The Annals of Statistics
2010, Vol. 38, No. 5, 2884–2915
DOI: 10.1214/10-AOS809

BOOTSTRAP CONSISTENCY FOR GENERAL SEMIPARAMETRIC *M*-ESTIMATION

BY GUANG CHENG[1] AND JIANHUA Z. HUANG[2]

Purdue University and Texas A&M University

Consider M-estimation in a semiparametric model that is characterized by a Euclidean parameter of interest and an infinite-dimensional nuisance parameter. As a general purpose approach to statistical inferences, the bootstrap has found wide applications in semiparametric M-estimation and, because of its simplicity, provides an attractive alternative to the inference approach based on the asymptotic distribution theory. The purpose of this paper is to provide theoretical justifications for the use of bootstrap as a semiparametric inferential tool. We show that, under general conditions, the bootstrap is asymptotically consistent in estimating the distribution of the M-estimate of Euclidean parameter; that is, the bootstrap distribution asymptotically imitates the distribution of the M-estimate. We also show that the bootstrap confidence set has the asymptotically correct coverage probability. These general conclusions hold, in particular, when the nuisance parameter is not estimable at root-n rate, and apply to a broad class of bootstrap methods with exchangeable bootstrap weights. This paper provides a first general theoretical study of the bootstrap in semiparametric models.

1. Introduction. Due to its flexibility, semiparametric modeling has provided a powerful statistical modeling framework for complex data, and proven to be useful in a variety of contexts, see [2, 7, 20, 44, 45]. Semiparametric models are indexed by a Euclidean parameter of interest $\theta \in \Theta \subset \mathbb{R}^d$ and an infinite-dimensional nuisance function η belonging to a Banach space $\mathcal{H}$ with a norm $\|\cdot\|$. M-estimation, including the maximum likelihood estimation as a special case, refers to a general method of estimation, where the estimates are obtained by optimizing some objective functions [10, 28, 42]. The asymptotic theories and inference procedures for semiparametric maximum likelihood estimation, or more generally M-estimation, have been extensively studied in [4, 11, 22, 24, 28, 32].

It is well known that the asymptotic inferences of semiparametric models often face practical challenges. In particular, the confidence set construction and the

Received October 2009.

[1]Supported by NSF Grant DMS-09-06497.

[2]Supported in part by NSF Grants DMS-06-06580, DMS-09-07170, NCI Grant CA57030 and Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

AMS 2000 subject classifications. Primary 62F40; secondary 62G20.

Key words and phrases. Bootstrap consistency, bootstrap confidence set, semiparametric model, M-estimation.

asymptotic variance estimation of the estimator for the Euclidean parameter both involve estimating and inverting a hard-to-estimate infinite-dimensional operator. The difficulty in dealing with such an infinite-dimensional operator motivated the development of the profile sampler [8, 9, 24], where the inference of the Euclidean parameter is based on sampling from the posterior of the profile likelihood [24]. However, because of the way it is designed, the profile sampler method has the typical caveats of the Bayesian methods. First, one needs to specify a prior distribution. Second, since the Markov chain Monte Carlo (McMC) is used for sampling from the posterior distribution, there are a number of controversial issues in generating the stationary Markov chain. For example, it is considerably difficult to determine the burn-in period and stopping time of the chain [16]. In particular, it may take a long time to run the Markov chain in order to give accurate inferences for θ when η is estimable at a slow convergence rate [8, 9]. Moreover, when the sample size is small, the profile likelihood may become nonsmooth or may not approximate well the desired parabolic form, violating the main theoretical basis of the profile sampler.

On the other hand, as a general data-resampling based statistical inference tool, the bootstrap method does not have the drawbacks of the profile sampler; see [6, 19, 21, 28, 37, 43] for its application in semiparametric models. In fact, the bootstrap method has several methodological advantages over the profile sampler: it is straightforward to implement; there is no need to specify a prior distribution and to check Markov chain convergence. In addition, the finite sample performance of the bootstrap can be controlled by choosing from a rich pool of resampling techniques; see Section 3 of [33]. Moreover, unlike the profile sampler which focuses on θ, one can make bootstrap inferences for both θ and η.

Unfortunately, a systematic theoretical study on the bootstrap inference in semiparametric models is almost nonexistent, especially when the nuisance function parameter η is not $\sqrt{n}$ estimable, despite the rich literature on the bootstrap theory for parametric models [3, 18, 30, 36]. The current literature only considered the bootstrap consistency for the joint estimator of (θ, η) in some special case of semiparametric models where η is $\sqrt{n}$-estimable, that is, [21]. In a recent monograph, Kosorok pointed out that "convergence rate and asymptotic normality results are quite difficult to establish for the nonparametric bootstrap (based on multinomial weights), especially for models with parameters not estimable at the $\sqrt{n}$ rate" [22]. In fact, the lack of theoretical justifications of the bootstrap in the semiparametric context is one of the main motivations for developing the profile sampler. The purpose of this paper is to develop a general theory on bootstrap consistency in semiparametric models, for a broad class of bootstrap methods including Efron's (nonparametric) bootstrap as a special case. We focus on the inference of the Euclidean parameter and leave study of the bootstrap inference of the nuisance parameter for future research, although we give some useful convergence rate results (see Section 5).

Our main results are summarized as follows. The semiparametric M-estimator $(\widehat{\theta}, \widehat{\eta})$ and the bootstrap M-estimator $(\widehat{\theta}^*, \widehat{\eta}^*)$ are obtained by optimizing the objective function $m(\theta, \eta)$ based on the i.i.d. observations $(X_1, \ldots, X_n)$ and the bootstrap sample $(X_1^*, \ldots, X_n^*)$, respectively:

$$(\widehat{\theta}, \widehat{\eta}) = \arg\sup_{\theta\in\Theta, \eta\in\mathcal{H}} \sum_{i=1}^{n} m(\theta, \eta)(X_i), \tag{1}$$

$$(\widehat{\theta}^*, \widehat{\eta}^*) = \arg\sup_{\theta\in\Theta, \eta\in\mathcal{H}} \sum_{i=1}^{n} m(\theta, \eta)(X_i^*), \tag{2}$$

where $(X_1^*, \ldots, X_n^*)$ are independent draws with replacement from the original sample. Note that we can express

$$(\widehat{\theta}^*, \widehat{\eta}^*) = \arg\sup_{\theta\in\Theta, \eta\in\mathcal{H}} \sum_{i=1}^{n} W_{ni} m(\theta, \eta)(X_i), \tag{3}$$

and the bootstrap weights $(W_{n1}, \ldots, W_{nn}) \sim \text{Multinomial}(n, (n^{-1}, \ldots, n^{-1}))$. In this paper, we consider the more general *exchangeable bootstrap weighting scheme* that includes Efron's bootstrap and its smooth alternative [27], for example, *Bayesian bootstrap*, as special cases. The general resampling scheme was first proposed in [34], and extensively studied by [1], who suggested the name "*weighted bootstrap*," and in [30, 33]. Note that other variations of Efron's bootstrap are also studied in [5] using the term "*generalized bootstrap*." The practical usefulness of the more general scheme is well-documented in the literature. For example, in semiparametric survival models, for example, Cox regression model, the nonparametric bootstrap often gives many ties when it is applied to censored survival data due to its "discreteness" and the general weighting scheme comes to the rescue. As one main contribution of the paper, we show that the nonparametric bootstrap distribution of $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$, conditional on the observed data, asymptotically imitates the distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$, where θ_0 is the true value of θ. As a consequence, we also establish the consistency of the bootstrap confidence set of θ, which means that the coverage probability converges to the nominal level. Our results hold when the estimate of the nuisance function has either root-n or slower than root-n convergence rate. This paper can also be viewed as a nontrivial extension of [5] to account for the presence of an infinite-dimensional nuisance parameter.

In a related paper, Ma and Kosorok [28] obtained some theoretical results when the bootstrap weights are assumed to be i.i.d. There is a crucial difference between their work and ours: They treated the bootstrap estimator as the regular weighted estimator and used the unconditional arguments rather than the usual conditional arguments as we employ in this paper. Note that the i.i.d. assumption rules out all interesting bootstrap schemes considered in this paper, and their theoretical approach cannot be extended to obtain our results. Indeed, they stated in the paper

that the independence assumption makes their proofs easier and the relaxation to the dependent weights appears to be quite difficult. Another related work is the piggyback bootstrap [11], which is invented solely to draw inferences for the functional parameter η when it is $\sqrt{n}$-estimable. The piggyback bootstrap is not the standard bootstrap and relies on a valid random draw from the asymptotic distribution of the estimate of θ, which is hard to estimate in general. Other related work includes interesting results on bootstrap (in)-consistency in nonparametric estimation; see [23, 35, 41]. An m out of n bootstrap was developed for nonstandard M-estimation with nuisance parameters in parametric models [25].

Section 2 provides the necessary background of M-estimation in semiparametric models. Our main results, including the bootstrap consistency theorem, are presented in Section 3. Sections 4 and 5 discuss how to verify various technical conditions needed for the main results. Section 6 illustrates the applications of our main results in three examples. Section 7 contains the proof of the main results in Section 3. Some useful lemmas and additional proofs are postponed to Appendix.

2. Background. We first introduce a paradigm for the semiparametric M-estimation [28, 42], which parallels the efficient influence function paradigm used for the MLEs [where $m(\theta, \eta)$ is the log likelihood]. Next, we present the model assumptions needed for the remainder of the paper, and, finally, we review some known results on the asymptotic distribution of semiparametric M-estimators, which are needed in studying the asymptotic properties of the bootstrap.

Let

$$m_1(\theta, \eta) = \frac{\partial}{\partial \theta} m(\theta, \eta) \quad \text{and} \quad m_2(\theta, \eta)[h] = \frac{\partial}{\partial t} m(\theta, \eta(t))\Big|_{t=0},$$

where h is a "direction" along which $\eta(t) \in \mathcal{H}$ approaches η as $t \to 0$, running through some index set $\mathbf{H} \subseteq L_2^0(P_{\theta,\eta})$. Similarly, we also define

$$m_{11}(\theta, \eta) = \frac{\partial}{\partial \theta} m_1(\theta, \eta) \quad \text{and} \quad m_{12}(\theta, \eta)[h] = \frac{\partial}{\partial t} m_1(\theta, \eta(t))\Big|_{t=0},$$

$$m_{21}(\theta, \eta)[h] = \frac{\partial}{\partial \theta} m_2(\theta, \eta)[h] \quad \text{and} \quad m_{22}(\theta, \eta)[h, g] = \frac{\partial}{\partial t} m_2(\theta, \eta_2(t))[h]\Big|_{t=0},$$

where $h, g \in \mathbf{H}$ and $(\partial/\partial t)\eta_2(t)|_{t=0} = g$. Define

$$m_2(\theta, \eta)[H] = (m_2(\theta, \eta)[h_1], \ldots, m_2(\theta, \eta)[h_d])',$$

$$m_{22}[H, h] = (m_{22}(\theta, \eta)[h_1, h], \ldots, m_{22}(\theta, \eta)[h_d, h])',$$

where $H = (h_1, \ldots, h_d)$ and $h_j \in \mathbf{H}$ for $j = 1, \ldots, d$. Assume there exists an

$$H^\dagger(\theta, \eta) = (h_1^\dagger(\theta, \eta), \ldots, h_d^\dagger(\theta, \eta))',$$

where each $h_j^\dagger(\theta, \eta) \in \mathbf{H}$, such that for any $h \in \mathbf{H}$

$$E_{\theta,\eta}\{m_{12}(\theta, \eta)[h] - m_{22}(\theta, \eta)[H^\dagger, h]\} = 0. \tag{4}$$

Following the idea of the *efficient score function*, we define the function

$$\widetilde{m}(\theta,\eta)=m_1(\theta,\eta)-m_2(\theta,\eta)[H^{\dagger}(\theta,\eta)].$$

We assume that the observed data are from the probability space $(\mathcal{X},\mathcal{A},P_X)$, and that

$$P_X\widetilde{m}(\theta_0,\eta_0)=0, \tag{5}$$

where P_Xf is the customary operator notation defined as $\int f\,dP_X$. The assumption (5) is common in semiparametric M-estimation [28, 42] and usually holds by the model specifications, for example, the semiparametric regression models with "panel count data" [42]. In particular, when $m(\theta,\eta)=\log\mathrm{lik}(\theta,\eta)$, (5) trivially holds and $\widetilde{m}(\theta,\eta)$ becomes the well studied *efficient score function* for θ in semiparametric models, see [4]. Since $(\widehat{\theta},\widehat{\eta})$ is assumed to be the maximizer of $\sum_{i=1}^n m(\theta,\eta)(X_i)$, $(\widehat{\theta},\widehat{\eta})$ satisfies

$$\mathbb{P}_n\widetilde{m}(\widehat{\theta},\widehat{\eta})=0, \tag{6}$$

where $\mathbb{P}_nf$ denotes $\sum_{i=1}^n f(X_i)/n$. The theory developed in this paper is general enough to deal with the case that $(\widehat{\theta},\widehat{\eta})$ is not the exact maximizer. Instead of (6), we only assume the following "nearly-maximizing" condition

$$\mathbb{P}_n\widetilde{m}(\widehat{\theta},\widehat{\eta})=o^o_{P_X}(n^{-1/2}), \tag{7}$$

where the superscript "o" denotes the outer probability.

Throughout the rest of the paper, we use the shortened notation $H_0^{\dagger}=H^{\dagger}(\theta_0,\eta_0)$, $\widetilde{m}_0=\widetilde{m}(\theta_0,\eta_0)$ and $\widehat{m}=\widetilde{m}(\widehat{\theta},\widehat{\eta})$. For a probability space $(\Omega,\mathcal{A},P)$ and a map $T:\Omega\mapsto\bar{\mathbb{R}}$ that need not be measurable, the notation E^oT, $O^o_P(1)$, and $o^o_P(1)$ represent the outer expectation of T w.r.t. P, bounded and converging to zero in outer probability, respectively. More precise definitions can be found on page 6 of [38]. Let $V^{\otimes2}$ represent VV' for any vector V. Define $x\vee y$ $(x\wedge y)$ to be the maximum (minimum) value of x and y.

We now state some general conditions that will be used throughout the whole paper. We assume that the true value θ_0 of the Euclidean parameter is an interior point of the compact set Θ. Define

$$A=P_X\{(\partial/\partial\theta)|_{\theta=\theta_0}\widetilde{m}(\theta,\eta_0)\}=P_X\{m_{11}(\theta_0,\eta_0)-m_{21}(\theta_0,\eta_0)[H_0^{\dagger}]\}, \tag{8}$$

$$B=\mathrm{Var}\{\widetilde{m}_0(X)\}=P_X\big[\{m_1(\theta_0,\eta_0)-m_2(\theta_0,\eta_0)[H_0^{\dagger}]\}^{\otimes2}\big]. \tag{9}$$

I. *Positive information condition*: the matrices A and B are both nonsingular.

Condition I above is used to ensure the nonsingularity of the asymptotic variance of $\widehat{\theta}$, which will be shown to be $A^{-1}B(A^{-1})'$; see Proposition 1.

For the empirical process $\mathbb{G}_n=\sqrt{n}(\mathbb{P}_n-P_X)$, denote its norm with respect to a function class $\mathcal{F}_n$ as $\|\mathbb{G}_n\|_{\mathcal{F}_n}=\sup_{f\in\mathcal{F}_n}|\mathbb{G}_nf|$. For any fixed $\delta_n>0$, define a class of functions $\mathcal{S}_n$ as

$$\mathcal{S}_n\equiv\mathcal{S}_n(\delta_n)=\left\{\frac{\widetilde{m}(\theta_0,\eta)-\widetilde{m}(\theta_0,\eta_0)}{\|\eta-\eta_0\|}:\|\eta-\eta_0\|\le\delta_n\right\} \tag{10}$$

and a shrinking neighborhood of (θ_0, η_0) as

$$\mathcal{C}_n \equiv \mathcal{C}_n(\delta_n) = \{(\theta, \eta) : \|\theta - \theta_0\| \le \delta_n, \|\eta - \eta_0\| \le \delta_n\}. \tag{11}$$

The next two conditions S1 and S2 imply that the empirical processes indexed by $\widetilde{m}(\theta, \eta)$ are well behaved and $\widetilde{m}(\theta, \eta)$ is smooth enough around (θ_0, η_0).

S1. *Stochastic equicontinuity condition*: for any $\delta_n \to 0$,

$$\|\mathbb{G}_n\|_{\mathcal{S}_n} = O^o_{P_X}(1) \tag{12}$$

and

$$\mathbb{G}_n\big(\widetilde{m}(\theta, \eta) - \widetilde{m}(\theta_0, \eta)\big) = O^o_{P_X}(\|\theta - \theta_0\|) \qquad \text{for } (\theta, \eta) \in \mathcal{C}_n. \tag{13}$$

S2. *Smoothness condition*:

$$P_X\big(\widetilde{m}(\theta, \eta) - \widetilde{m}_0\big) = A(\theta - \theta_0) + O(\|\theta - \theta_0\|^2 \vee \|\eta - \eta_0\|^2) \tag{14}$$

for (θ, η) in some neighborhood of (θ_0, η_0).

For any fixed θ, define

$$\widehat{\eta}_\theta = \arg\sup_{\eta \in \mathcal{H}} \mathbb{P}_n m(\theta, \eta).$$

The next condition says that $\widehat{\eta}_\theta$ should be close to η_0 if θ is close to θ_0.

S3. *Convergence rate condition*: there exists a $\gamma \in (1/4, 1/2]$ such that

$$\|\widehat{\eta}_{\widetilde{\theta}} - \eta_0\| = O^o_{P_X}(\|\widetilde{\theta} - \theta_0\| \vee n^{-\gamma}) \tag{15}$$

for any consistent $\widetilde{\theta}$.

The above range requirement of γ is always satisfied for regular semiparametric models; see Section 3.4 of [38]. Verifications of conditions S1–S3 will be discussed in Sections 4 and 5, and illustrated with examples in Section 6.

The following proposition summarizes a known result on the the asymptotic normality of the semiparametric M-estimator $\widehat{\theta}$ [22, 28, 42], which plays an important role in proving bootstrap consistency in next section.

PROPOSITION 1. *Suppose that conditions* I, S1–S3 *hold and that* $(\widehat{\theta}, \widehat{\eta})$ *satisfies* (7). *If* $\widehat{\theta}$ *is consistent, then*

$$\sqrt{n}(\widehat{\theta} - \theta_0) = -\sqrt{n} A^{-1} \mathbb{P}_n \widetilde{m}_0 + o^o_{P_X}(1). \tag{16}$$

Consequently,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma), \tag{17}$$

where $\Sigma \equiv A^{-1} B (A^{-1})'$, A *and* B *are given in* (8) *and* (9), *respectively.*

We assume consistency of $\widehat{\theta}$ in Proposition 1. The consistency can usually be guaranteed under the following "well-separated" condition

$$P_X m(\theta_0, \eta_0) > \sup_{(\theta,\eta)\notin G} P_X m(\theta, \eta) \tag{18}$$

for any open set $G \subset \Theta \times \mathcal{H}$ containing (θ_0, η_0), see Theorem 5.7 in [39]. For maximum likelihood estimation, that is, $m(\theta, \eta) = \log \mathrm{lik}(\theta, \eta)$, it is easy to see that $A = -B$ and $\Sigma = B^{-1}$, and thus Σ^{-1} becomes the *efficient information matrix*.

REMARK 1. Given any consistent estimator $\widehat{\Sigma}$ of Σ, we have

$$\sqrt{n}\widehat{\Sigma}^{-1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, I) \tag{19}$$

by Proposition 1 and Slutsky's theorem. In practice, a consistent $\widehat{\Sigma}$ can be obtained via either the observed profile information approach [31] or the profile sampler approach [24].

3. Main results: Bootstrap consistency. In this section, we establish the consistency of bootstrapping θ under general conditions in the framework of semiparametric M-estimation. Define

$$\mathbb{P}_n^* f = (1/n)\sum_{i=1}^{n} W_{ni} f(X_i),$$

where W_{ni}'s are the bootstrap weights defined on the probability space $(\mathcal{W}, \Omega, P_W)$. In view of (3), the bootstrap estimator can be rewritten as

$$(\widehat{\theta}^*, \widehat{\eta}^*) = \arg\sup_{\theta\in\Theta,\eta\in\mathcal{H}} \mathbb{P}_n^* m(\theta, \eta). \tag{20}$$

The definition of $(\widehat{\theta}^*, \widehat{\eta}^*)$, that is, (20), implies that

$$\mathbb{P}_n^* \widetilde{m}(\widehat{\theta}^*, \widehat{\eta}^*) = 0. \tag{21}$$

Similar to (7), we weaken (21) to the following "nearly-maximizing" condition

$$\mathbb{P}_n^* \widetilde{m}(\widehat{\theta}^*, \widehat{\eta}^*) = o^o_{P_{XW}}(n^{-1/2}), \tag{22}$$

where P_{XW} is a probability measure on a product space that we will formally define later.

The bootstrap weights W_{ni}'s are assumed to belong to the class of exchangeable bootstrap weights introduced in [33]. Specifically, they satisfy:

W1. The vector $W_n = (W_{n1}, \ldots, W_{nn})'$ is exchangeable for all $n = 1, 2, \ldots$, that is, for any permutation $\pi = (\pi_1, \ldots, \pi_n)$ of $(1, 2, \ldots, n)$, the joint distribution of $\pi(W_n) = (W_{n\pi_1}, \ldots, W_{n\pi_n})'$ is the same as that of W_n.

W2. $W_{ni} \geq 0$ for all n, i and $\sum_{i=1}^{n} W_{ni} = n$ for all n.

W3. For some positive constant $C < \infty$, $\limsup_{n\to\infty} \|W_{n1}\|_{2,1} \le C$, where $\|W_{n1}\|_{2,1} = \int_0^\infty \sqrt{P_W(W_{n1} \ge u)}\, du$.

W4. $\lim_{\lambda\to\infty} \limsup_{n\to\infty} \sup_{t\ge\lambda} t^2 P_W(W_{n1} > t) = 0$.

W5. $(1/n)\sum_{i=1}^n (W_{ni} - 1)^2 \xrightarrow{P_W} c^2 > 0$.

The bootstrap weights corresponding to Efron's nonparametric bootstrap satisfy W1–W5. Another important class of bootstrap whose weights satisfy W1–W5 is the *multiplier bootstrap* in which $W_{ni} = \omega_i/\bar{\omega}_n$ and $(\omega_1, \ldots, \omega_n)$ are i.i.d. positive r.v.s with $\|\omega_1\|_{2,1} < \infty$. By taking $\omega_i \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$, we obtain the *Bayesian bootstrap* of [34]. The multiplier bootstrap is often thought to be a smooth alternative to the nonparametric bootstrap [27]. In general, conditions W3–W5 are easily satisfied under some moment conditions on W_{ni}; see Lemma 3.1 of [33]. The sampling schemes that satisfy conditions W1–W5 include *the double bootstrap*, *the urn bootstrap and the grouped or delete-h Jackknife* [13]; see [33]. The value of c in W5 is independent of n and depends on the resampling method, for example, $c = 1$ for the nonparametric bootstrap and Bayesian bootstrap, and $c = \sqrt{2}$ for the double bootstrap.

There exist two sources of randomness for the bootstrapped quantity, for example, $\widehat{\theta}^*$ and $\widehat{\eta}^*$: one comes from the observed data; another comes from the resampling done by the bootstrap, that is, randomness in W_{ni}'s. Therefore, in order to rigorously state our theoretical results for the bootstrap, we need to specify relevant probability spaces and define the related stochastic orders.

We view X_i as the ith coordinate projection from the canonical probability space $(\mathcal{X}^\infty, \mathcal{A}^\infty, P_X^\infty)$ onto the ith copy of $\mathcal{X}$. For the joint randomness involved, the product probability space is defined as

$$(\mathcal{X}^\infty, \mathcal{A}^\infty, P_X^\infty) \times (\mathcal{W}, \Omega, P_W) = (\mathcal{X}^\infty \times \mathcal{W}, \mathcal{A}^\infty \times \Omega, P_{XW}).$$

In this paper, we assume that the bootstrap weights W_{ni}'s are independent of the data X_i's, thus $P_{XW} = P_X^\infty \times P_W$. We write P_X^∞ as P_X for simplicity thereafter. Define E_{XW}^o as the outer expectation w.r.t. P_{XW}. The notation $E_{W|X}^o$, E_X^o and E_W are defined similarly.

Given a real-valued function Δ_n defined on the above product probability space, for example, $\widehat{\theta}^*$, we say that Δ_n is of an order $o_{P_W}^o(1)$ in P_X^o-probability if for any $\varepsilon, \delta > 0$,

$$P_X^o\{P_{W|X}^o(|\Delta_n| > \varepsilon) > \delta\} \longrightarrow 0 \qquad \text{as } n \to \infty, \tag{23}$$

and that Δ_n is of an order $O_{P_W}^o(1)$ in P_X^o-probability if for any $\delta > 0$, there exists a $0 < M < \infty$ such that

$$P_X^o\{P_{W|X}^o(|\Delta_n| \ge M) > \delta\} \longrightarrow 0 \qquad \text{as } n \to \infty. \tag{24}$$

Given a function Γ_n defined only on $(\mathcal{X}^\infty, \mathcal{A}^\infty, P_X^\infty)$, if it is of an order $o_{P_X}^o(1)$ $[O_{P_X}^o(1)]$, then it is also of an order $o_{P_{XW}}^o(1)$ $[O_{P_{XW}}^o(1)]$ based on the following

argument:

$$P_{XW}^o(|\Gamma_n| > \varepsilon) = E_{XW}^o 1\{|\Gamma_n| > \varepsilon\} = E_X E_{W|X} 1\{|\Gamma_n| > \varepsilon\}^o$$
$$= E_X 1\{|\Gamma_n| > \varepsilon\}^o = P_X^o\{|\Gamma_n| > \varepsilon\},$$

where the third equation holds since Γ_n does not depend on the bootstrap weight. More results on transition of various stochastic orders are given in Lemma 3 of the Appendix. Such results are used repeatedly in proving our bootstrap consistency theorem.

To establish the bootstrap consistency, we need some additional conditions. The first condition is the measurability condition, denoted as $M(P_X)$. We say a class of functions $\mathcal{F} \in M(P_X)$ if $\mathcal{F}$ possesses enough measurability so that $\mathbb{P}_n$ can be randomized, that is, we can replace $(\delta_{X_i} - P_X)$ by $(W_{ni} - 1)\delta_{X_i}$, and Fubini's theorem can be used freely. The detailed description for $M(P_X)$ is spelled out in [17] and also given in the Appendix of this paper. Define $\mathcal{T} = \{\widetilde{m}(\theta, \eta) : \|\theta - \theta_0\| + \|\eta - \eta_0\| \le R\}$ for some $R > 0$. For the rest of the paper, we assume $\mathcal{T} \in M(P_X)$.

The second class of conditions parallels conditions S1–S3 used for obtaining asymptotic normality of $\widehat{\theta}$ and is only slightly stronger. Thus, the bootstrap consistency for θ is almost automatically guaranteed once $\widehat{\theta}$ is shown to be asymptotically normal. Let $S_n(x)$ be the envelop function of the class $\mathcal{S}_n = \mathcal{S}_n(\delta_n)$ defined in (10), that is,

$$S_n(x) = \sup_{\|\eta - \eta_0\| \le \delta_n} \left| \frac{\widetilde{m}(\theta_0, \eta) - \widetilde{m}_0}{\|\eta - \eta_0\|} \right|.$$

The next condition controls the tail of this envelop function.

SB1. *Tail probability condition*:

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{t \ge \lambda} t^2 P_X^o\big(S_n(X_1) > t\big) = 0 \tag{25}$$

for any sequence $\delta_n \to 0$.

Let $\dot{\mathcal{T}} = \{\partial \widetilde{m}(\theta, \eta)/\partial\theta : (\theta, \eta) \in \mathcal{C}_n\}$, where $\mathcal{C}_n = \mathcal{C}_n(\delta_n)$ is defined in (11).

SB2. We assume that $\dot{\mathcal{T}} \in M(P_X) \cap L_2(P_X)$ and that $\dot{\mathcal{T}}$ is P-Donsker.

Condition SB2 ensures that the size of the function class $\dot{\mathcal{T}}$ is reasonable so that the bootstrapped empirical processes $\mathbb{G}_n^* \equiv \sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$ indexed by $\dot{\mathcal{T}}$ has a limiting process conditional on the observations; see Theorem 2.2 in [33].

For any fixed θ, define

$$\widehat{\eta}_\theta^* = \arg\max_{\eta \in \mathcal{H}} \mathbb{P}_n^* m(\theta, \eta).$$

The next condition says that $\widehat{\eta}_\theta^*$ should be close to η_0 if θ is close to θ_0.

SB3. *Bootstrap convergence rate condition*: there exists a $\gamma \in (1/4, 1/2]$ such that

$$\|\widehat{\eta}^*_{\widetilde{\theta}} - \eta_0\| = O^o_{P_W}(\|\widetilde{\theta} - \theta_0\| \vee n^{-\gamma}) \qquad \text{in } P^o_X\text{-probability} \tag{26}$$

for any $\widetilde{\theta} \xrightarrow{P^o_{XW}} \theta_0$.

Verifications of conditions SB1–SB2 will be discussed in Section 4. Two general theorems are given in Section 5 to aid verification of condition SB3.

Now we are ready to present our main results. Theorem 1 below says that the bootstrap distribution of $(\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta})$, conditional on the observations, asymptotically imitates the unconditional distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$. Let $P_{W|\mathcal{X}_n}$ denote the conditional distribution given the observed data $\mathcal{X}_n$.

THEOREM 1. *Suppose that* $\widehat{\theta}$ *and* $\widehat{\theta}^*$ *satisfy* (7) *and* (22), *respectively. Assume that* $\widehat{\theta} \xrightarrow{P_X} \theta_0$ *and* $\widehat{\theta}^* \xrightarrow{P^o_W} \theta_0$ *in* P^o_X*-probability. In addition, assume that conditions* I, S1–S3, SB1–SB3 *and* W1–W5 *hold. We have that*

$$\|\widehat{\theta}^* - \theta_0\| = O^o_{P_W}(n^{-1/2}) \tag{27}$$

in P^o_X*-probability. Furthermore*,

$$\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) = -A^{-1}\mathbb{G}^*_n\widetilde{m}_0 + o^o_{P_W}(1) \tag{28}$$

in P^o_X*-probability. Consequently*,

$$\sup_{x\in\mathbb{R}^d} \big|P_{W|\mathcal{X}_n}\big((\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta}) \le x\big) - P\big(N(0, \Sigma) \le x\big)\big| = o^o_{P_X}(1), \tag{29}$$

where "$\le$*" is taken componentwise, c is given in* W5 *and* $\Sigma \equiv A^{-1}B(A^{-1})'$ *with A and B given in* (8) *and* (9), *respectively. Thus, we have*

$$\sup_{x\in\mathbb{R}^d} \big|P_{W|\mathcal{X}_n}\big((\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta}) \le x\big) - P_X\big(\sqrt{n}(\widehat{\theta} - \theta_0) \le x\big)\big| \xrightarrow{P^o_X} 0. \tag{30}$$

The consistency assumption for $\widehat{\theta}^*$ can be established by adapting the Argmax theorem, that is, Corollary 3.2.3 in [38]. Briefly, we need two conditions for accomplishing this. The first one is the "well-separated" condition (18). The second one is

$$\sup_{(\theta,\eta)\in\Theta\times\mathcal{H}} |\mathbb{P}^*_n m(\theta, \eta) - P_X m(\theta, \eta)| \xrightarrow{P^o_{XW}} 0. \tag{31}$$

By the multiplier Glivenko–Cantelli theorem, that is, Lemma 3.6.16 in [38], and (69) in the Appendix, we know that (31) holds if $\{m(\theta, \eta) : \theta \in \Theta, \eta \in \mathcal{H}\}$ is shown to be P-Glivenko–Cantelli.

REMARK 2. For any consistent $\widehat{\Sigma}^* \xrightarrow{P^o_{XW}} \Sigma$ and $\widehat{\Sigma} \xrightarrow{P_X} \Sigma$, we have

$$\sup_{x\in\mathbb{R}^d} \big| P_{W|\mathcal{X}_n}\big((\sqrt{n}/c)(\widehat{\Sigma}^*)^{-1/2}(\widehat{\theta}^* - \widehat{\theta}) \le x\big) - P_X\big(\sqrt{n}\widehat{\Sigma}^{-1/2}(\widehat{\theta} - \theta_0) \le x\big)\big| \xrightarrow{P^o_X} 0 \tag{32}$$

by the arguments in proving Theorem 1, Slutsky's theorem and Lemma 3. A possible candidate for the consistent $\widehat{\Sigma}^*$ is the block jackknife proposed in [29].

REMARK 3. Our arguments in proving Theorem 1 can also be used to improve the remainder term in (28) from "$o^o_{P_W}(1)$ in P^o_X-probability" to "$O^o_{P_W}(n^{-2\gamma+1/2})$ in P^o_X-probability" if we strengthen the "nearly maximizing" condition (22) to the exactly maximizing condition (21). A similar result holds in Proposition 1 where the remainder term $o^o_{P_X}(1)$ in (16) can be improved to $O_{P_X}(n^{-2\gamma+1/2})$ if (7) is strengthened to (6). It is interesting to note that the rate of convergence of the remainder term depends on how accurately the nuisance function parameter η can be estimated. In particular, if η is $\sqrt{n}$-estimable, then the remainder is of the order of $O(n^{-1/2})$.

The distribution consistency result of the bootstrap estimator $\widehat{\theta}^*$ proven in (30) can be used to prove the consistency of a variety of bootstrap confidence sets, that is, *percentile*, *hybrid* and t types.

A lower αth quantile of bootstrap distribution is any quantity $\tau^*_{n\alpha} \in \mathbb{R}^d$ satisfying $\tau^*_{n\alpha} = \inf\{\varepsilon : P_{W|\mathcal{X}_n}(\widehat{\theta}^* \le \varepsilon) \ge \alpha\}$, where ε is an infimum over the given set only if there does not exist a $\varepsilon_1 < \varepsilon$ in $\mathbb{R}^d$ such that $P_{W|\mathcal{X}_n}(\widehat{\theta}^* \le \varepsilon_1) \ge \alpha$. Because of the assumed smoothness of the criterion function $m(\theta, \eta)$ in our setting, we can, without loss of generality, assume $P_{W|\mathcal{X}_n}(\widehat{\theta}^* \le \tau^*_{n\alpha}) = \alpha$. Due to the distribution consistency result proven in (30), we can approximate the αth quantile of the distribution of $(\widehat{\theta} - \theta_0)$ by $(\tau^*_{n\alpha} - \widehat{\theta})/c$. Thus, we define the *percentile*-type bootstrap confidence set as

$$\mathrm{BC}_p(\alpha) = \left[\widehat{\theta} + \frac{\tau^*_{n(\alpha/2)} - \widehat{\theta}}{c}, \widehat{\theta} + \frac{\tau^*_{n(1-\alpha/2)} - \widehat{\theta}}{c}\right].$$

Similarly, we can approximate the αth quantile of $\sqrt{n}(\widehat{\theta} - \theta_0)$ by $\kappa^*_{n\alpha}$, where $\kappa^*_{n\alpha}$ is the αth quantile of the hybrid quantity $(\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta})$, that is, $P_{W|\mathcal{X}_n}((\sqrt{n}/c) \times (\widehat{\theta}^* - \widehat{\theta}) \le \kappa^*_{n\alpha}) = \alpha$. Thus, we define the *hybrid*-type bootstrap confidence set as

$$\mathrm{BC}_h(\alpha) = \left[\widehat{\theta} - \frac{\kappa^*_{n(1-\alpha/2)}}{\sqrt{n}}, \widehat{\theta} - \frac{\kappa^*_{n(\alpha/2)}}{\sqrt{n}}\right].$$

Note that $\tau^*_{n\alpha}$ and $\kappa^*_{n\alpha}$ are not unique since θ is assumed to be a vector.

We now prove the consistency of the above bootstrap confidence sets by using the arguments in Lemma 23.3 of [39]. First, it follows from (17) and (29) that, for any $x \in \mathbb{R}^d$,

$$P_X\big(\sqrt{n}(\widehat{\theta} - \theta_0) \le x\big) \longrightarrow \Psi(x), \tag{33}$$

$$P_{W|\mathcal{X}_n}\big((\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta}) \le x\big) \xrightarrow{P_X^o} \Psi(x), \tag{34}$$

where $\Psi(x) = P(N(0, \Sigma) \le x)$. The quantile convergence theorem, that is, Lemma 21.1 in [39], applied to (34) implies that $\kappa^*_{n\alpha} \to \Psi^{-1}(\alpha)$ almost surely. When applying quantile convergence theorem, we use the almost sure representation Theorem 2.19 in [39] and argue along subsequences. Then the Slutsky's lemma implies that $\sqrt{n}(\widehat{\theta} - \theta_0) - \kappa^*_{n(\alpha/2)}$ weakly converges to $N(0, \Sigma) - \Psi^{-1}(\alpha/2)$. Thus,

$$\begin{aligned} P_{XW}\left(\theta_0 \le \widehat{\theta} - \frac{\kappa^*_{n(\alpha/2)}}{\sqrt{n}}\right) &= P_{XW}\big(\sqrt{n}(\widehat{\theta} - \theta_0) \ge \kappa^*_{n(\alpha/2)}\big) \\ &\to P_{XW}\big(N(0, \Sigma) \ge \Psi^{-1}(\alpha/2)\big) \\ &= 1 - \alpha/2. \end{aligned}$$

This argument yields the consistency of the *hybrid*-type bootstrap confidence set, that is, (36) below, and can also be applied to justify the *percentile*-type bootstrap confidence set, that is, (35) below. The following Corollary 1 summarizes the above discussion.

COROLLARY 1. *Under the conditions in Theorem* 1, *we have*

$$P_{XW}\big(\theta_0 \in \mathrm{BC}_p(\alpha)\big) \longrightarrow 1 - \alpha, \tag{35}$$

$$P_{XW}\big(\theta_0 \in \mathrm{BC}_h(\alpha)\big) \longrightarrow 1 - \alpha \tag{36}$$

as $n \to \infty$.

It is well known that the above bootstrap confidence sets can be computed easily through routine bootstrap sampling.

Investigating the consistency of the bootstrap variance estimator is also of great interest. However, the usual sufficient condition for moment consistency, that is, uniform integrability condition, becomes very hard to verify due to the existence of an infinite-dimensional parameter η. An alternative resampling method to obtain the variance estimator in semiparametric models is the block jackknife approach, which was proposed and theoretically justified in [29]. We do not pursue this topic further in this paper.

REMARK 4. Provided consistent variance estimators $\widehat{\Sigma}^*$ and $\widehat{\Sigma}$ are available, we can define the t-type bootstrap confidence set as

$$\mathrm{BC}_t(\alpha) = \left[\widehat{\theta} - \frac{\widehat{\Sigma}^{1/2}\omega^*_{n(1-\alpha/2)}}{\sqrt{n}}, \widehat{\theta} - \frac{\widehat{\Sigma}^{1/2}\omega^*_{n(\alpha/2)}}{\sqrt{n}}\right],$$

where $\omega^*_{n\alpha}$ satisfies $P_{W|\mathcal{X}_n}((\sqrt{n}/c)(\widehat{\Sigma}^*)^{-1/2}(\widehat{\theta}^* - \widehat{\theta}) \le \omega^*_{n\alpha}) = \alpha$. By applying again the arguments in Lemma 23.3 of [39] to (19) and (32), we can prove that

$$P_{XW}\big(\theta_0 \in \mathrm{BC}_t(\alpha)\big) \longrightarrow 1 - \alpha$$

as $n \to \infty$.

4. Verifications of conditions S1, S2 and SB1, SB2.

4.1. *Verifications of conditions* S1 *and* S2. The continuity modulus condition (12) in S1 can be checked via one of the following two approaches. The first approach is to show the boundedness of $E^o_X\|\mathbb{G}_n\|_{\mathcal{S}_n}$ by using Lemma 3.4.2 in [38]. The second approach is to calculate the bracketing entropy number of $\mathcal{S}_n$ and apply Lemma 5.13 in [40] if L_2-norm is used on the nuisance parameter. As for (13), we can verify it easily if we can show that the class of functions $\{(\partial/\partial\theta)\widetilde{m}(\theta,\eta) : (\theta,\eta) \in \mathcal{C}_n\}$ is P-Donsker.

Next, we discuss how to verify the smoothness condition S2. We first write $P_X(\widetilde{m}(\theta,\eta) - \widetilde{m}_0)$ as the sum of $P_X(\widetilde{m}(\theta,\eta) - \widetilde{m}(\theta_0,\eta))$ and $P_X(\widetilde{m}(\theta_0,\eta) - \widetilde{m}_0)$. We apply the Taylor expansion to obtain

$$\begin{aligned} &P_X\big(\widetilde{m}(\theta,\eta) - \widetilde{m}(\theta_0,\eta)\big) \\ &\quad = P_X\{m_{11}(\theta_0,\eta) - m_{21}(\theta_0,\eta)[H^\dagger(\theta_0,\eta)]\}(\theta - \theta_0) + O(\|\theta - \theta_0\|^2) \\ &\quad = A(\theta - \theta_0) + (\theta - \theta_0)O(\|\eta - \eta_0\|) + O(\|\theta - \theta_0\|^2), \end{aligned}$$

where A is defined in (8), the first and second equality follows from the Taylor expansion of $\theta \mapsto P_X\widetilde{m}(\theta,\eta)$ around θ_0 and

$$\eta \mapsto P_X\{m_{11}(\theta_0,\eta) - m_{21}(\theta_0,\eta)[H^\dagger(\theta_0,\eta)]\}$$

around η_0, respectively. By applying the second-order Taylor expansion to $\eta \mapsto P_X\widetilde{m}(\theta_0,\eta)$ around η_0 and considering (4), we can show that $P(\widetilde{m}(\theta_0,\eta) - \widetilde{m}_0) = O(\|\eta - \eta_0\|^2)$. In summary, condition S2 usually holds in models where the map $\eta \mapsto \widetilde{m}(\theta_0,\eta)$ is smooth in the sense that the Fréchet derivative of $\eta \mapsto P_X((\partial/\partial\theta)\widetilde{m}(\theta_0,\eta))$ around η_0 and the second order Fréchet derivative of $\eta \mapsto P_X\widetilde{m}(\theta_0,\eta)$ around η_0 are bounded as discussed above.

4.2. *Verifications of conditions* SB1 *and* SB2. We can verify condition SB1 by showing either $S_n(x)$ is uniformly bounded, that is, $\limsup_{n\to\infty} S_n(x) \le M < \infty$ for every $x \in \mathcal{X}$, or more generally, $\limsup_{n\to\infty} E[\{S_n(X_1)\}^{2+\delta}] < \infty$ for some $\delta > 0$. That the moment condition implies condition SB1 follows from the Chebyshev's inequality. In our examples in Section 6, the uniformly boundedness condition is usually satisfied. Hence, we focus on how to show $S_n(x)$ is uniformly bounded here. By the Taylor expansion in a Banach space, we can write $\widetilde{m}(\theta_0, \eta) - \widetilde{m}_0 = D_{\widetilde{\eta}}[\eta - \eta_0]$, where $\widetilde{\eta}$ lies on the line segment between η and η_0, and $D_\xi[h]$ is the Fréchet derivative of $\eta \mapsto \widetilde{m}(\theta_0, \eta)$ at ξ along the direction h. Since we require $\|\eta - \eta_0\| \le \delta_n \to 0$, the bounded Fréchet derivative at η_0 will imply that $S_n(x)$ is uniformly bounded. The method in verifying (13) of condition S1 can be applied to check condition SB2; see the discussion in the previous subsection.

5. Convergence rates of bootstrap estimate of functional parameter. In this section, we present two general theorems for calculating the convergence rate of the bootstrap estimate of the functional parameter. These results can be applied to verify condition SB3. Condition S3 can also be verified based on these theorems by assuming the weights $W_{ni} = 1$. Note that both theorems extend general results on M-estimators [31, 38] to bootstrap M-estimators and are also of independent interest. Separate treatments are given to the cases that the estimate η has $\sqrt{n}$ convergence rate, that is, Section 5.1, and has slower than $\sqrt{n}$ rate, that is, Section 5.2.

5.1. *Root-n rate.* We consider a collection of measurable objective functions $x \mapsto k(\theta, \eta)[g](x)$ indexed by the parameter $(\theta, \eta) \in \Theta \times \mathcal{H}$ and an arbitrary index set $g \in \mathbf{G}$. For example, $k(\theta, \eta)[g]$ can be the score function for η given any fixed θ indexed by $g \in \mathbf{G}$. Define

$$U_n^*(\theta, \eta)[g] = \mathbb{P}_n^* k(\theta, \eta)[g],$$
$$U_n(\theta, \eta)[g] = \mathbb{P}_n k(\theta, \eta)[g],$$
$$U(\theta, \eta)[g] = P_X k(\theta, \eta)[g].$$

We assume that the maps $g \mapsto U_n^*(\theta, \eta)[g]$, $g \mapsto U_n(\theta, \eta)[g]$ and $g \mapsto U(\theta, \eta)[g]$ are uniformly bounded, so that U_n^*, U_n and U are viewed as maps from the parameter set $\Theta \times \mathcal{H}$ into $\ell^\infty(\mathbf{G})$. The following conditions are assumed in Theorem 2 below:

$$\{k(\theta, \eta)[g] : \|\theta - \theta_0\| + \|\eta - \eta_0\| \le \delta, g \in \mathbf{G}\} \in M(P_X) \cap L_2(P_X) \tag{37}$$

and is P-Donsker for some $\delta > 0$,

$$\sup_{g\in\mathbf{G}} P_X\{k(\theta, \eta)[g] - k(\theta_0, \eta_0)[g]\}^2 \to 0 \qquad \text{as } \|\theta - \theta_0\| + \|\eta - \eta_0\| \to 0. \tag{38}$$

Let

$$\mathcal{D}_n = \left\{ \frac{k(\theta,\eta)[g] - k(\theta_0,\eta_0)[g]}{1+\sqrt{n}\|\theta-\theta_0\|+\sqrt{n}\|\eta-\eta_0\|} : g \in \mathbf{G}, \|\theta-\theta_0\|+\|\eta-\eta_0\| \le \delta_n \right\}$$

and $D_n(X)$ be the envelop function of the class of functions $\mathcal{D}_n$. For any sequence $\delta_n \to 0$, we assume that $D_n(X)$ satisfies

$$\lim_{\lambda\to\infty} \lim_{n\to\infty} \sup_{t\ge\lambda} t^2 P_X^o\big(D_n(X_1) > t\big) = 0. \tag{39}$$

Now we consider the convergence rate of $\widehat{\eta}_{\widetilde{\theta}}^*$ satisfying:

$$U_n^*(\widetilde{\theta}, \widehat{\eta}_{\widetilde{\theta}}^*)[g] = O_{P_{XW}}^o(n^{-1/2}) \tag{40}$$

for any $\widetilde{\theta} \xrightarrow{P_{XW}^o} \theta_0$ and g ranging over $\mathbf{G}$. In Theorem 2 below, we will show that $\widehat{\eta}_{\widetilde{\theta}}^*$ has the root-n convergence rate under conditions (37)–(39).

THEOREM 2. *Suppose that $U : \Theta \times \mathcal{H} \mapsto \ell^\infty(\mathbf{G})$ is Fréchet differentiable at (θ_0, η_0) with bounded derivative $\dot{U} : \mathbb{R}^d \times \operatorname{lin}\mathcal{H} \mapsto \ell^\infty(\mathbf{G})$ such that the map $\dot{U}(0,\cdot) : \operatorname{lin}\mathcal{H} \mapsto \ell^\infty(\mathbf{G})$ is invertible with an inverse that is continuous on its range. Furthermore, assume that (37)–(39) hold, and that $U(\theta_0,\eta_0) = 0$, then*

$$\|\widehat{\eta}_{\widetilde{\theta}}^* - \eta_0\| = O_{P_W}^o(\|\widetilde{\theta}-\theta_0\| \vee n^{-1/2}) \tag{41}$$

in P_X^o-probability, given that $\widetilde{\theta} \xrightarrow{P_{XW}^o} \theta_0$ and $\widehat{\eta}_{\widetilde{\theta}}^* \xrightarrow{P_{XW}^o} \eta_0$.

The proof of Theorem 2 is given in Appendix A.4.

5.2. *Slower than root-n rate.* We next present a result that deals with slower than $\sqrt{n}$ convergence rate for the bootstrap M-estimate of the functional parameter. This result is so general that it can be applied to the sieve estimate of nuisance parameter [15]. The essence of the sieve method is that a sequence of increasing spaces (sieves), that is, $\mathcal{H}_n$, is employed to approximate the large parameter space, for example, $\mathcal{H}$. In other words, for any $\eta \in \mathcal{H}$, there exists a $\pi_n\eta \in \mathcal{H}_n$ such that $\|\eta - \pi_n\eta\| \to 0$ as $n \to \infty$.

Now, we consider the M-estimate $\widehat{\eta}_\theta^* \in \mathcal{H}_n$ satisfying

$$\mathbb{P}_n^* v(\theta, \widehat{\eta}_\theta^*) \ge \mathbb{P}_n^* v(\theta, \eta_n) \qquad \text{for any } \theta \in \Theta \text{ and some } \eta_n \in \mathcal{H}_n, \tag{42}$$

where $x \mapsto v(\theta,\eta)(x)$ is a measurable objective function. Let "$\lesssim$" and "$\gtrsim$" denote greater than or smaller than, up to an universal constant. We assume the following conditions hold for every $\delta > 0$:

$$E_X\big(v(\theta,\eta) - v(\theta,\eta_n)\big) \lesssim -d^2(\eta,\eta_n) + \|\theta-\theta_0\|^2, \tag{43}$$

$$E_X^o \sup_{\theta\in\Theta, \eta\in\mathcal{H}_n, \|\theta-\theta_0\|\le\delta, d(\eta,\eta_n)\le\delta} \big|\mathbb{G}_n\big(v(\theta,\eta) - v(\theta,\eta_n)\big)\big| \lesssim \psi_n(\delta), \tag{44}$$

$$E_{XW}^o \sup_{\theta\in\Theta, \eta\in\mathcal{H}_n, \|\theta-\theta_0\|\le\delta, d(\eta,\eta_n)\le\delta} \big|\mathbb{G}_n^*\big(v(\theta,\eta) - v(\theta,\eta_n)\big)\big| \lesssim \psi_n^*(\delta). \tag{45}$$

Here $d^2(\eta, \eta_n)$ may be thought of as the square of a distance, for example, $\|\eta - \eta_n\|^2$, but our theorem is also true for any arbitrary function $\eta \mapsto d^2(\eta, \eta_n)$.

THEOREM 3. *Suppose that conditions* (43)–(45) *hold. We assume* (44) [*and* (45)] *is valid for functions* ψ_n (ψ_n^*) *such that* $\delta \mapsto \psi_n(\delta)/\delta^\alpha$ $[\delta \mapsto \psi_n^*(\delta)/\delta^\alpha]$ *is decreasing for some* $0 < \alpha < 2$. *Then for every* $(\widetilde{\theta}, \widehat{\eta}_{\widetilde{\theta}}^*)$ *satisfying* $P(\widetilde{\theta} \in \Theta, \widehat{\eta}_{\widetilde{\theta}}^* \in \mathcal{H}_n) \to 1$, *we have*

$$d(\widehat{\eta}_{\widetilde{\theta}}^*, \eta_n) \le O_{P_W}^o(\delta_n \vee \|\widetilde{\theta} - \theta_0\|)$$

in P_X^o*-probability, for any sequence of positive numbers* δ_n *satisfying both* $\psi_n(\delta_n) \le \sqrt{n}\delta_n^2$ *and* $\psi_n^*(\delta_n) \le \sqrt{n}\delta_n^2$ *for large* n.

The proof of Theorem 3 is given in Appendix A.5.

In application of Theorem 3, the parameter η_n is taken to be some element in $\mathcal{H}_n$ that is very close to η_0. When $\mathcal{H}_n = \mathcal{H}$, a natural choice for η_n is η_0 and we can directly use Theorem 3 to derive the convergence rate $d(\widehat{\eta}_{\widetilde{\theta}}^*, \eta_0)$ as shown in the examples of Section 6. In general, η_n may be taken as the maximizer of the mapping $\eta \mapsto P_X v(\theta_0, \eta)$ over $\mathcal{H}_n$, the projection of η_0 onto $\mathcal{H}_n$. Then we need to consider the approximation rate of the sieve space $\mathcal{H}_n$ to $\mathcal{H}$, that is, $d(\eta_n, \eta_0)$, since $d(\widehat{\eta}_{\widetilde{\theta}}^*, \eta_0) \le d(\widehat{\eta}_{\widetilde{\theta}}^*, \eta_n) + d(\eta_n, \eta_0)$. The approximation rate $d(\eta_n, \eta_0)$ depends on the choices of sieves and is usually derived in the mathematical literature.

Now we discuss verification of the nontrivial conditions (43)–(45). The smoothness condition for $v(\theta, \eta)$, that is, (43), is implied by

$$E_X\big(v(\theta, \eta) - v(\theta_0, \eta_n)\big) \lesssim -d^2(\eta, \eta_n) - \|\theta - \theta_0\|^2, \tag{46}$$

$$E_X\big(v(\theta, \eta_n) - v(\theta_0, \eta_n)\big) \gtrsim -\|\theta - \theta_0\|^2. \tag{47}$$

The two conditions depict the quadratic behaviors of the criterion functions $(\theta, \eta) \mapsto E_X v(\theta, \eta)$ and $\theta \mapsto E_X v(\theta, \eta_n)$ around the maximum point (θ_0, η_n) and θ_0, respectively. We next present one useful lemma for verifying the continuity modulus of (bootstrapped) empirical processes, that is, (44) and (45). Denote

$$\mathcal{V}_\delta = \{x \mapsto [v(\theta, \eta)(x) - v(\theta, \eta_n)(x)] : d(\eta, \eta_n) \le \delta, \|\theta - \theta_0\| \le \delta\} \tag{48}$$

and define the bracketing entropy integral of $\mathcal{V}_\delta$ as

$$K(\delta, \mathcal{V}_\delta, L_2(P_X)) = \int_0^\delta \sqrt{1 + \log N_{[\cdot]}(\varepsilon, \mathcal{V}_\delta, L_2(P_X))}\, d\varepsilon, \tag{49}$$

where $\log N_{[\cdot]}(\delta, \mathcal{A}, d)$ is the δ-bracketing entropy number for the class $\mathcal{A}$ under the distance measure d.

LEMMA 1. *Suppose that the functions* $(x, \theta, \eta) \mapsto v_{\theta,\eta}(x)$ *are uniformly bounded for* (θ, η) *ranging over some neighborhood of* (θ_0, η_n) *and that*

$$E_X(v_{\theta,\eta} - v_{\theta,\eta_n})^2 \lesssim d^2(\eta, \eta_n) + \|\theta - \theta_0\|^2. \tag{50}$$

Then condition (44) *is satisfied for any functions* ψ_n *such that*

$$\psi_n(\delta) \geq K(\delta, \mathcal{V}_\delta, L_2(P_X))\left(1 + \frac{K(\delta, \mathcal{V}_\delta, L_2(P_X))}{\delta^2\sqrt{n}}\right). \tag{51}$$

Let $V_n(X)$ *be the envelop function of the class* $\mathcal{V}_{\delta_n}$. *If we further assume that, for each sequence* $\delta_n \to 0$, *the envelop functions* V_n *satisfies*

$$\lim_{\lambda\to\infty} \lim_{n\to\infty} \sup_{t\geq\lambda} t^2 P_X^o\big(V_n(X_1) > t\big) = 0, \tag{52}$$

then condition (45) *is satisfied for any functions* ψ_n^* *such that*

$$\psi_n^*(\delta) \geq K(\delta, \mathcal{V}_\delta, L_2(P_X))\left(1 + \frac{K(\delta, \mathcal{V}_\delta, L_2(P_X))}{\delta^2\sqrt{n}}\right). \tag{53}$$

REMARK 5. Note that the inequalities $\psi_n(\delta) \lesssim \sqrt{n}\delta^2$ and $\psi_n^*(\delta) \lesssim \sqrt{n}\delta^2$ are equivalent to $K(\delta, \mathcal{V}_\delta, L_2(P_X)) \lesssim \sqrt{n}\delta^2$ when we let ψ_n and ψ_n^* be equal to the right-hand side of (51) and (53), respectively. Consequently, the convergence rate of $\widehat{\eta}_{\widehat{\theta}}^*$ calculated in Theorem 3, that is, δ_n, is determined by the bracketing entropy integral of $\mathcal{V}_{\delta_n}$.

REMARK 6. The assumptions of Lemma 1 are relaxable to great extent. For example, we can drop the uniform bounded condition on the class of functions $v(\theta, \eta)$ by using the "Bernstein norm," that is, $\|f\|_{P,B} = (2P(e^{|f|} - 1 - |f|))^{1/2}$, instead of the L_2-norm. In some cases, the bracketing entropy integral diverges at zero. Then we can change the limit of the integration in (49) from $[0, \delta]$ to $[a\delta^2 \wedge \delta/3, \delta]$ for some small positive constant a, see Lemma 3.4.3 and page 326 in [38].

6. Examples. In this section, we apply the main results in Section 3 to justify the bootstrap validity of drawing semiparametric inferences in three examples of semiparametric models. In the Cox regression models with censored data, we use the log-likelihood as the criterion function, while in the partially linear model, the least squares criterion is used. The M-estimate of the nuisance functional parameters have different convergence rates in these examples. Indeed, the advantages of using bootstrap approach in all of the three examples were considered in the literature, for example, [14, 26]. This section also serves the purpose of illustration on verification of the technical conditions used in the general results.

6.1. *Cox regression model with right censored data.* In the Cox regression model, the hazard function of the survival time T of a subject with covariate Z is modeled as

$$\lambda(t|z) \equiv \lim_{\Delta\to 0} \frac{1}{\Delta} P(t \leq T < t + \Delta | T \geq t, Z = z) = \lambda(t)\exp(\theta' z), \tag{54}$$

where λ is an unspecified baseline hazard function and θ is a regression vector. In this model, we are usually interested in θ while treating the cumulative hazard function $\eta(y) = \int_0^y \lambda(t)\,dt$ as the nuisance parameter. The MLE for θ is proven to be semiparametric efficient and widely used in applications. Here we consider bootstrapping $\widehat{\theta}$, which corresponds to treating log-likelihood as the criterion function $m(\theta, \eta)$ in our general formulation.

With right censoring of survival time, the data observed is $X = (Y, \delta, Z)$, where $Y = T \wedge C$, C is a censoring time, $\delta = I\{T \le C\}$, and Z is a regression covariate belonging to a compact set $\mathbb{Z} \subset \mathbb{R}^d$. We assume that C is independent of T given Z. The log-likelihood is obtained as

$$m(\theta, \eta) = \delta\theta' z - \exp(\theta' z)\eta(y) + \delta \log \eta\{y\}, \tag{55}$$

where $\eta\{y\} = \eta(y) - \eta(y-)$ is a point mass that denotes the jump of η at point y. The parameter space $\mathcal{H}$ is restricted to a set of nondecreasing cadlag functions on the interval $[0, \tau]$ with $\eta(\tau) \le M$ for some constant M. By some algebra, we have

$$\begin{aligned}\widetilde{m}(\theta, \eta)(x) &= m_1(\theta, \eta) - m_2(\theta, \eta)[H^\dagger(\theta, \eta)] \\ &= [\delta z - z\exp(\theta' z)\eta(y)] \\ &\quad - \left[\delta H^\dagger(\theta, \eta)(y) - \exp(\theta' z)\int_0^y H^\dagger(\theta, \eta)(u)\,d\eta(u)\right],\end{aligned}$$

where

$$H^\dagger(\theta, \eta)(y) = \frac{E_{\theta,\eta} Z\exp(\theta' Z)1\{Y \ge y\}}{E_{\theta,\eta}\exp(\theta' Z)1\{Y \ge y\}}.$$

Conditions I, S1–S3 in guaranteeing the asymptotic normality of $\widehat{\theta}$ have been verified in [8]. In particular, the convergence rate of the estimated nuisance parameter is established in Theorem 3.1 of [31], that is,

$$\|\widehat{\eta}_{\widetilde{\theta}_n} - \eta_0\|_\infty = O_{P_X}(n^{-1/2} + \|\widetilde{\theta}_n - \theta_0\|), \tag{56}$$

where $\|\cdot\|_\infty$ denotes the supreme norm. We next verify the bootstrap consistency conditions, that is, SB1–SB3. Condition SB1 trivially holds since it is easy to show that $\eta \mapsto \widetilde{m}(\theta_0, \eta)$ has bounded Fréchet derivative around η_0. The P-Donsker condition SB2 has been verified when verifying (13) in condition S1. In the end, we will verify the bootstrap convergence rate condition $\|\widehat{\eta}^*_{\widetilde{\theta}} - \eta_0\|_\infty = O^o_{P_{XW}}(\|\widetilde{\theta} - \theta_0\| \vee n^{-1/2})$ via Theorem 2. Since $\widehat{\eta}^*_\theta$ maximizes $\mathbb{P}^*_n m(\theta, \eta)$ for fixed θ, we set $k(\theta, \eta)[g] = m_2(\theta, \eta)[g]$ and have $U^*_n(\theta, \widehat{\eta}^*_\theta)[g] = \mathbb{P}^*_n m_2(\theta, \widehat{\eta}^*_\theta)[g] = 0$. The invertibility of $\dot{W}(0, \cdot)$, conditions (37) and (38) have been verified in [31] when they showed (56). Now we only need to consider condition (39): for n so

large that $\delta_n \le R$

$$\begin{aligned} D_n(x) &\equiv \sup\Biggl\{ \frac{|(m_2(\theta,\eta)[g]) - m_2(\theta_0,\eta_0)[g]|}{1+\sqrt{n}(\|\theta-\theta_0\| + \|\eta-\eta_0\|_\infty)}, g \in \mathbf{G}, \\ &\qquad\qquad \|\theta-\theta_0\| + \|\eta-\eta_0\|_\infty \le \delta_n \Biggr\} \\ &\le 2\sup\{|m_2(\theta,\eta)[g]|, g \in \mathbf{G}, \|\theta-\theta_0\| + \|\eta-\eta_0\|_\infty \le R\} \\ &\le \text{some constant.} \end{aligned}$$

The last inequality follows from the assumption that $\mathbf{G}$ is a class of functions of bounded total variation and the inequality that $\int_0^y g(u)\,d\eta(u) \le \eta(\tau)\|g\|_{\mathrm{BV}}$, where $\|g\|_{\mathrm{BV}}$ is the total variation of the function g. Thus, condition (39) holds trivially.

6.2. *Cox regression model with current status data.* We next consider the current status data when each subject is observed at a single examination time C to determine if an event has occurred. The event time T cannot be known exactly. Then the observed data are n i.i.d. realizations of $X = (C, \delta, Z) \in R^+ \times \{0,1\} \times \mathbb{Z}$, where $\delta = I\{T \le C\}$. The corresponding criterion function, that is, the log-likelihood, is derived as

$$m(\theta,\eta) = \delta \log[1 - \exp(-\eta(c)\exp(\theta' z))] - (1-\delta)\exp(\theta' z)\eta(c). \tag{57}$$

We make the following assumptions throughout the rest of this subsection: (i) T and C are independent given Z; (ii) the covariance of $Z - E(Z|C)$ is positive definite, which guarantees the efficient information to be positive definite; (iii) C possesses a Lebesgue density which is continuous and positive on its support $[\sigma,\tau]$, for which the true nuisance parameter η_0 satisfies $\eta_0(\sigma-) > 0$ and $\eta_0(\tau) < M < \infty$, and this density is continuously differentiable on $[\sigma,\tau]$ with derivative bounded above and bounded below by zero. The form of $\widetilde{m}(\theta,\eta)$ can be found in [9] as follows

$$\begin{aligned} \widetilde{m}(\theta,\eta) &= m_1(\theta,\eta) - m_2(\theta,\eta)[H^\dagger(\theta,\eta)] \\ &= \bigl(z\eta(c) - H^\dagger(\theta,\eta)(c)\bigr) Q(x;\theta,\eta), \end{aligned}$$

where

$$Q(x;\theta,\eta) = e^{\theta' z}\biggl[\frac{\delta}{\exp(e^{\theta' z}\eta(c)) - 1} - (1-\delta)\biggr]$$

and the form of $H^\dagger(\theta,\eta)(c)$ is given in (4) of [9].

Conditions I and S1–S3 are verified in [9]. Conditions SB1 and SB2 can be checked similarly as in the previous example. Note that the convergence rate for the nuisance parameter becomes slower, that is,

$$\|\widehat{\eta}_{\widetilde{\theta}_n} - \eta_0\|_2 = O_{P_X}(\|\widetilde{\theta}_n - \theta_0\| + n^{-1/3}), \tag{58}$$

where $\|\cdot\|_2$ denotes the regular L_2-norm, as shown in [31]. By Theorem 3, we can show that the same convergence rate, that is, $n^{-1/3}$, also holds for $\widehat{\eta}_\theta^*$. The assumptions (43) and (44) in Theorem 3 are verified in [31] when showing (58). We apply Lemma 1 to verify assumption (45). We show that condition (52) on the envelop function $V_n(x)$ holds: for n so large that $\delta_n \le R$

$$\begin{aligned} V_n(x) &\equiv \sup\{|m(\theta,\eta) - m(\theta,\eta_0)| : \|\eta-\eta_0\|_2 \le \delta_n, \|\theta-\theta_0\| \le \delta_n\} \\ &\le 2\sup\{|m(\theta,\eta)| : \|\eta-\eta_0\|_2 \le R, \|\theta-\theta_0\| \le R\} \\ &\le \text{some constant.} \end{aligned}$$

6.3. *Partially linear models.* In this example, a continuous outcome variable Y, depending on the covariates $(W, Z) \in [0,1]^2$, is modeled as

$$Y = \theta W + f(Z) + \xi,$$

where ξ is independent of (W, Z) and f is an unknown smooth function belonging to $\mathcal{H} \equiv \{f : [0,1] \mapsto [0,1], \int_0^1 (f^{(k)}(u))^2\,du \le M\}$ for a fixed $0 < M < \infty$. In addition, we assume $E(\mathrm{Var}(W|Z))$ is positive definite and $E\{f(Z)\} = 0$. We want to estimate (θ, f) using the least square criterion:

$$m(\theta, f) = -\big(y - \theta w - f(z)\big)^2. \tag{59}$$

Note that the above model would be more flexible if we did not require knowledge of M. A sieve estimator could be obtained if we replaced M with a sequence $M_n \to \infty$. The theory we develop in this paper will be applicable in this setting, but, in order to maintain clarity of exposition, we have elected not to pursue this more complicated situation here. Another approach is to use penalization, the study of which is beyond the scope of this paper.

Simple calculations give

$$\begin{aligned} \widetilde{m}(\theta,\eta)(x) &= m_1(\theta,\eta) - m_2(\theta,\eta)[H^\dagger(\theta,\eta)] \\ &= 2\big(y - \theta w - f(z)\big)\big(w - H^\dagger(\theta,\eta)(z)\big), \end{aligned}$$

where

$$H^\dagger(\theta,\eta)(z) = \frac{E_{\theta,\eta}(W(Y-\theta W - f(Z))^2 | Z = z)}{E_{\theta,\eta}((Y-\theta W - f(Z))^2 | Z = z)}.$$

The finite variance condition I follows from $E[W\{W - H^\dagger(\theta_0,\eta_0)(Z)\}] > 0$. The distribution of ξ is assumed to have finite second moment and satisfy (5), for example, $\xi \sim N(0,1)$. Conditions S1–S3 and SB2 can be verified using similar arguments in Example 3 of [9], in particular, $\|\widehat{f}_{\widetilde{\theta}} - f_0\|_2 = O_{P_X}(\|\widetilde{\theta} - \theta_0\| \vee n^{-k/(2k+1)})$ in (15). It is easy to show that the Fréchet derivative of $\eta \mapsto \widetilde{m}(\theta_0,\eta)$ is bounded around η_0, and thus the tail condition SB1 holds. To prove $\|\widehat{f}^*_{\widetilde{\theta}} - f_0\|_2 = O^o_{P_{XW}}(\|\widetilde{\theta} - \theta_0\| \vee n^{-k/(2k+1)})$ via Theorem 3, we proceed as in the previous example, checking assumption (52) using similar arguments, that is, $V_n(x)$ is uniformly bounded.

7. Proof of Theorem 1 (bootstrap consistency theorem). To prove Theorem 1, we need the following lemma whose proof is given in Appendix A.3.

LEMMA 2. *Under the assumptions of Theorem 1, we have*

$$\mathbb{G}_n^*\big(\widetilde{m}(\theta,\eta)-\widetilde{m}(\theta_0,\eta_0)\big)=O_{P_W}^o(\|\theta-\theta_0\|\vee\|\eta-\eta_0\|) \tag{60}$$

in P_X^o-probability for $(\theta,\eta)\in\mathcal{C}_n$.

We shall use repeatedly Lemma 3 in the Appendix, which concerns about the transition of stochastic orders among different probability spaces.

We first prove (27). Recall that $\mathbb{G}_n=\sqrt{n}(\mathbb{P}_n-P_X)$ and $\mathbb{G}_n^*=\sqrt{n}(\mathbb{P}_n^*-\mathbb{P}_n)$. Define $\widehat{m}^*$ as $\widetilde{m}(\widehat{\theta}^*,\widehat{\eta}^*)$. By some algebra, we have

$$\begin{aligned}&\mathbb{G}_n^*\widetilde{m}_0+\mathbb{G}_n\widetilde{m}_0+\sqrt{n}P_X(\widehat{m}^*-\widetilde{m}_0)\\&\quad=\mathbb{G}_n^*(\widetilde{m}_0-\widehat{m}^*)+\mathbb{G}_n(\widetilde{m}_0-\widehat{m}^*)+\sqrt{n}\mathbb{P}_n^*\widehat{m}^*,\end{aligned}$$

since $P_X\widetilde{m}_0=0$. Thus, we have the following inequality:

$$\begin{aligned}\big\|\sqrt{n}P_X(\widehat{m}^*-\widetilde{m}_0)\big\|&\le\|\mathbb{G}_n^*\widetilde{m}_0\|+\|\mathbb{G}_n\widetilde{m}_0\|+\|\mathbb{G}_n^*(\widetilde{m}_0-\widehat{m}^*)\|\\&\quad+\|\mathbb{G}_n(\widetilde{m}_0-\widehat{m}^*)\|+\big\|\sqrt{n}\mathbb{P}_n^*\widehat{m}^*\big\|\\&\equiv L_1+L_2+L_3+L_4+L_5.\end{aligned}\tag{61}$$

Based on Theorem 2.2 in [33], we have $L_1=O_{P_W}^o(1)$ in P_X^o-probability. The CLT implies $L_2=O_{P_X}^o(1)$. We next consider L_3 and L_4. By condition SB3, we can show that $\|\widehat{\eta}^*-\eta_0\|=o_{P_W}^o(1)$ in P_X^o-probability since $\widehat{\theta}^*$ is assumed to be consistent, that is, $\|\widehat{\theta}^*-\theta_0\|=o_{P_W}^o(1)$ in P_X^o-probability, and by (69) and (73) in Lemma 3. Then, we have $L_3=o_{P_W}^o(1)$ in P_X^o-probability based on Lemma 2 and (73) in Lemma 3. Next, we obtain that $L_4=o_{P_W}^o(1)$ in P_X^o-probability based on condition S1 and (71) in Lemma 3. Finally, $L_5=o_{P_{XW}}^o(1)$ based on (22). In summary, (61) can be rewritten as:

$$\big\|\sqrt{n}P_X(\widehat{m}^*-\widetilde{m}_0)\big\|\le O_{P_W}^o(1)+O_{P_X}^o(1) \tag{62}$$

in P_X^o-probability.

Let $\alpha_n=\|\widehat{\theta}^*-\theta_0\|$. Combining (14) with (62) and noticing (26), we have

$$\sqrt{n}\|A\alpha_n\|\le O_{P_W}^o(1)+O_{P_X}^o(1)+O_{P_W}^o\big(\sqrt{n}\alpha_n^2\vee n^{-2\gamma+1/2}\big) \tag{63}$$

in P_X^o-probability. By considering the consistency of $\widehat{\theta}^*$ and condition I, we complete the proof of (27) based on (63).

We next prove (28). Write

$$\begin{aligned}I_1&=-\mathbb{G}_n^*(\widehat{m}^*-\widetilde{m}_0)=\sqrt{n}(\mathbb{P}_n^*-\mathbb{P}_n)(\widetilde{m}_0-\widehat{m}^*),\\I_2&=\mathbb{G}_n(\widehat{m}-\widetilde{m}_0)=\sqrt{n}(\mathbb{P}_n-P_X)(\widehat{m}-\widetilde{m}_0),\\I_3&=-\mathbb{G}_n(\widehat{m}^*-\widetilde{m}_0)=\sqrt{n}(\mathbb{P}_n-P_X)(\widetilde{m}_0-\widehat{m}^*),\\I_4&=\sqrt{n}\mathbb{P}_n^*\widehat{m}^*-\sqrt{n}\mathbb{P}_n\widehat{m}.\end{aligned}$$

By some algebra, we obtain that $\sqrt{n}P_X(\widehat{m}^* - \widehat{m}) + \mathbb{G}_n^*\widetilde{m}_0 = \sum_{j=1}^4 I_j$.

By the definition (24), we can show that $A_n \times B_n = O_{P_W}^o(1)$ in P_X^o-probability if A_n and B_n are both of the order $O_{P_W}^o(1)$ in P_X^o-probability. Then the root-n consistency of $\widehat{\theta}^*$ proven in (27) together with SB3 implies

$$\|\widehat{\eta}^* - \eta_0\| \vee \|\widehat{\theta}^* - \theta_0\| = O_{P_W}^*(n^{-\gamma}) \tag{64}$$

in P_X^o-probability. Thus, by Lemma 2, we know $I_1 = O_{P_W}^o(n^{-\gamma})$ in P_X^o-probability. Note that (12) and (13) of condition S1 imply

$$\mathbb{G}_n\big(\widetilde{m}(\theta, \eta) - \widetilde{m}_0\big) = O_{P_X}^o(\|\theta - \theta_0\| \vee \|\eta - \eta_0\|) \tag{65}$$

for (θ, η) in the shrinking neighborhood $\mathcal{C}_n$ of (θ_0, η_0). Considering (65), S3 and Proposition 1, we have $I_2 = O_{P_X}^o(n^{-\gamma})$. By (64), (65) and (72), we know the order of I_3 is $O_{P_W}^o(n^{-\gamma})$ in P_X^o-probability. We also obtain $I_4 = o_{P_X}^o(1) + o_{P_{XW}}^o(1)$ by using (7) and (22).

Therefore, we have established

$$\sqrt{n}P_X(\widehat{m}^* - \widehat{m}) = -\mathbb{G}_n^*\widetilde{m}_0 + o_{P_X}^o(1) + o_{P_W}^o(1) \tag{66}$$

in P_X^o-probability. To analyze the left-hand side of (66), we rewrite it as $\sqrt{n}P_X(\widehat{m}^* - \widetilde{m}_0) - \sqrt{n}P_X(\widehat{m} - \widetilde{m}_0)$. Applying condition S2, we obtain

$$\begin{aligned}
&\sqrt{n}P_X\big(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[H_0^\dagger]\big)(\widehat{\theta}^* - \widehat{\theta}) \\
&\quad = -\mathbb{G}_n^*\widetilde{m}_0 + o_{P_X}^o(1) + o_{P_W}^o(1) + O_{P_X}^o(n^{1/2-2\gamma}) + O_{P_W}^o(n^{1/2-2\gamma}) \\
&\quad = -\mathbb{G}_n^*\widetilde{m}_0 + o_{P_X}^o(1) + o_{P_W}^o(1)
\end{aligned} \tag{67}$$

in P_X^o-probability, by considering condition S3, SB3 and the range of γ. Note that $o_{P_X}^o(1)$ in (67) is also of the order $o_{P_{XW}}^o(1)$, and thus is of the order $o_{P_W}^o(1)$ in P_X^o-probability by (69). Moreover, according to condition I we have that $A = P_X(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[H_0^\dagger])$ is nonsingular. We obtain (28) by multiplying A^{-1} on both sides of (67).

By applying Lemma 4.6 in [33] under the bootstrap weight conditions, we obtain (29). Proposition 1 together with Lemma 2.11 in [39] implies that

$$\sup_{x \in \mathbb{R}^d} \big|P_X\big(\sqrt{n}(\widehat{\theta} - \theta_0) \le x\big) - P\big(N(0, \Sigma) \le x\big)\big| = o(1). \tag{68}$$

Combining (29) and (68), we obtain (30).

APPENDIX

A.1. Measurability and stochastic orders. *Measurability condition* $M(P)$: we say that a class of random functions $\mathcal{F} \in M(P)$ if $\mathcal{F}$ is nearly linearly deviation measurable for P and that both $\mathcal{F}^2$ and $\mathcal{F}'^2$ are nearly linearly supremum measurable for P. Here $\mathcal{F}^2$ and $\mathcal{F}'^2$ denote the classes of squared functions and

squared differences of functions from $\mathcal{F}$, respectively. It is known that if $\mathcal{F}$ is countable, or if $\{\mathbb{P}_n\}_{n=1}^{\infty}$ are stochastically separable in $\mathcal{F}$, or if $\mathcal{F}$ is image admissible Suslin [12], then $\mathcal{F} \in M(P)$. More precise descriptions can be found in pages 853 and 854 of [17].

The following lemma is very important since it accurately describes the transition of stochastic orders among different probability spaces. We implicitly assume the random quantities in Lemma 3 posses enough measurability so that the usual Fubini theorem can be used freely.

LEMMA 3. *Suppose that*

$$Q_n = o^o_{P_W}(1) \qquad \text{in } P^o_X\text{-probability},$$
$$R_n = O^o_{P_W}(1) \qquad \text{in } P^o_X\text{-probability}.$$

We have

$$A_n = o^o_{P_{XW}}(1) \iff A_n = o^o_{P_W}(1) \qquad \text{in } P^o_X\text{-probability}, \tag{69}$$
$$B_n = O^o_{P_{XW}}(1) \iff B_n = O^o_{P_W}(1) \qquad \text{in } P^o_X\text{-probability}, \tag{70}$$
$$C_n = Q_n \times O^o_{P_X}(1) \implies C_n = o^o_{P_W}(1) \qquad \text{in } P^o_X\text{-probability}, \tag{71}$$
$$D_n = R_n \times O^o_{P_X}(1) \implies D_n = O^o_{P_W}(1) \qquad \text{in } P^o_X\text{-probability}, \tag{72}$$
$$E_n = Q_n \times R_n \implies E_n = o^o_{P_W}(1) \qquad \text{in } P^o_X\text{-probability}. \tag{73}$$

PROOF. To verify (69), we have for every $\varepsilon, \nu > 0$,

$$\begin{aligned} P^o_X\{P^o_{W|X}(|A_n| \geq \varepsilon) \geq \nu\} &\leq \frac{1}{\nu} E^o_X P^o_{W|X}(|A_n| \geq \varepsilon) \\ &\leq \frac{1}{\nu} E^o_X E^o_{W|X} 1\{|A_n| \geq \varepsilon\} \end{aligned} \tag{74}$$

by Markov's inequality. According to Lemmas 6.5 and 6.14 in [22], we have $E^o_X E^o_{W|X} 1\{|A_n| \geq \varepsilon\} \leq E^o_{XW} 1\{|A_n| \geq \varepsilon\} = P^o_{XW}(|A_n| \geq \varepsilon)$, and thus

$$P^o_X\{P^o_{W|X}(|A_n| \geq \varepsilon) \geq \nu\} \leq \frac{1}{\nu} P^o_{XW}(|A_n| \geq \varepsilon). \tag{75}$$

From (75), we can conclude that if $A_n = o^o_{P_{XW}}(1)$, then $A_n = o^o_{P_W}(1)$ in P^o_X-probability. Another direction of (69) follows from the following inequalities: for any $\varepsilon, \eta > 0$,

$$\begin{aligned} P^o_{XW}(|A_n| \geq \varepsilon) &= E^o_X\{P^o_{W|X}(|A_n| \geq \varepsilon)\} \\ &= E^o_X\big\{P^o_{W|X}(|A_n| \geq \varepsilon) 1\{P^o_{W|X}(|A_n| \geq \varepsilon) \geq \eta\}\big\} \\ &\quad + E^o_X\big\{P^o_{W|X}(|A_n| \geq \varepsilon) 1\{P^o_{W|X}(|A_n| \geq \varepsilon) < \eta\}\big\} \\ &\leq E^o_X\big\{1\{P^o_{W|X}(|A_n| \geq \varepsilon) \geq \eta\}\big\} + \eta \\ &\leq P^o_X\{P^o_W(|A_n| \geq \varepsilon) \geq \eta\} + \eta. \end{aligned} \tag{76}$$

Note that the first term in (76) can be made arbitrarily small by the assumption that $A_n = o^o_{P_W}(1)$ in P^o_X-probability. Since η can be chosen arbitrarily small, we can show $\lim_{n\to\infty} P^o_{XW}(|A_n| \geq \varepsilon) = 0$ for any $\varepsilon > 0$. This completes the proof of (69). (70) can be shown similarly by using the inequalities (74) and (76).

As for (71), we establish the following inequalities:

$$\begin{aligned}
&P^o_X\{P^o_{W|X}(|Q_n \times O^o_{P_X}(1)| \geq \varepsilon) \geq \nu\} \\
&\quad \leq P^o_X\{P^o_{W|X}(|Q_n| \geq \varepsilon/|O^o_{P_X}(1)|) \geq \nu\} \\
&\quad \leq P^o_X\{P^o_{W|X}(|Q_n| \geq \varepsilon/M) + P^o_{W|X}(|O^o_{P_X}(1)| \geq M) \geq \nu\} \\
&\quad \leq P^o_X\{P^o_{W|X}(|Q_n| \geq \varepsilon/M) \geq \nu/2\} + P^o_X\{P^o_{W|X}(|O^o_{P_X}(1)| \geq M) \geq \nu/2\} \\
&\quad \leq P^o_X\{P^o_{W|X}(|Q_n| \geq \varepsilon/M) \geq \nu/2\} + \frac{2}{\nu} P^o_X(|O^o_{P_X}(1)| \geq M)
\end{aligned}$$

for any $\varepsilon, \nu, M > 0$. Since M can be chosen arbitrarily large, we can show (71) by considering the definition of $O^o_{P_X}(1)$. The proof of (72) is similar by using the above set of inequalities. The proof of (71) can be carried over to prove (73). Similarly, we establish the following inequalities:

$$\begin{aligned}
&P^o_X\{P^o_{W|X}(|Q_n \times R_n| \geq \varepsilon) \geq \eta\} \\
&\quad \leq P^o_X\{P^o_{W|X}(|Q_n| \geq \varepsilon/M) \geq \eta/2\} + P^o_X\{P^o_{W|X}(|R_n| \geq M) \geq \eta/2\}
\end{aligned}$$

for any $\varepsilon, \eta, M > 0$. Then by selecting sufficiently large M, we can show that

$$P^o_X\{P^o_{W|X}(|Q_n \times R_n| \geq \varepsilon) \geq \eta\} \to 0$$

as $n \to \infty$ for any $\varepsilon, \eta > 0$. □

A.2. Two useful inequalities. Here we give two key inequalities used in proving Lemmas 1 and 2.

Multiplier inequality (*Lemma* 4.1 *of* [41]).

Let $W_n = (W_{n1}, \ldots, W_{nn})'$ be nonnegative exchangeable random variables on $(\mathcal{W}, \Omega, P_W)$ such that, for every n, $R_n = \int_0^\infty \sqrt{P_W(W_{n1} \geq u)}\, du < \infty$. Let Z_{ni}, $i = 1, 2, \ldots, n$, be i.i.d. random elements in $(\mathcal{X}^\infty, \mathcal{A}^\infty, P^\infty_X)$ with values in $\ell^\infty(\mathcal{F}_n)$, and write $\|\cdot\|_n = \sup_{f\in\mathcal{F}_n} |Z_{ni}(f)|$. It is assumed that Z_{ni}'s are independent of W_n. Then for any n_0 such that $1 \leq n_0 < \infty$ and any $n > n_0$, the following inequality holds:

$$\begin{aligned}
E^o_{XW}\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n W_{ni} Z_{ni} \right\|_n &\leq n_0 E^o_X \|Z_{n1}\|_n \cdot \frac{E_W(\max_{1\leq i\leq n} W_{ni})}{\sqrt{n}} \\
&\quad + R_n \cdot \max_{n_0 < i \leq n} \left\{ E^o_X \frac{1}{\sqrt{i}} \left\| \sum_{j=n_0+1}^{i} Z_{nj} \right\|_n \right\}.
\end{aligned} \tag{77}$$

Hoffmann–Jorgensen inequality for moments (*Proposition* A.1.5 *in* [38]).

Let $1 \le p < \infty$ and suppose that $V_1, \ldots, V_n$ are independent stochastic processes with mean zero indexed by an arbitrary index set T. Then there exist constants K_p and $0 < v_p < 1$ such that

$$E^o \left\| \sum_{i=1}^n V_i \right\|^p \le K_p \Big\{ E^o \max_{1 \le k \le n} \|V_k\|^p + [G^{-1}(v_p)]^p \Big\},$$

where $\|Y\| = \sup_t |Y_t|$ denotes the supremum of a stochastic process $\{Y_t, t \in T\}$, and $G^{-1}(v) = \inf\{u : P^o(\|\sum_{i=1}^n V_i\| \le v) \ge u\}$.

A.3. Proof of Lemma 2. We first write $\mathbb{G}_n^*(\widetilde{m}(\theta, \eta) - \widetilde{m}_0)$ as the sum of $\mathbb{G}_n^*(\widetilde{m}(\theta, \eta) - \widetilde{m}(\theta_0, \eta))$ and $\mathbb{G}_n^*(\widetilde{m}(\theta_0, \eta) - \widetilde{m}_0)$. By the Taylor expansion, the first term becomes $(\theta - \theta_0)' \mathbb{G}_n^*(\partial/\partial\theta)\widetilde{m}(\bar{\theta}, \eta)$, where $\bar{\theta}$ is between θ and θ_0. By SB2 and Theorem 2.2 in [33], we know that the first term is of the order $O_{P_W}^o(\|\theta - \theta_0\|)$ in P_X^o-probability. We next consider the second term. Let

$$\Delta_n = \sup_{\eta \in U_n} \left\{ \frac{\|\mathbb{G}_n^*(\widetilde{m}(\theta_0, \eta) - \widetilde{m}_0)\|}{\|\eta - \eta_0\|} \right\}, \tag{78}$$

where $U_n = \{\eta : \|\eta - \eta_0\| \le \delta_n\}$ for any $\delta_n \to 0$. Note that we can write $\Delta_n = \|\mathbb{G}_n^*\|_{\mathcal{S}_n}$, where $\|\mathbb{G}_n^*\|_{\mathcal{S}_n} = \sup_{f \in \mathcal{S}_n} |\mathbb{G}_n^* f|$. By (70), to verify the bootstrap equicontinuity condition that $\mathbb{G}_n^*(\widetilde{m}(\theta_0, \eta) - \widetilde{m}_0) = O_{P_W}^o(\|\eta - \eta_0\|)$ in P_X^o-probability, it suffices to show

$$\lim_{n \to \infty} \sup E_{XW}^o \Delta_n < \infty. \tag{79}$$

Note that

$$\mathbb{G}_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n (W_{ni} - 1)\delta_{X_i} = \frac{1}{\sqrt{n}} \sum_{i=1}^n (W_{ni} - 1)(\delta_{X_i} - P_X)$$

by condition W2. Let $W_n' = (W_{n1}', \ldots, W_{nn}')$ be exchangeable bootstrap weights generated from $P_{W'}$, an independent copy of P_W. The bootstrap weight conditions W1 and W2 imply that $E_{W'} W_{ni}' = 1$ for $i = 1, \ldots, n$. Let

$$m_n(\eta, \eta_0) = \frac{\widetilde{m}(\theta_0, \eta) - \widetilde{m}_0}{\|\eta - \eta_0\|}.$$

Then we have

$$\begin{aligned} E_{XW}^o \Delta_n &= E_{XW}^o \sup_{\eta \in U_n} \|\mathbb{G}_n^* m_n(\eta, \eta_0)\| \\ &= E_{XW}^o \sup_{\eta \in U_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (W_{ni} - 1)(\delta_{X_i} - P_X) m_n(\eta, \eta_0) \right\| \end{aligned}$$

$$= E^o_{XW} \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (W_{ni} - E_{W'} W'_{ni})(\delta_{X_i} - P_X) m_n(\eta, \eta_0) \right\|$$

$$\le E^o_{XW} E^o_{W'} \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (W_{ni} - W'_{ni})(\delta_{X_i} - P_X) m_n(\eta, \eta_0) \right\|.$$

To further bound $E^o_{XW}\Delta_n$, we employ the symmetrization argument familiar in the empirical process literature to obtain

$$\begin{aligned} E^o_{XW}\Delta_n &\le E^o_{XW} \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n W_{ni}(\delta_{X_i} - P_X) m_n(\eta, \eta_0) \right\| \\ &\quad + E^o_{XW} E^o_{W'} \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n W'_{ni}(\delta_{X_i} - P_X) m_n(\eta, \eta_0) \right\| \\ &= 2E^o_{XW} \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n W_{ni}(\delta_{X_i} - P_X) m_n(\eta, \eta_0) \right\|. \end{aligned} \tag{80}$$

We next apply the multiplier inequality (77) to (80) with $Z_{ni} = \{(\delta_{X_i} - P_X)m_n(\eta, \eta_0) : \eta \in U_n\}$. Define

$$\|Z_{ni}\|_n = \sup_{\eta\in U_n} \|(\delta_{X_i} - P_X)m_n(\eta, \eta_0)\|.$$

To show (79), we need only to show

$$E_W\Big(\max_{1\le i\le n} W_{ni}\Big)\Big/\sqrt{n} \to 0, \tag{81}$$

$\limsup_n E^o_X \|Z_{n1}\|_n < \infty$, and

$$\limsup_n \max_{n_0<i\le n} E^o_X \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{i}} \sum_{j=n_0+1}^i Z_{ni} \right\| < \infty \tag{82}$$

for some $n_0 < \infty$. The bootstrap weight conditions W3 and W4 together with Lemma 4.7 in [33] imply (81). Note that

$$\begin{aligned} E^o_X \|Z_{n1}\|_n &= E^o_X \sup_{\eta\in U_n} \|(\delta_{X_1} - P_X)m_n(\eta, \eta_0)\| \\ &\le E^o_X \sup_{\eta\in U_n} \|m_n(\eta, \eta_0)(X_1)\| + E^o_X \sup_{\eta\in U_n} \|E_X m_n(\eta, \eta_0)\| \\ &\le 2E^o_X S_n(X_1), \end{aligned}$$

where S_n is the envelop of the class $\mathcal{S}_n$ defined in (10), and the first inequality follows from the Fatou's lemma. Condition SB1 implies

$$\frac{1}{\sqrt{n}} E^o_X \max_{1\le k\le n} S_n(X_k) \longrightarrow 0, \tag{83}$$

$$\limsup_{n\to\infty} E^o_X S_n(X_1) < \infty; \tag{84}$$

see page 120 of [38]. The result (84) implies $\limsup_n E_X^o \|Z_{n1}\|_n < \infty$.

It remains to show (82). We apply the Hoffmann–Jorgensen inequality with $p = 1$ in Appendix A.2. First, we establish

$$E_X^o \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{ni} \right\| \le K_1 \left\{ \frac{1}{\sqrt{n}} E_X^o \max_{1\le k\le n} \|Z_{nk}\|_n + G_n^{-1}(v_1) \right\} \le I_1 + I_2, \tag{85}$$

where K_1 and $0 < v_1 < 1$ are constants and

$$G_n(t) = P_X^o \left(n^{-1/2} \left\| \sum_{i=1}^n Z_{ni} \right\|_n \le t \right).$$

Obviously, (83) implies that $I_1 \to 0$. We next consider I_2. Note that assumption S1 implies $\|\mathbb{G}_n\|_{\mathcal{S}_n} = \|n^{-1/2} \sum_{i=1}^n Z_{ni}\|_n = O_{P_X}^o(1)$. Hence, there exists a finite constant M_t such that $\liminf_n G_n(M_t) \ge t$ for every $1 > t > 0$. It follows that $\limsup_n G_n^{-1}(v_1) \le M_{v_1} < \infty$ since $0 < v_1 < 1$. Thus, the left-hand side of (85) is bounded away from infinity, and therefore (82) holds in light of the following result from the triangular inequality

$$\max_{n_0<i\le n} E_X^o \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{i}} \sum_{j=n_0+1}^i Z_{nj} \right\| \le \max_{n_0<i\le n} E_X^o \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{i}} \sum_{j=1}^i Z_{nj} \right\| + E_X^o \sup_{\eta\in U_n} \left\| \frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} Z_{nj} \right\|.$$

The proof of Lemma 2 is complete.

A.4. Proof of Theorem 2. Using (40) and the fact that $U(\theta_0, \eta_0) = 0$, we have

$$\begin{aligned} &U(\widetilde{\theta}, \widehat{\eta}_{\widetilde{\theta}}^*) - U(\theta_0, \eta_0) \\ &\quad = U(\widetilde{\theta}, \widehat{\eta}_{\widetilde{\theta}}^*) - U_n^*(\widetilde{\theta}, \widehat{\eta}_{\widetilde{\theta}}^*) + O_{P_{XW}}^o(n^{-1/2}) \\ &\quad = -(U_n^* - U_n)(\widetilde{\theta}, \widehat{\eta}_{\widetilde{\theta}}^*) - (U_n - U)(\widetilde{\theta}, \widehat{\eta}_{\widetilde{\theta}}^*) + O_{P_{XW}}^o(n^{-1/2}) \\ &\quad = L_1 + L_2 + O_{P_{XW}}^o(n^{-1/2}). \end{aligned} \tag{86}$$

Further, based on conditions (37) and (39), we apply Lemma 4.2 in [41] to obtain that $L_1 = -(U_n^* - U_n)(\theta_0, \eta_0) + o_{P_{XW}}^o(n^{-1/2} \vee \|\widetilde{\theta} - \theta_0\| \vee \|\widehat{\eta}_{\widetilde{\theta}}^* - \eta_0\|)$. By Lemma 3.3.5 in [38] given (37) and (38), we have $L_2 = -(U_n - U)(\theta_0, \eta_0) + o_{P_{XW}}^o(n^{-1/2} \vee \|\widetilde{\theta} - \theta_0\| \vee \|\widehat{\eta}_{\widetilde{\theta}}^* - \eta_0\|)$. By applying CLT and Theorem 2.2 in [33] under condition (37) to L_1 and L_2, we have

$$U(\widetilde{\theta}, \widehat{\eta}_{\widetilde{\theta}}^*) - U(\theta_0, \eta_0) = O_{P_{XW}}^o(n^{-1/2}) + o_{P_{XW}}^o(\|\widetilde{\theta} - \theta_0\| \vee \|\widehat{\eta}_{\widetilde{\theta}}^* - \eta_0\|). \tag{87}$$

We next apply the Taylor expansion to get

$$\begin{aligned} &U(\widetilde{\theta}, \widehat{\eta}^*_{\widetilde{\theta}}) - U(\theta_0, \eta_0) \\ &\quad = \dot{U}(\widetilde{\theta} - \theta_0, \widehat{\eta}^*_{\widetilde{\theta}} - \eta_0) + o(\|\widetilde{\theta} - \theta_0\| \vee \|\widehat{\eta}^*_{\widetilde{\theta}} - \eta_0\|) \\ &\quad = \dot{U}(\widetilde{\theta} - \theta_0, 0) + \dot{U}(0, \widehat{\eta}^*_{\widetilde{\theta}} - \eta_0) + o(\|\widetilde{\theta} - \theta_0\| \vee \|\widehat{\eta}^*_{\widetilde{\theta}} - \eta_0\|) \end{aligned}$$

by the assumed Fréchet differentiability of U and linearity of $\dot{U}$. Note that U has bounded Fréchet derivative and $\dot{U}(0, \cdot)$ is continuously invertible. Thus, we can conclude that

$$U(\widetilde{\theta}, \widehat{\eta}^*_{\widetilde{\theta}}) - U(\theta_0, \eta_0) \geq c\|\widehat{\eta}^*_{\widetilde{\theta}} - \eta_0\| + O(\|\widetilde{\theta} - \theta_0\|) + o(\|\widetilde{\theta} - \theta_0\| \vee \|\widehat{\eta}^*_{\widetilde{\theta}} - \eta_0\|)$$

for some $c > 0$. Combining the above inequality with (87), we can establish the following inequality:

$$\|\widehat{\eta}^*_{\widetilde{\theta}} - \eta_0\| \lesssim O^o_{P_{XW}}(\|\widetilde{\theta} - \theta_0\| \vee n^{-1/2}) + o^o_{P_{XW}}(\|\widehat{\eta}^*_{\widetilde{\theta}} - \eta_0\|),$$

which implies (41).

A.5. Proof of Theorem 3. According to (70), we need only to show that

$$P^o_{XW}\big(d(\widehat{\eta}^*_{\widetilde{\theta}}, \eta_n) \geq 2^{M_n}(\delta_n \vee \|\widetilde{\theta} - \theta_0\|), \widetilde{\theta} \in \Theta, \widehat{\eta}^*_{\widetilde{\theta}} \in \mathcal{H}_n\big) \longrightarrow 0 \tag{88}$$

as $n \to \infty$ and $M_n \to \infty$. The basic idea in proving (88) is first to partition the whole parameter space into "shells," and then bound the probability of each shell under conditions (43)–(45).

For now we fix $M = M_n$ and then allow it to increase to infinity. We first define the shell $S_{n,j,M}$ as

$$S_{n,j,M} = \{(\theta, \eta) \in \Theta \times \mathcal{H}_n : 2^{j-1}\delta_n < d(\eta, \eta_n) \leq 2^j \delta_n, d(\eta, \eta_n) \geq 2^M \|\theta - \theta_0\|\}$$

with j ranging over the integers and $M > 0$. Obviously, the event $\{\widetilde{\theta} \in \Theta, \widehat{\eta}^*_{\widetilde{\theta}} \in \mathcal{H}_n : d(\widehat{\eta}^*_{\widetilde{\theta}}, \eta_n) \geq 2^M(\delta_n \vee \|\widetilde{\theta} - \theta_0\|)\}$ is contained in the union of the events $\{(\widetilde{\theta}, \widehat{\eta}^*_{\widetilde{\theta}}) \in S_{n,j,M}\}$ for $j \geq M$. Thus, we have

$$\begin{aligned} &P^o_{XW}\big(d(\widehat{\eta}^*_{\widetilde{\theta}}, \eta_n) \geq 2^M(\delta_n \vee \|\widetilde{\theta} - \theta_0\|), \widetilde{\theta} \in \Theta, \widehat{\eta}^*_{\widetilde{\theta}} \in \mathcal{H}_n\big) \\ &\quad \leq \sum_{j \geq M} P^o_{XW}\big((\widetilde{\theta}, \widehat{\eta}^*_{\widetilde{\theta}}) \in S_{n,j,M}\big) \\ &\quad \leq \sum_{j \geq M} P^o_{XW}\Big(\sup_{(\theta,\eta) \in S_{n,j,M}} \mathbb{P}^*_n\big(v(\theta, \eta) - v(\theta, \eta_n)\big) \geq 0\Big). \end{aligned}$$

The second inequality follows from the definition of $\widehat{\eta}^*_{\widetilde{\theta}}$. By the smoothness condition on $v(\theta, \eta)$, that is, (43), we have the following inequality when $(\theta, \eta) \in S_{j,n,M}$ for $j \geq M$:

$$P_X\big(v(\theta, \eta) - v(\theta, \eta_n)\big) \lesssim -d(\eta, \eta_n)^2 + \|\theta - \theta_0\|^2 \lesssim -2^{2j-2}\delta_n^2 \tag{89}$$

for sufficiently large M.

Considering (89), we have

$$
\begin{aligned}
&P^o_{XW}\big(d(\widehat{\eta}^*_{\widetilde{\theta}}, \eta_n) \geq 2^M(\delta_n \vee \|\widetilde{\theta} - \theta_0\|), \widetilde{\theta} \in \Theta, \widehat{\eta}^*_{\widetilde{\theta}} \in \mathcal{H}_n\big) \\
&\quad \leq \sum_{j \geq M} P^o_{XW}\Big(\sup_{(\theta,\eta) \in S_{n,j,M}} \sqrt{n}(\mathbb{P}^*_n - P_X)\big(v(\theta,\eta) - v(\theta,\eta_n)\big) \gtrsim \sqrt{n} 2^{2j-2} \delta_n^2 \Big) \\
&\quad \leq \sum_{j \geq M} P^o_{XW}\Big(\sup_{(\theta,\eta) \in S_{n,j,M}} \big|\mathbb{G}^*_n\big(v(\theta,\eta) - v(\theta,\eta_n)\big)\big| \gtrsim \sqrt{n} 2^{2j-3} \delta_n^2 \Big) \\
&\qquad + P^o_X\Big(\sup_{(\theta,\eta) \in S_{n,j,M}} \big|\mathbb{G}_n\big(v(\theta,\eta) - v(\theta,\eta_n)\big)\big| \gtrsim \sqrt{n} 2^{2j-3} \delta_n^2 \Big) \\
&\quad \lesssim \sum_{j \geq M} \frac{\psi^*_n(2^j \delta_n)}{\sqrt{n}\delta_n^2 2^{2j}} + \frac{\psi_n(2^j \delta_n)}{\sqrt{n}\delta_n^2 2^{2j}} \\
&\quad \lesssim \sum_{j \geq M} 2^{j(\alpha-2)},
\end{aligned}
$$

where the third inequality follows from the Markov inequality and (44) and (45). Note that the assumption that $\delta \mapsto \psi_n(\delta)/\delta^\alpha$ [$\delta \mapsto \psi^*_n(\delta)/\delta^\alpha$] is decreasing for some $0 < \alpha < 2$ implies that $\psi_n(c\delta) \leq c^\alpha \psi_n(\delta)$ for every $c > 1$. Combining these with the assumption that $\psi_n(\delta_n) \leq \sqrt{n}\delta_n^2$ and $\psi^*_n(\delta_n) \leq \sqrt{n}\delta_n^2$, we obtain the last inequality in the above display. By letting $M = M_n \to \infty$, we complete the proof of (88), and thus Theorem 3.

A.6. Proof of Lemma 1. The result (51) is an immediate consequence of Lemma 3.4.2 in [38]. To show (53), we first apply the symmetrization arguments used in the proof of Lemma 2. For sufficiently small δ, the left-hand side of (45) is bounded by

$$
2E^o_{XW} \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^n W_{ni} Y_{ni} \Big\|_{\mathcal{V}_\delta}, \tag{90}
$$

where W_{ni}'s are the assumed bootstrap weights and

$$
Y_{ni} = \big\{(\delta_{X_i} - P_X)\big(v(\theta,\eta) - v(\theta,\eta_n)\big) : d(\eta,\eta_n) \leq \delta, \|\theta - \theta_0\| \leq \delta\big\}.
$$

Next, the multiplier inequality (77) is employed to further bound (90). In view of (77), we need only to figure out the upper bound for

$$
E^o_X \|Y_{n1}\|_{\mathcal{V}_\delta} \tag{91}
$$

and

$$
\max_{n_0 \leq i \leq n} E^o_X \Big\| \frac{1}{\sqrt{i}} \sum_{j_0+1}^{i} Y_{nj} \Big\|_{\mathcal{V}_\delta} \tag{92}
$$

for some $n_0 \geq 1$ given assumptions W3 and W4 on the bootstrap weights. By a similar argument as in the proof of Lemma 2, we know

$$E_X^o \|Y_{n1}\|_{\mathcal{V}_\delta} \leq 2E_X^o V_n(X_1),$$

where V_n is the envelop function of the class $\mathcal{V}_\delta$ defined in (48). The assumption (52), together with the analysis of assumption SB1, implies that $\limsup_n E_X^o \|Y_{n1}\|_{\mathcal{V}_\delta} < \infty$. Next, Lemma 3.4.2 in [38] implies that

$$E_X^o \|\mathbb{G}_n\|_{\mathcal{V}_\delta} \leq K(\delta, \mathcal{V}_\delta, L_2(P))\left(1 + \frac{K(\delta, \mathcal{V}_\delta, L_2(P))}{\delta^2\sqrt{n}}\right).$$

By the triangular inequality, we know that (92) has the same upper bound as $E_X^o \|\mathbb{G}_n\|_{\mathcal{V}_\delta}$. This concludes the proof of (53).

Acknowledgments. The authors thank Professor Anirban DasGupta for continuous encouragement and Professors Michael Kosorok and Jon Wellner for many helpful discussions. The authors also thank the Co-editor Susan Murphy and two referees for insightful comments which led to important improvements over an earlier draft.

REFERENCES

[1] BARBE, P. and BERTAIL, P. (1995). *The Weighted Bootstrap*. *Lecture Notes in Statistics* **98**. Springer, New York. MR2195545

[2] BANERJEE, M., MUKHERJEE, D. and MISHRA, S. (2009). Semiparametric binary regression models under shape constraints with an application to Indian schooling data. *J. Econometrics* **149** 101–117. MR2518501

[3] BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103

[4] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. MR1623559

[5] CHATTERJEE, S. and BOSE, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.* **33** 414–436. MR2157808

[6] CHEN, X. and POUZO, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *J. Econometrics* **152** 46–60. MR2562763

[7] CHENG, G. (2008). Semiparametric additive isotonic regression. *J. Statist. Plann. Inference* **100** 345–362. MR2497554

[8] CHENG, G. and KOSOROK, M. (2008). Higher order semiparametric frequentist inference with the profile sampler. *Ann. Statist.* **36** 1786–1818. MR2435456

[9] CHENG, G. and KOSOROK, M. (2008). General frequentist properties of the posterior profile distribution. *Ann. Statist.* **36** 1819–1853. MR2435457

[10] DELECROIX, M., HRISTACHE, M. and PATILEA, V. (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference* **136** 730–769. MR2181975

[11] DIXON, J., KOSOROK, M. and LEE, B. L. (2005). Functional inference in semiparametric models using the piggyback bootstrap. *Ann. Inst. Statist. Math.* **57** 255–277. MR2160650

[12] DUDLEY, R. M. (1984). *A Course on Empirical Processes*. *Lecture Notes in Math.* **1097** 2–142. Springer, Berlin. MR0876079

[13] EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia. MR0659849
[14] EFRON, B. and TIBSHIRANI, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* **1** 54–75. MR0833275
[15] GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York. MR0599175
[16] GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D. (2003). *Bayesian Data Analysis*, 2nd ed. Chapman and Hall, London. MR1385925
[17] GINE, E. and ZINN, J. (1990). Bootstrapping general empirical functions. *Ann. Probab.* **18** 851–869. MR1055437
[18] HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York. MR1145237
[19] HARDLE, W., HUET, S., MAMMEN, E. and SPERLICH, S. (2004). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory* **20** 265–300. MR2044272
[20] HUANG, J. (1999). Efficient estimation of the partly linear Cox model. *Ann. Statist.* **27** 1536–1563. MR1742499
[21] KOSOROK, M., LEE, B. L. and FINE, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.* **32** 1448–1491. MR2089130
[22] KOSOROK, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.
[23] KOSOROK, M. (2008). Boostrapping the Grenander estimator. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*. *IMS Collections* **1** 282–292. IMS, Beachwood, OH. MR2462212
[24] LEE, B. L., KOSOROK, M. R. and FINE, J. P. (2005). The profile sampler. *J. Amer. Statist. Assoc.* **100** 960–969. MR2201022
[25] LEE, S. M. S. and PUN, M. C. (2006). On m out of n bootstrapping for nonstandard M-estimation with nuisance parameters. *J. Amer. Statist. Assoc.* **101** 1185–1197. MR2328306
[26] LIANG, H., HÄRDLE, W. and SOMMERFELD, V. (2000). Bootstrap approximations in a partially linear regression model. *J. Statist. Plann. Inference* **91** 413–426. MR1814793
[27] LO, A. Y. (1993). A Bayesian bootstrap for censored data. *Ann. Statist.* **21** 100–123. MR1212168
[28] MA, S. and KOSOROK, M. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.* **96** 190–217. MR2202406
[29] MA, S. and KOSOROK, M. (2005). Penalized log-likelihood estimation for partly linear transformation models with current status data. *Ann. Statist.* **33** 2256–2290. MR2211086
[30] MASON, D. and NEWTON, M. (1992). A rank statistic approach to the consistency of a general bootstrap. *Ann. Statist.* **20** 1611–1624. MR1186268
[31] MURPHY, S. A. and VAN DER VAART, A. W. (1999). Observed information in semiparametric models. *Bernoulli* **5** 381–412. MR1693616
[32] MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 1461–1474. MR1803168
[33] PRAESTGAARD, J. and WELLNER, J. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. MR1245301
[34] RUBIN, D. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134. MR0600538
[35] SEN, B., BANERJEE, M. and WOODROOFE, M. B. (2010). Inconsistency of bootstrap: The Grenander estimator. *Ann. Statist.* **38** 1953–1977.
[36] SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195. MR0630102
[37] STRAWDERMAN, R. (2006). A regression model for dependent gap times. *Int. J. Biostat.* **2** Article 1, 34 pp. (electronic). MR2275896

[38] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

[39] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge. MR1652247

[40] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

[41] WELLNER, J. A. and ZHAN, Y. (1996). Bootstrapping Z-estimators. Technical Report 308, Univ. Washington.

[42] WELLNER, J. A. and ZHANG, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35** 2106–2142. MR2363965

[43] YOUNG, J. G., JEWELL, N. P. and SAMUELS, S. J. (2008). Regression analysis of a disease onset distribution using diagnosis data. *Biometrics* **64** 20–28. MR2422815

[44] ZENG, D. L. and LIN, D. Y. (2007). Maximum likelihood estimation in semiparametric models with censored data (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 507–564. MR2370068

[45] ZHANG, C. M. and YU, T. (2008). Semiparametric detection of significant activation for brain FMRI. *Ann. Statist.* **36** 1693–1725. MR2435453

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907-2066
USA
E-MAIL: chengg@purdue.edu

DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143
USA
E-MAIL: jianhua@stat.tamu.edu