theta)\right| \leq \kappa_2\rho^2(\theta,\Theta_0) + n^{-2/3}M_n^2 \tag{10}$$

for $\theta \in \Theta_0^{\eta_2}$. Combining these results for $\kappa_2 = \kappa_1/2$ and using (7) and Assumption A3 yields

$$Q_n(\theta) \leq \sup_\Theta Q - \frac{\kappa_1}{2}\rho^2(\theta,\Theta_0) + n^{-2/3}M_n^2 \text{ for all } \theta \in \Theta_0^{\eta_1\wedge\eta_2}.$$

Notice that when $n^{-2/3}M_n^2$ is smaller than $(\kappa_1/4)\rho^2(\theta,\Theta_0)$, we have

$$Q_n(\theta) \leq \sup_\Theta Q - \frac{\kappa_1}{4}\rho^2(\theta,\Theta_0), \tag{11}$$

which is of the form required by Assumption A4. This is true whenever $\rho(\theta,\Theta_0) \geq 4\kappa_1^{-1/2}n^{-1/3}M_n$. Since $M_n = O_p(1)$, for any $\varepsilon_3 \in (0,1)$, there exists a κ_3 and n_3 such that for all $n \geq n_3$, $\rho(\theta,\Theta_0) \geq \kappa_3 n^{-1/3} \geq 4\kappa_1^{-1/2}n^{-1/3}M_n$ and the bound in (11) holds uniformly on $\Theta_0^\eta \setminus \Theta_0^{\kappa_3 n^{-1/3}}$ with probability at least $1-\varepsilon_3$. (Note that we can always choose n_3 large enough so that $\kappa_3 n^{-1/3}$ is smaller than $\eta < \eta_1 \wedge \eta_2$, ensuring that the relevant region of the domain is nonempty.)

To show that Assumption A4 holds, let $\varepsilon \in (0,1)$ be given. For $\varepsilon_1 = \varepsilon/2$, choose n_1 and δ_η as above so that (9) holds uniformly on $\Theta \setminus \Theta_0^\eta$ with probability at least $1-\varepsilon_1$. Then, for $\varepsilon_3 = \varepsilon/2$, choose n_3 and κ_3 such that (11) holds uniformly on $\Theta_0^\eta \setminus \Theta_0^{\kappa_3 n^{-1/3}}$ with probability at least $1-\varepsilon_3$.

To summarize, we have shown that

$$Q_n(\theta) \leq \sup_\Theta Q - \max\left\{\frac{\kappa_1}{4}\rho^2(\theta,\Theta_0), \delta_\eta\right\}$$

uniformly on $\Theta \setminus \Theta_0^{\kappa_3 n^{-1/3}}$ with probability at least $1-\varepsilon$. It follows that Assumption A4 holds with $a_n = n^{1/2}$, $\delta = \delta_\eta$, $c = \kappa_1/4$, $\gamma_1 = 2$, $\gamma_2 = 2/3$, $c_\varepsilon = \kappa_3$, and $n_\varepsilon = \max\{n_1, n_3\}$. Therefore, for any sequence r_n such that $r_n = o(n^{1/3})$, let $\tau_n \propto r_n^{-3/2}$. Since $n^{1/3}r_n^{-1} \to \infty$, we have $(n^{1/3}r_n^{-1})^{3/2} = n^{1/2}r_n^{-3/2} \propto n^{1/2}\tau_n \xrightarrow{p} \infty$ and therefore Theorem 2 implies $d_{\mathrm{H}}(\hat{\Theta}_n,\Theta_0) = O_p(\tau_n^{2/3}) = O_p(r_n^{-1})$. ■

B.3. Sufficient Conditions for Arbitrarily Fast Convergence

Lemma 10. *Suppose that Assumptions B1–B3 hold and that Q is a step function. If $\tau_n \xrightarrow{\mathrm{p}} 0$ and $n^{1/2}\tau_n \xrightarrow{\mathrm{p}} \infty$, then for any positive sequence r_n with $r_n \to \infty$, $r_n d_{\mathrm{H}}(\hat{\Theta}_n, \Theta_0) \xrightarrow{\mathrm{p}} 0$.*

Proof of Lemma 10. As established by Lemma 8, Assumptions B1–B3 are sufficient for Assumptions A1–A3 with $a_n = n^{1/2}$. Since Q is a step function, Assumption A4' holds for all $\delta < \sup_\Theta Q - \sup_{\Theta\setminus\Theta_0} Q$. The result follows from Theorem 3. ■

C. Proofs of Results for the Semiparametric Binary Response Model

Proof of Lemma 3. We verify the conditions of Lemma 8 with $a_n = n^{1/2}$. Since Θ is compact, Assumption B1 is verified. Q is continuous when Assumption C1 holds and Q is a step function when Assumption C2 holds, so Assumption B2 is satisfied. It remains to verify Assumption B3.

Let $\alpha, \gamma \in \mathbb{R}$ and $\delta \in \mathbb{R}^K$. For each $(x, y, t) \in \mathscr{X} \times \{0,1\} \times \mathbb{R}$, define $g(x, y, t; \alpha, \gamma, \delta) = \alpha t + \gamma y + \delta' x$ and $\mathscr{G} = \{g(\cdot,\cdot,\cdot;\alpha,\gamma,\delta) : \alpha, \gamma \in \mathbb{R} \text{ and } \delta \in \mathbb{R}^K\}$. Since $\mathscr{G}$ is a finite-dimensional vector space of real-valued functions on $\mathscr{X} \times \{0,1\} \times \mathbb{R}$, classes of sets of the form $\{g \geq r\}$ or $\{g > r\}$ with $g \in \mathscr{G}$ and $r \in \mathbb{R}$ are VC classes (Pakes and Pollard, 1989, Lemma 2.4). We will use particular choices of α, γ, and δ to show that $\mathscr{F}$ is VC subgraph class. First, note that we can rewrite f as

$$\begin{aligned} f(x, y, \theta) &= \left(1\{y > 0\} - 1\{y \leq 0\}\right)\left(1\{x'\theta \geq 0\} - 1\{x'\theta < 0\}\right) \\ &= 1\{y > 0,\, x'\theta \geq 0\} - 1\{y > 0,\, x'\theta < 0\} - 1\{y \leq 0,\, x'\theta \geq 0\} + 1\{y \leq 0,\, x'\theta < 0\}. \end{aligned}$$

For any $\theta \in \Theta$,

$$\begin{aligned} \operatorname{subgraph}\left(f(\cdot,\cdot,\theta)\right) &= \left\{(x, y, t) \in \mathscr{X} \times \{0,1\} \times \mathbb{R} : 0 < t < f(x, y, \theta) \text{ or } 0 > t > f(x, y, \theta)\right\} \\ &= \left(\{y > 0\} \cap \{x'\theta \geq 0\} \cap \{t \geq 1\}^{\mathrm{c}} \cap \{t > 0\}\right) \\ &\quad \cup \left(\{y > 0\} \cap \{x'\theta \geq 0\}^{\mathrm{c}} \cap \{t > -1\} \cap \{t \geq 0\}^{\mathrm{c}}\right) \\ &\quad \cup \left(\{y \geq 0\}^{\mathrm{c}} \cap \{x'\theta \geq 0\} \cap \{t > -1\} \cap \{t \geq 0\}^{\mathrm{c}}\right) \\ &\quad \cup \left(\{y \geq 0\}^{\mathrm{c}} \cap \{x'\theta < 0\} \cap \{t \geq 1\}^{\mathrm{c}} \cap \{t > 0\}\right). \end{aligned}$$

Now, we construct three functions of the form $g_j(x, y, t) = \alpha_j t + \gamma_j z + \delta_j' x$ with $g_j \in \mathscr{G}$ for $j = 1, 2, 3$ by choosing $\alpha_1 = 0$, $\gamma_1 = 1$, $\delta_1 = 0$, $\alpha_2 = 0$, $\gamma_2 = 0$, $\delta_2 = \theta$, $\alpha_3 = 1$, $\gamma_3 = 0$, and $\delta_3 = 0$. Then

$$\begin{aligned} \operatorname{subgraph}\left(f(\cdot,\cdot,\theta)\right) &= \left(\{g_1 > 0\} \cap \{g_2 \geq 0\} \cap \{g_3 \geq 1\}^{\mathrm{c}} \cap \{g_3 > 0\}\right) \\ &\quad \cup \left(\{g_1 > 0\} \cap \{g_2 \geq 0\}^{\mathrm{c}} \cap \{g_3 > -1\} \cap \{g_3 \geq 0\}^{\mathrm{c}}\right) \\ &\quad \cup \left(\{g_1 \geq 0\}^{\mathrm{c}} \cap \{g_2 \geq 0\} \cap \{g_3 > -1\} \cap \{g_3 \geq 0\}^{\mathrm{c}}\right) \\ &\quad \cup \left(\{g_1 \geq 0\}^{\mathrm{c}} \cap \{g_2 < 0\} \cap \{g_3 \geq 1\}^{\mathrm{c}} \cap \{g_3 > 0\}\right). \end{aligned}$$

Since sets of the form $\{g \geq 0\}$ or $\{g > 0\}$ for $g \in \mathscr{G}$ are VC classes and since this property is preserved over complements, unions, and intersections (Pakes and Pollard, 1989, Lemma 2.5), it follows that

$\{\text{subgraph}(f) : f \in \mathscr{F}\}$ is a VC class. By Lemma 2.12 of Pakes and Pollard (1989), $\mathscr{F}$ is Euclidean for every envelope including $F = 1$. Since $\mathscr{F}$ is Euclidean it is also manageable in the sense of Pollard (1989) (cf. Pakes and Pollard, 1989, p. 1033), verifying Assumption B3. ■

Proof of Lemma 4. Following Abrevaya and Huang (2005), we re-normalize the objective function using $f(x, y, \beta) = (2y-1)(1\{\tilde{x}'\beta + x_K \geq 0\} - 1\{\tilde{x}'\bar{\beta} + x_K \geq 0\})$ for some $\bar{\beta} \in B_0$. Assumptions B1–B3 hold by Lemma 3. We verify Assumptions B4 and B5 here and appeal to Theorem 4 to establish the rate.

Under assumptions a–f, it follows from Abrevaya and Huang (2005, p. 1200) that $\nabla_{\beta\beta'} Q(\beta) = -V(\beta)$ for all $\beta \in \text{bd}(B_0)$. Therefore, in a neighborhood $\mathscr{N}$ of B_0, Q is approximately quadratic and for some $c > 0$, $Q(\beta) \leq \sup Q - c\rho^2(\beta, B_0)$. This verifies Assumption B4.

To show that Assumption B5 holds, let $\eta > 0$ and define $\mathscr{F}_\eta \equiv \{f(\cdot, \beta) \in \mathscr{F} : \rho(\beta, B_0) \leq \eta\}$. We will show that the natural envelope F_η of $\mathscr{F}_\eta$ is such that $PF_\eta^2 = O(\eta)$. First, note that by definition of $\mathscr{F}_\eta$ if $f(\cdot, \beta) \in \mathscr{F}_\eta$, then $\beta \in B_0^\eta$. Then,

$$\begin{aligned}
F_\eta(x, y) &= \sup_{\beta \in B_0^\eta} \left| f(x, y, \beta) \right| = \sup_{\beta \in B_0^\eta} \left| 1\{\tilde{x}'\beta \geq -x_K > \tilde{x}'\bar{\beta}\} + 1\{\tilde{x}'\bar{\beta} \geq -x_K > \tilde{x}'\beta\} \right| \\
&\leq \sup_{\beta \in B_0^\eta} 1\left\{ - \|\tilde{x}\| \cdot \left\|\beta - \bar{\beta}\right\| \leq \tilde{x}'\bar{\beta} + x_K \leq \|\tilde{x}\| \cdot \left\|\beta - \bar{\beta}\right\| \right\}.
\end{aligned}$$

Now, for all $\beta \in B_0^\eta$ and $\bar{\beta} \in B_0$, we have

$$\left\|\beta - \bar{\beta}\right\| \leq \rho(\beta, \text{bd}(B_0)) + d_{\text{H}}(\text{bd}(B_0), \text{int}(B_0)) \leq \eta + \delta$$

where $\delta \equiv d_{\text{H}}(\text{bd}(B_0), \text{int}(B_0))$. Therefore $F_\eta(x, y) \leq 1\left\{-(\eta + \delta) \|\tilde{x}\| \leq \tilde{x}'\bar{\beta} + x_K \leq (\eta + \delta) \|\tilde{x}\|\right\}$ and

$$\begin{aligned}
PF_\eta^2 &\leq \int_{\tilde{\mathscr{X}}} \int_{\{x_K \in \mathscr{X}_K : -(\eta+\delta)\|\tilde{x}\| \leq \tilde{x}'\bar{\beta} + x_K \leq (\eta+\delta)\|\tilde{x}\|\}} dF_{x_K|\tilde{x}}(x_K \mid \tilde{x})\, dF_{\tilde{x}}(\tilde{x}) \\
&\leq 2(\eta + \delta) \int_{\tilde{\mathscr{X}}} \sup_{x_K \in \mathscr{X}_K} f_{x_K|\tilde{x}}(x_K \mid \tilde{x}) \cdot \|\tilde{x}\|\, dF_{\tilde{x}}(\tilde{x}).
\end{aligned}$$

Since $\tilde{x}$ has finite first absolute moments (assumption a), $f_{x_K|\tilde{x}}(x_K \mid \tilde{x})$ is finite for almost every $\tilde{x}$ (assumption b), and $\delta < \infty$ since B_0 is compact (assumption e), we have $PF_\eta^2 = O(\eta)$. ■

Proof of Lemma 6. We verify the conditions of Theorem 2.2 of Romano and Shaikh (2010) by showing that R_n converges in distribution to R and that the distribution of R is continuous at the $1 - \alpha$ quantile.

Recall that the binary choice model the parameter vector θ is normalized up to scale so that $\theta = (\beta', \theta_K)'$ for $\beta \in B$ and $\theta_K \in \{-1, 1\}$. As in the proof of Theorem 4, we use the re-normalized objective function $Q_n(\beta) = P_n f(\cdot, \beta)$ for $f(\cdot, \beta) \in \mathscr{F}$ so that $f(\cdot, \beta) = 0$ for $\beta \in B_0$. We can restate the inferential statistic as $R_n = n^{2/3} \sup_{\beta \in B} Q_n(\beta) - n^{2/3} \inf_{\beta \in B_0} Q_n(\beta)$.

It follows from a slight extension of Lemma 4.1 of Kim and Pollard (1990), following the proof of Lemma B.14 of Wan and Xu (2015), that $\inf_{\beta \in B_0} Q_n(\beta) = o_p(n^{-2/3})$. Hence, $R_n = \sup_{\beta \in B} n^{2/3} Q_n(\beta) + o_p(1)$. From (11) in the proof of Theorem 4 above, there exists a $\kappa > 0$ such that $\sup_{\beta \in B} Q_n(\beta) = \sup_{\beta \in B_0^{\kappa n^{-1/3}}} Q_n(\beta)$ with probability approaching one. Hence, we focus on rescaled parameter sequences $\beta = \bar{\beta} + t n^{-1/3}$ for $\bar{\beta} \in \text{bd}(B_0)$ and $t \in \mathbb{R}^{K-1}$. We use the following decomposition to analyze the remaining term of R_n:

$n^{2/3}Q_n(\bar{\beta}+tn^{-1/3}) = n^{2/3}Pf(\cdot,\bar{\beta}+tn^{-1/3}) + n^{2/3}(P_n - P)f(\cdot,\bar{\beta}+tn^{-1/3})$. The analysis is again similar to the arguments of Kim and Pollard (1990, Theorem 4.7) and Wan and Xu (2015, Lemma B.13). By condition f of Lemma 4, the first component contributes a quadratic trend: $n^{2/3}Pf(\cdot,\bar{\beta}+tn^{-1/3}) \to -t'V(\bar{\beta})t$. The second component converges in distribution to a mean-zero Gaussian process $W(\bar{\beta},\cdot)$ with covariance kernel $H(\bar{\beta},t_1,t_2) = \lim_{n\to\infty} nPf(\cdot,\bar{\beta}+t_1/n)f(\cdot,\bar{\beta}+t_2/n)$.

Finally, by Theorem 11.1 of Davydov, Lifshits, and Smorodina (1998), the distribution of R is continuous except possibly at the separation point $r_0 = \inf\{r \mid \Pr(R \le r) > 0\}$. For some $(\bar{\beta}, t) \in \mathrm{bd}(B_0) \times \mathbb{R}^{K-1}$, we have $\Pr(R \le 0) \le \Pr(W(\bar{\beta}, t) \le 0) = 1/2$ and therefore $r_0 \le 0$. Hence, the distribution of R is absolutely continuous on $(0,\infty)$ and at all $1-\alpha$ quantiles for $\alpha < 1/2$.

■