Scandinavian Journal of Statistics Theory and Applications

doi: 10.1002/sjos.12005

How Many Iterations are Sufficient for Efficient Semiparametric Estimation?

GUANG CHENG

Department of Statistics, Purdue University

ABSTRACT. A common practice in obtaining an efficient semiparametric estimate is through iteratively maximizing the (penalized) full log-likelihood w.r.t. its Euclidean parameter and functional nuisance parameter. A rigorous theoretical study of this semiparametric iterative estimation approach is the main purpose of this study. We first show that the grid search algorithm produces an initial estimate with the proper convergence rate. Our second contribution is to provide a formula in calculating the minimal number of iterations k^* needed to produce an efficient estimate $\hat{\theta}_n^{(k^*)}$. We discover that (i) k^* depends on the convergence rates of the initial estimate and the nuisance functional estimate, and (ii) k^* iterations are also sufficient for recovering the estimation sparsity in high dimensional data. The last contribution is the novel construction of $\hat{\theta}_n^{(k)}$ which does not require knowing the explicit expression of the efficient score function. The above general conclusions apply to semiparametric models estimated under various regularizations, for example, kernel or penalized estimation. As far as we are aware, this study provides a first general theoretical justification for the 'one-/two-step iteration' phenomena observed in the semiparametric literature.

Key words: generalized profile likelihood, higher order asymptotic efficiency, k-step estimation, Newton–Raphson algorithm, semiparametric models

1. Introduction

Semiparametric models indexed by a Euclidean parameter of interest $\theta \in \Theta \subset \mathbb{R}^d$ and an infinite-dimensional nuisance parameter $\eta \in \mathcal{H}$ are proven to be useful in a variety of contexts (e.g. Severini & Staniswalis, 1994; Fan *et al.*, 1995; Huang, 1996; Roeder *et al.*, 1996; Carroll *et al.*, 1997). The semiparametric MLE for θ can be viewed as a solution of the implicitly defined efficient score function whose non-parametric estimation is only possible in some special cases (e.g. Huang, 1996). Therefore, it is generally hard to solve the MLE from the efficient score function analytically or numerically. A common practice is to maximize the log-profile likelihood

$$\log pl_n(\theta) = \sup_{\eta \in \mathcal{H}} \log lik_n(\theta, \eta), \tag{1}$$

where $lik_n(\theta, \eta) = \prod_{i=1}^n lik(X_i; \theta, \eta)$ is the likelihood based on i.i.d. data $(X_1, \ldots, X_n)$, via some optimization algorithm. For example, the Newton–Raphson algorithm is applied to the partial likelihood of the Cox model in the software **R** (with the command *coxph*).

A general approach of obtaining an efficient semiparametric estimate of θ is as follows:

General semiparametric iterative estimation approach

i. Identify an initial estimate $\hat{\theta}_n^{(0)}$.
ii. Apply the Newton–Raphson (NR) or other optimization algorithm to the generalized profile likelihood ([32]):

$$\hat{S}_n(\theta) = \log lik_n(\theta, \hat{\eta}(\theta)), \tag{2}$$

at $\theta = \hat{\theta}_n^{(0)}$to obtain $\hat{\theta}_n^{(1)}$. The above $\hat{\eta}(\theta)$ is defined as the nuisance estimate for any fixed $\theta \in \Theta$. We can construct it either by pure non-parametric approach, for example, isotonic estimation, or under some regularization, for example, kernel or sieve estimation.

iii. Update k^* times the value of θ in the optimization algorithm of Step II until

$$\|\hat{S}_n\left(\hat{\theta}_n^{(k^*)}\right) - \hat{S}_n\left(\hat{\theta}_n^{(k^*-1)}\right)\| \leq \epsilon$$

for some norm $\|\cdot\|$ and pre-determined small $\epsilon > 0$.

Note that, in step II, we first obtain $\hat{\eta}(\theta)$ by maximizing some log-likelihood-based criterion function, for example (31), given the current value θ, and then update the value of θ according to the criterion function $\hat{S}_n(\theta)$ built upon $\hat{\eta}(\theta)$. Hence, the whole algorithm iteratively maximizes the log-likelihood w.r.t. θ and η in an implicit manner. If $\hat{\eta}(\theta)$ is the non-parametric MLE (NPMLE), then $\hat{S}_n(\theta)$ is just the profile likelihood defined in (1). The above likelihood estimation procedure or its M-estimation variant has been extensively implemented in the literature. Here is an incomplete list: (i) Odds-Rate Regression Model under Survival Data (e.g. Huang, 1996);I (ii) Semiparametric Regression under Shape Constraints (e.g. Cheng, 2009); (iii) Logistic Regression with Missing Covariates (e.g. Roeder *et al.*, l996); (iv) Generalized Partly Linear (Single Index) Model (e.g. Fan *et al.*, 1995; Carroll *et al.*, 1997); (v) Conditionally Parametric Model (Severini and Wong, 1992; Severini and Stainswalis, 1994). The above iterative procedure is also widely applied to the sparse semiparametric estimation, that is, some components of the true value θ_0 are zero, by using a penalized version of (2), see section 5.2.

Unfortunately, the rigorous statistical analyses are unavailable for the above semiparametric estimation approach. This is mainly because the higher order asymptotic analysis on the implicitly defined $\hat{S}_n(\theta)$ is very hard to derive. In this study, by using the empirical processes theories, we have made the following three contributions.

It is well known that identifying the proper initial estimate $\hat{\theta}_n^{(0)}$ is critical in guaranteeing the fast convergence of the above approach. Occasionally, we can exploit some simple semiparametric models to produce a $\sqrt{n}$-consistent $\hat{\theta}_n^{(0)}$, for example, partly linear model (Yatchew, 1997). However, a general strategy is to conduct a search of $\hat{S}_n(\theta)$ at finitely many θ-value and use the maximizer as $\hat{\theta}_n^{(0)}$ (e.g. Swann, 1972). Our first contribution is to provide very weak sufficient conditions for the above grid search to produce $\hat{\theta}_n^{(0)}$ with the desired convergence rate. When the dimension of θ is large, $\hat{\theta}_n^{(0)}$ may have the slower than root-n rate. This motivates our second contribution below, which is particularly useful for the sub-optimal $\hat{\theta}_n^{(0)}$.

Our second contribution is to answer the title of this article from a theoretical point of view. We provide a formula in calculating the minimal number of iterations k^* needed to produce a semiparametric efficient $\hat{\theta}_n^{(k^*)}$. Specifically, we discover that (i) k^* depends on the convergence rates of $\hat{\theta}_n^{(0)}$ and $\hat{\eta}(\theta)$; (ii) more than k^* iterations will not change the limiting distribution but improve the higher order asymptotic efficiency of the iterative estimate; (iii) k^* iterations are also sufficient for recovering the estimation sparsity, that is, estimating the zero components as exactly zero with large probability, under high dimensional data. Surprisingly, the value of k^* could be quite large when η is estimated at a very slow rate, for example, $k^* = 8$ in conditionally exponential models; (see Table 3).

The last contribution is the novel construction of $\hat{\theta}_n^{(k)}$. In contrast with the literature (i.e. Bickel, 1982; Schick, 1986) our construction does not require knowing the explicit expression/characterization of the efficient score function or applying the sample splitting (drop-one-out) technique.

All the above conclusions apply to a wide range of semiparametric models estimated under various regularizations, for example, kernel or penalized estimation. A special case of our general theory is the simple case of (parametric) GEE setting when the iteration is needed between the mean parameter estimates and variance/covariance parameter estimates (see Jiang *et al.*, 2007). It is known that one iteration is sufficient to achieve the asymptotical efficiency. The bootstrap results in this setup have also been studied by Lipsitz *et al.* (1990) and Cheng *et al.* (2012). Note that our theory does not directly cover the semiparametric models with bundled parameters (a terminology used by Huang & Wellner, 1997) in which the parameter of interest and the nuisance parameter are bundled together, that is, the function η also depends on θ. An example is the semiparametric binary regression model studied by (Cosslett, 1983; Dominitz & Sherman, 2005).

On the other hand, one has to be careful in applying the theoretical results of this study, because we focus more on the general theoretical explorations. In specific models, it might be possible to modify the grid search of $\hat{\theta}_n^{(0)}$ or the construction of $\hat{\theta}_n^{(k)}$ to better capture the model features so that the finite sample behaviours become better. Due to the space limitation, we only consider the NR algorithm in this study, but notice that the extensions to the modified NR are possible by considering the discussions in Robinson (1988, p. 534).

Section 2 provides some necessary review on the semiparametric efficient estimation. In section 3, we propose two grid search algorithms for identifying the initial estimate whose convergence rate will be rigorously proven. In section 4, we consider the semiparametric maximum likelihood estimation in which $\hat{S}_n(\theta)$ is the possibly non-differentiable profile likelihood (1). In section 5, we consider the semiparametric estimation under two types of regularization, that is, kernel estimation and penalized estimation. Some simulation experiment is also performed to empirically confirm our theory. As an example, we also consider the sparse and efficient estimation of the partial linear models. Several semiparametric models ranging from survival models, mixture models to conditionally exponential models are treated to illustrate the applicability of our theories. All the proofs are postponed to the Appendix or Supporting information. The latter is available on the journal's web site.

2. Preliminary

We assume that the data $X_1, \dots, X_n$ are i.i.d. throughout the article. In what follows, we first briefly review the concepts of the efficient score function and least favourable submodel (LFS), and then relate the estimation of LFS to that of θ as discussed in Severini & Wong (1992). Unless otherwise specified, the notation E is reserved for the expectation taken under the true value (θ_0, η_0).

The score functions for θ and η are defined as, respectively,

$$\begin{aligned} \dot{\ell}_0(X_i) &= \frac{\partial}{\partial\theta}|_{\theta=\theta_0} \log\ lik(X_i; \theta, \eta_0), \\ A_{\theta_0,\eta_0} h(X_i) &= \frac{\partial}{\partial t}|_{t=\theta_0} \log\ lik(X_i; \theta_0, \eta(t)), \end{aligned} \tag{3}$$

where h is a 'direction' along which $\eta(t) \in \mathcal{H}$ approaches η_0 as $t \to \theta_0$. $A_{\theta_0,\eta_0} : \mathbf{H} \mapsto L_2^0(P_{\theta_0,\eta_0})$ is the score operator for η, where $\mathbf{H}$ is some closed and linear direction set. The efficient score function $\tilde{\ell}_0$ is defined as the orthocomplement projection of $\dot{\ell}_0$ onto the tangent space $\mathcal{T}$, that is, the closed linear span of tangent set $\{A_{\theta_0,\eta_0} H = (A_{\theta_0,\eta_0} h_1, \dots, A_{\theta_0,\eta_0} h_d)' : h_j \in \mathbf{H}\}$. Therefore, we can write the efficient score function at (θ_0, η_0) as $\tilde{\ell}_0 = \dot{\ell}_0 - \Pi_0 \dot{\ell}_0$, where $\Pi_0 \dot{\ell}_0 = \arg\min_{t \in \mathcal{T}} E\|\dot{\ell}_0 - t\|^2$. The variance of $\tilde{\ell}_0$ is defined as the efficient information matrix $\tilde{I}_0$. The inverse of $\tilde{I}_0$ is shown to be Cramér–Rao bound for estimating θ in the presence of an infinite dimensional η (see Bickel *et al.*, 1998).

A main idea of estimating θ is to reduce a high-dimensional semiparametric model to a low-dimensional random submodel of the same dimension as θ called the LFS. The LFS is constructed as $t \mapsto \log\, lik(x; t, \eta_*(t))$ which satisfies

$$\eta_*(\theta_0) = \eta_0 \quad \text{and} \quad \frac{\partial}{\partial t}|_{t=\theta_0} \log\, lik(x; t, \eta_*(t)) = \tilde{\ell}_0(x). \tag{4}$$

The least favourable curve $\eta_*(t)$ turns out to be

$$\eta_*(t) = \arg\sup_{\eta \in \mathcal{H}} E \,\log\, lik(X; t, \eta) \quad \text{for any fixed } t \in \Theta. \tag{5}$$

The existence of LFS is implied by the closedness of the tangent set. By (5) and standard arguments, we can establish that the maximizer of $S_n(\theta) \equiv \sum_{i=1}^n \log\, lik(X_i; \theta, \eta_*(\theta))$ is semiparametric efficient. In addition, based on (4), we can derive that

$$\tilde{I}_0 = E\left(\frac{\partial \log\, lik(X; t, \eta_*(t))}{\partial t}|_{t=\theta_0}\right)^{\otimes 2} = -E\left(\frac{\partial^2 \log\, lik(X; t, \eta_*(t))}{\partial t_i \partial t_j}|_{t=\theta_0}\right)_{i,j=1,2,\ldots,d}. \tag{6}$$

Define

$$\hat{\theta}_n = \arg\sup_{\theta \in \Theta} \log\, lik_n(\theta, \hat{\eta}(\theta)) = \arg\sup_{\theta \in \Theta} \hat{S}_n(\theta). \tag{7}$$

In view of the above discussions, we can show that $\hat{\theta}_n$ is semiparametric efficient if $\hat{\eta}(\theta)$ is a consistent estimate of $\eta_*(\theta)$; see section 4 of Severini and Wong (1992) for more details. Note that the form of $\hat{\theta}_n$ depends on how we estimate the abstract $\eta_*(\theta)$ defined in (5). For example, $\hat{\theta}_n$ is just the semiparametric MLE if $\hat{\eta}(\theta)$ is the NPMLE. When the infinite dimensional $\mathcal{H}$ is too large, we may consider estimating $\eta_*(\theta)$ under some form of regularization, for example, penalization. In this study, we will consider two types of $\hat{\theta}_n$ defined in (7) according to the way we estimate $\eta_*(\theta)$: (i) pure non-parametric estimation in section 4; (ii) non-parametric estimation under regularization in section 5.

We use $\mathcal{N}(\theta_0)$ to denote some neighbourhood of θ_0. Let v_i denote the ith unit vector in $\mathbb{R}^d$. Define the ith ((*i,j*)th) element of a vector V (Matrix M) as V_i (M_{ij}). For a tensor $T^{(3)}(\theta)$ (Hoffmann, 1966), we define $V^{\mathrm{T}} \otimes T^{(3)}(\theta) \otimes V$ as a d-dimensional vector with ith element $V^{\mathrm{T}}(\partial^2/\partial\theta^2)(\dot{T}(\theta))_i V$, where $\dot{T}(\theta)$ is the gradient of $T(\theta)$. Denote $\mathrm{int}[x]$ and $\widetilde{\mathrm{int}}[x]$ as the smallest non-negative integer $\geq x$ and $> x$, respectively. The symbols $\mathbb{P}_n$ and $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$ are used for the empirical distribution and the empirical process of the observations, respectively.

3. Initial estimate

In this study, we assume the initial estimate $\hat{\theta}_n^{(0)}$ to be n^{ψ}-consistent for some $0 < \psi \leq 1/2$. We will prove that the grid search of $\hat{S}_n(\theta)$ will produce such an initial estimate in this section. The proof is non-trivial since $\hat{S}_n(\theta)$ usually has no explicit form and is possibly non-differentiable. In fact, our theoretical results on searching $\hat{\theta}_n^{(0)}$, that is, theorem 1, can be applied to any objective functions satisfying the below conditions I1–I2, and are thus of independent interest.

We first state two primary conditions I1–I2 on $\hat{S}_n(\theta)$.

I1. [Asymptotic uniqueness] For any random sequence $\{\tilde{\theta}_{xn}\} \in \Theta$,

$$[\hat{S}_n(\tilde{\theta}_n) - \hat{S}_n(\hat{\theta}_n)]/n = o_P(1) \text{ implies that } \tilde{\theta}_n - \theta_0 = o_P(1). \tag{8}$$

I2. [Asymptotic expansion] For any consistent $\tilde{\theta}_n$ and sufficiently large n, $\hat{S}_n$ satisfies

$$\hat{S}_n(\tilde{\theta}_n) = \hat{S}_n(\theta_0) + n(\tilde{\theta}_n - \theta_0)'\mathbb{P}_n\tilde{\ell}_0 - \frac{n}{2}(\tilde{\theta}_n - \theta_0)'\tilde{I}_0(\tilde{\theta}_n - \theta_0) + \Delta_n(\tilde{\theta}_n), \tag{9}$$

where $\Delta_n(\theta) = n\|\theta - \theta_0\|^3 \vee n^{1-2v}\|\theta - \theta_0\|$, for some $1/4 < v \leq 1/2$.

Conditions I1–I2 are very mild. Condition I1 is usually implied by the model identifiability conditions. Condition I2 is very weak since we only assume the existence of the asymptotic expansion (9) but not require the differentiability of $\hat{S}_n(\cdot)$. When $\hat{S}_n(\cdot)$ is the possibly non-differentiable log $pl_n(\cdot)$, I2 is implied by model assumptions M1–M4 in section 4 with v equal to r in (12). As for the differentiable $\hat{S}_n$, we can verify I2 using a three-term Taylor expansion of $\hat{S}_n(\cdot)$ with v being g in condition G of section 5.

Now we consider two types of grid search: deterministic type and stochastic type. In the former, we form a grid of cubes with sides of length $sn^{-\psi}$ over $\mathbb{R}^d$ for some $s>0$ and $0<\psi\le 1/2$, and thus obtain a set of points $\mathcal{D}_n=\{\theta_{iD}\}$ regularly spaced throughout Θ with cardinality $\text{card}(\mathcal{D}_n)\ge Cn^{d\psi}$ for some $C>0$. The grid point maximizing $\hat{S}_n(\theta)$ is thought of as $\hat{\theta}_n^{(0)}$. However, this deterministic search could be very slow if the dimension of θ is high. This motivates us to propose the stochastic search in which the search points are the realizations of some independent random variable $\bar{\theta}$. The magnitude of the stochastic search points remains n^{ψ} no matter how large the dimension d is. In theory, the stochastic grid search has significant computational savings over the deterministic approach. Theorem 1 below rigorously prove the convergence rates of the above numerical outcomes.

Theorem 1. *Let $\mathcal{D}_n$ be a set of points regularly spaced throughout Θ with* $\text{card}(\mathcal{D}_n)\ge Cn^{d\psi}$ *for some $C>0$ and $0<\psi\le 1/2$. Assume that $\bar{\theta}$ is independent of the data and admits a density having support Θ and bounded away from zero in some neighbourhood of θ_0. Let $\mathcal{S}_n$ be a set of realizations of $\bar{\theta}$ with* $\text{card}(\mathcal{S}_n)\ge\tilde{C}n^{\psi}$ *for some $\tilde{C}>0$ and $0<\psi\le 1/2$. Suppose that conditions I1–I2 hold, and that the parameter space Θ is compact. Then, if $\hat{\theta}_n$ defined in* (7) *is consistent and $\tilde{I}_0$ is non-singular, we have*

$$\theta_n^D-\theta_0=O_P(n^{-\psi}), \tag{10}$$

$$\theta_n^S-\theta_0=O_P(n^{-\psi}), \tag{11}$$

where $\theta_n^D=\arg\max_{\theta\in\mathcal{D}_n}\hat{S}_n(\theta)$ *and* $\theta_n^S=\arg\max_{\theta\in\mathcal{S}_n}\hat{S}_n(\theta)$.

Similar theorem is also proven in Robinson (1988) for parametric models. The strictly positive density assumption on $\bar{\theta}$ is reasonable. For example, $\bar{\theta}$ can be assumed to follow uniform or truncated normal distribution over the compact Θ. In practice, the search for the initial estimate is usually done over some compact set. Thus, the compactness of Θ is also reasonable.

In the following two models, there is no theoretically justified initial estimate available in the literature. Hence, we naturally apply the above grid search to obtain $\hat{\theta}_n^{(0)}$.

Example 1: Cox model under current status data. In the Cox proportional hazards model, the hazard function of the survival time T of a subject with covariate Z is expressed as:

$$\lambda(t\,|\,z)\equiv\lim_{\Delta\to 0}\frac{1}{\Delta}\Pr(t\le T<t+\Delta\,|\,T\ge t, Z=z)=\lambda(t)\exp(\theta' z),$$

where λ is an unspecified baseline hazard function. We consider the current status data where each subject is observed at a single examination time Y to determine if an event has occurred, but the event time T cannot be known exactly. Specifically, the observed data are n realizations of $X=(Y,\delta,Z)\in R^+\times\{0,1\}\times R$, where $\delta=I\{T\le Y\}$. The cumulative hazard function $\eta(y)=\int_0^y\lambda(t)\,dt$ is considered as the nuisance parameter. The parameter space $\mathcal{H}$ for η

is restricted to a set of non-decreasing and cadlag functions on some compact interval. In this model, it is well known that both $\hat{\eta}(\theta)$ and $\hat{S}_n(\theta)=\log pl_n(\theta)$ have no explicit forms, and can only be calculated numerically via the iterative convex minorant algorithm (see Huang, 1996). As for the convergence rate of η, Murphy & van der Vaart (1999) showed $\|\hat{\eta}(\tilde{\theta}_n)-\eta_0\|_2=O_P(\|\tilde{\theta}_n-\theta_0\|\vee n^{-1/3})$, where $\|\cdot\|_2$ is the L_2 norm. According to (S1) in supplementary material, we know that I2 is satisfied with $v=1/3$. Condition I1 is verified in lemma 2 of Lee *et al.* (2005) for this model.

Example 2: Semiparametric mixture model in case–control studies. Roeder *et al.* (1996) consider the logistic regression model with a missing covariate for case–control studies. In this model, they observe two independent random samples: one complete component $Y_C=(D_C, W_C)$ and Z_C of the size n_C, and one reduced component $Y_R=(D_R, W_R)$ of the size n_R. Following the assumptions given in Roeder *et al.* (1996), the likelihood for $x=(y_C, y_R, z_C)$ is given as

$$lik(\theta', \eta)(x)=p_{\theta'}(y_C\,|\,z_C)\eta\{z_C\}\int p_{\theta'}(y_R\,|\,z)\,\mathrm{d}\eta(z),$$

where $d\eta$ denotes the density of η w.r.t. some dominating measure, and

$$p_{\theta'}(y\,|\,z)=\left(\frac{\exp(\gamma+\theta e^z)}{1+\exp(\gamma+\theta e^z)}\right)^d\left(\frac{1}{1+\exp(\gamma+\theta e^z)}\right)^{1-d}\phi_\sigma(w-\alpha_0-\alpha_1 z),$$

where $\phi_\sigma(\cdot)$ denotes the density for $N(0,\sigma)$. The unknown parameters are $\theta'=(\theta,\alpha_0,\alpha_1,\gamma,\sigma)$ over the compact $\Theta'\subset\mathbb{R}^4\times(0,\infty)$ and the distribution η of the regression variable restricted to the set of non-degenerate probability distributions with a compact support. In this model, we will concentrate on θ while treating $\theta_2=(\alpha_0,\alpha_1,\ \gamma,\sigma)$ and η as nuisance parameters. The NPMLE $\hat{\eta}(\theta)$ is a weighted average of two empirical distributions and $\hat{S}_n(\theta)=\log pl_n(\theta)$ has no explicit form, both of which is computed efficiently via the iterative algorithm in [29]. For each fixed θ, denote the profile estimate as $(\hat{\theta}_{2,\theta},\hat{\eta}(\theta))$ so that $\hat{\theta}'_\theta=(\theta,\hat{\theta}_{2,\theta})$. Murphy & van der Vaart (1999) showed that, for any $\tilde{\theta}_n\xrightarrow{P}\theta_0$, $\|\hat{\eta}(\tilde{\theta}_n)-\eta_0\|_{BL_1}+\|\hat{\theta}'_{\tilde{\theta}_n}-\theta'_0\|=O_P(|\tilde{\theta}_n-\theta_0|\vee n^{-\frac{1}{2}})$, where $\|\cdot\|_{BL_1}$ is the weak topology. According to (S1) in supplementary materials, we know that I2 is satisfied with $v=1/2$. Condition I1 is verified in lemma 3 of Lee *et al.* (2005).

4. Semiparametric maximum likelihood estimation

In this section, we consider the MLE of θ corresponding to the case that $\hat{\eta}(\theta)$ is the well-defined NPMLE for $\eta_*(\theta)$ and $\hat{S}_n(\theta)=\log pl_n(\theta)$. We first discuss the construction of $\hat{\theta}_n^{(k)}$ even when the profile likelihood is unnecessarily differentiable, and then show that the minimal number of iterations k^* is jointly determined by the convergence rates of $\hat{\theta}_n^{(0)}$ and $\hat{\eta}(\theta)$.

In this section, we first assume the following convergence rate condition (12) and then the LFS conditions M1–M4. For any random sequence $\tilde{\theta}_n\xrightarrow{P}\theta_0$, we assume

$$\|\hat{\eta}(\tilde{\theta}_n)-\eta_0\|=O_P(\|\tilde{\theta}_n-\theta_0\|\vee n^{-r}), \tag{12}$$

where $\|\cdot\|$ is some norm in $\mathcal{H}$ and $1/4<r\leq 1/2$. Of course, we take the largest such r in the following and call it the convergence rate for estimating η. The above range of r holds in regular semiparametric models, which we can define without loss of generality to be models where the entropy integral converges. The value of r depends on the entropy number of $\mathcal{H}$; see theorems 3.1–3.2 in Murphy & van der Vaart (1999). In this section, we construct the LFS by following Cheng & Kosorok (2008b). Specifically, we first assume the

existence of a smooth map from the neighbourhood of θ into $\mathcal{H}$, which has the form $t \mapsto \eta_*(t, \theta, \eta)$ and satisfies $\eta_*(\theta, \theta, \eta) = \eta$ for any fixed $(\theta, \eta) \in \Theta \times \mathcal{H}$ (with a bit abuse of notation), and then define the map $t \mapsto \ell(x; t, \theta, \eta)$ as follows: $\ell(x; t, \theta, \eta) = \log lik(x; t, \eta_*(t, \theta, \eta))$. Thus, $\log pl_n(\theta) = \sum_{i=1}^n \ell(X_i; \theta, \theta, \hat{\eta}(\theta))$. From now on, we use the notation $\ell(t, \theta, \eta)$ for simplicity. We define $\dot{\ell}(t, \theta, \eta)$, $\ddot{\ell}(t, \theta, \eta)$ and $\ell^{(3)}(t, \theta, \eta)$ as the first, second and third derivative of $\ell(t, \theta, \eta)$ with respect to t, respectively. Also denote $\ell_{t,\theta}(t, \theta, \eta)$ as $(\partial^2/\partial t \partial \theta)\ell(t, \theta, \eta)$.

M1. We assume that the derivatives $(\partial^{l+m}/\partial t^l \partial \theta^m)\ell(t, \theta, \eta)$ have integrable envelope functions in $L_1(P)$ for $(l+m) \le 3$, and that the Fréchet derivatives of $\eta \mapsto \ddot{\ell}(\theta_0, \theta_0, \eta)$ and $\eta \mapsto \ell_{t,\theta}(\theta_0, \theta_0, \eta)$ are bounded around η_0.

M2. $E\dot{\ell}(\theta_0, \theta_0, \eta) = O(\|\eta - \eta_0\|^2)$ for all η around η_0.

M3. $\mathbb{G}_n(\dot{\ell}(\theta_0, \theta_0, \hat{\eta}(\tilde{\theta}_n)) - \dot{\ell}(\theta_0, \theta_0, \eta_0)) = O_P(n^{-2r+1/2} \vee n^{1/2-r}\|\tilde{\theta}_n - \theta_0\|)$ for any $\tilde{\theta}_n \xrightarrow{P} \theta_0$.

M4. The classes of functions $\{\ddot{\ell}(t, \theta, \eta)(x) : (t, \theta, \eta) \in V\}$ and $\{\ell_{t,\theta}(t, \theta, \eta)(x) : (t, \theta, \eta) \in V\}$ are P-Donsker, and $\{\ell^{(3)}(t, \theta, \eta)(x) : (t, \theta, \eta) \in V\}$ is P-Glivenko-Cantelli, where V is some neighbourhood of $(\theta_0, \theta_0, \eta_0)$.

See section 2.2 of Cheng & Kosorok (2008b) for the discussions on M1–M4.

Under conditions M1–M4 and (12), Cheng & Kosorok (2008b) showed the following second-order asymptotic linear expansion result:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{I}_0^{-1} \tilde{\ell}_0(X_i) + O_P(n^{-2r+1/2}). \tag{13}$$

We need to estimate $\sum_{i=1}^n \tilde{\ell}_0(X_i)/n$ and $\tilde{I}_0$ in (13) to construct the NR estimate $\hat{\theta}_n^{(k)}$. In view of (4) and (6), we can estimate them based on the following numerical derivatives of the log-profile likelihood:

$$\left[\hat{\ell}_n(\theta, s_n)\right]_i = \frac{\log pl_n(\theta + s_n v_i) - \log pl_n(\theta)}{n s_n}, \tag{14}$$

$$\left[\hat{I}_n(\theta, t_n)\right]_{i,j} = -\frac{\log pl_n(\theta + t_n(v_i + v_j)) + \log pl_n(\theta)}{n t_n^2} + \frac{\log pl_n(\theta + t_n v_i) + \log pl_n(\theta + t_n v_j)}{n t_n^2}. \tag{15}$$

Note that the above $\hat{I}_n$ is exactly the observed profile information proposed in Murphy & van der Vaart (1999). Lemma A1 of appendix implies that (14) and (15) are indeed consistent. Now we can construct $\hat{\theta}_n^{(k)}$ as

$$\hat{\theta}_n^{(k)} = \hat{\theta}_n^{(k-1)} + \left[\hat{I}_n\left(\hat{\theta}_n^{(k-1)}, t_n^{(k-1)}\right)\right]^{-1} \hat{\ell}_n\left(\hat{\theta}_n^{(k-1)}, s_n^{(k-1)}\right), \tag{16}$$

where step sizes $s_n^{(k-1)} \vee t_n^{(k-1)} = o(1)$. *A close inspection of* (16) *reveals that we have constructed* $\hat{\theta}_n^{(k)}$ *even without knowing the forms of* $\tilde{\ell}_0$ *and* $\tilde{I}_0$. *Therefore, it is a general construction approach.*

The convergence of $\hat{\theta}_n^{(k)}$ to $\hat{\theta}_n$, which is exactly the maximizer of $\log pl_n(\theta)$, as $k \to \infty$ is guaranteed by the asymptotic parabolic form of $\log pl_n(\theta)$ proven in Murphy & van der Vaart (2000). However, to figure out the minimal k^* such that $\|\hat{\theta}_n^{(k^*)} - \hat{\theta}_n\| = o_P(n^{-1/2})$, we need to make use of the second-order asymptotic quadratic expansion of $\log pl_n(\theta)$ derived in Cheng & Kosorok (2008b). As seen from (16), the orders of step sizes $(s_n^{(k-1)}, t_n^{(k-1)})$ are critical in determining the convergence rate of $\hat{\theta}_n^{(k)}$ to $\hat{\theta}_n$, and thus need to be properly chosen at each iteration. Lemma 1 below presents the theoretically optimal step sizes, under which the fastest convergence rate is achieved, at each iteration. The data-dependent choice of step sizes is important, and will be dealt in a separate study due to the space limitation.

Denote the convergence rate of $\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|$ as $O_P(n^{-r_{k-1}})$. Define $R_n \asymp r_n$ if $r_n/M \le R_n \le r_n M$ for some $M \ge 1$.

Lemma 1. *Suppose* (12) *and conditions M1–M4 hold. Also suppose that $\hat{\theta}_n$ is consistent and $\tilde{I}_0$ is non-singular. The convergence rate of $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\|$ is improved through the following three stages:*

(i) $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{3/2})$ *when* $r_{k-1} < r$ *and* $(s_n^{(k-1)}, t_n^{(k-1)}) \asymp (n^{-3r_{k-1}/2}, n^{-r_{k-1}/2})$;

(ii) $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{1/2} n^{-r})$ *when* $r \le r_{k-1} < 1/2$ *and* $(s_n^{(k-1)}, t_n^{(k-1)}) \asymp (n^{-r-r_{k-1}/2}, n^{-r_{k-1}/2})$;

(iii) $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(n^{-r-1/4})$ *when* $r_{k-1} \ge 1/2$ *and* $(s_n^{(k-1)}, t_n^{(k-1)}) \asymp (n^{-r-1/4}, n^{-r_{k-1}/2})$.

Now we present our second main theorem, that is, theorem 2, proving $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(n^{-S(\psi,r,k)})$, where $\hat{\theta}_n^{(0)}$ is n^{ψ}-consistent. From the form of $S(\psi, r, k)$, we can figure out the value of k^*. For example, according to the above lemma 1, it is easily seen that $S(1/2, r, k) = r + 1/4$ for any $1/4 < r \le 1/2$ and $k \ge 1$ (thus $k^* = 1$); and $S(1/3, 1/2, 1) = 1/2$ and $S(1/3, 1/2, k) = 3/4$ for any $k \ge 2$ (thus $k^* = 2$). This former case is the one-step iteration result in the semiparametric literature (given $\sqrt{n}$-consistent $\hat{\theta}_n^{(0)}$). As discussed in section 1, $\hat{\theta}_n^{(0)}$ may also have the sub-optimal rate. This explains why we are interested in deriving the general form of $S(\psi, r, k)$ for $\psi \le 1/2$.

Let $S_1(\psi, k) = \psi(3/2)^k$, $K_1(\psi, r) = \mathrm{int}[\log\,(r/\psi)/\log\,(3/2)]$ and $\tilde{S}_1(\psi, r) = S_1(\psi, K_1(\psi, r))$. Define, if $\tilde{S}_1(\psi, r) \ge 1/2$,

$$S(\psi, r, k) = \begin{cases} S_1(\psi, k) & k \le K_1(\psi, r) \\ r + 1/4 & k \ge K_1(\psi, r) + 1, \end{cases}$$

and, if $r \le \tilde{S}_1(\psi, r) < 1/2$,

$$S(\psi, r, k) = \begin{cases} S_1(\psi, k) & k \le K_1(\psi, r) \\ S_2(\tilde{S}_1(\psi, r), r, k - K_1(\psi, r)) & K_1(\psi, r) < k \le K_1(\psi, r) + \tilde{K}_2(\psi, r), \\ r + 1/4 & k \ge K_1(\psi, r) + \tilde{K}_2(\psi, r) + 1, \end{cases}$$

where $S_2(\psi, r, k) = 2r + 2^{-k}(\psi - 2r)$, $K_2(\psi, r) = \mathrm{int}[\log\{(2r - \psi)/(2r - 1/2)\}/\log 2]$ and $\tilde{K}_2(\psi, r) = K_2(\tilde{S}_1(\psi, r), r)$.

Theorem 2. *Suppose that conditions in lemma 1 hold and proper step sizes are chosen according to lemma 1. Let $\hat{\theta}_n^{(k)}$ be the k-step estimator defined in* (16) *and $\hat{\theta}_n^{(0)}$ be n^{ψ}-consistent for $0 < \psi \le 1/2$. Recall that $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(n^{-r_k})$. We show that r_k increases from ψ to $(r + 1/4)$ as $k \to \infty$. Specifically, we have*

$$\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(n^{-S(\psi,r,k)}). \tag{17}$$

This implies that

$$\|\hat{\theta}_n^{(k^*)} - \hat{\theta}_n\| = o_P(n^{-1/2}), \tag{18}$$

where $k^* = K_1(\psi, r) + \widetilde{\mathrm{int}}[\log\,((2r - \tilde{S}_1(\psi, r))/(2r - 1/2))/\log\,2]$.

Interestingly, we notice that the optimal bound of $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\|$, that is, $O_P(n^{-r-1/4})$, is intrinsically determined by how accurately we estimate the nuisance parameter, that is, the value of r. This bound cannot be further improved unless we are willing to make stronger

Table 1. *Cox model under current status data* ($r=1/3$)

	$\psi=1/2$	$\psi=1/3$	$\psi=1/4$
Cox models	$r_1=7/12$ $k^*=1$	$r_1=1/2, r_2=7/12$ $k^*=2$	$r_1=3/8, r_2=25/48, r_3=7/12$ $k^*=2$

Remark: Define $\|\hat{\theta}_n^{(0)}-\hat{\theta}_n\|=O_P(n^{-\psi})$ and $\|\hat{\theta}_n^{(k)}-\hat{\theta}_n\|=O_P(n^{-r_k})$.

Table 2. *Semiparametric mixture model in case–control studies* ($r=1/2$)

	$\psi=1/2$	$\psi=1/3$	$\psi=1/4$
Mixture models	$r_1=3/4$ $k^*=1$	$r_1=1/2, r_2=3/4$ $k^*=2$	$r_1=3/8, r_2=9/16, r_3=3/4$ $k^*=2$

Remark: Define $\|\hat{\theta}_n^{(0)}-\hat{\theta}_n\|=O_P(n^{-\psi})$ and $\|\hat{\theta}_n^{(k)}-\hat{\theta}_n\|=O_P(n^{-r_k})$.

assumptions than M1–M4, which seem unrealistic. The form of $S(\psi, r, k)$ implies that more accurate initial estimate leads to more efficient $\hat{\theta}_n^{(k)}$ (in terms of closer distance to $\hat{\theta}_n$).

We apply theorem 2 to the previous two examples, and the required conditions are verified in Cheng & Kosorok (2008a, b) for them.

Example 1: Cox model under current status data (cont'). According to theorem 2, we establish Table 1 to depict the convergence of $\hat{\theta}_n^{(k)}$ to $\hat{\theta}_n$ given different initial estimates until it reaches the lower bound $O_P(n^{-7/12})$.

Example 2: Semiparametric mixture model in case–control studies (cont'). The following Table 2 is similar as Table 1. In Table 2, we notice that $\hat{\theta}_n^{(k)}$ converges to $\hat{\theta}_n$ at a faster rate due to the larger r.

5. Semiparametric estimation under regularization

In this section, we consider the semiparametric estimation under two types of regularizations, that is, kernel estimation and penalized estimation. In contrast with the profile likelihood estimation, the regularized $\hat{S}_n(\theta)$ is usually smooth and differentiable although its form may vary under different regularizations. We first present a unified framework for studying $\hat{\theta}_n^{(k)}$ when $\hat{S}_n(\theta)$ is third-order differentiable, and then give easy-to-verify sufficient conditions for kernel estimate and penalized estimation, respectively. In the end, we discuss the iterative sparse estimation of partly linear models as an extension of the penalized estimation.

In this section, we construct $\hat{\theta}_n^{(k)}$ as follows:

$$\hat{\theta}_n^{(k)}=\hat{\theta}_n^{(k-1)}+\left[\hat{I}_n(\hat{\theta}_n^{(k-1)})\right]^{-1}\hat{\ell}_n(\hat{\theta}_n^{(k-1)}), \tag{19}$$

where $\hat{\ell}_n(\cdot)=\hat{S}_n^{(1)}(\cdot)/n$ and

$$\hat{I}_n(\cdot)=-\hat{S}_n^{(2)}(\cdot)/n, \tag{20}$$

where $\hat{S}_n^{(j)}(\cdot)$ is the jth derivative of $\hat{S}_n(\cdot)$. When $\hat{S}_n^{(2)}(\theta)$ has no explicit form or is hard to compute, we may prefer constructing $[\hat{I}_n(\theta)]_{ij}$ as

$$-\frac{1}{n}\times\frac{[\hat{S}_n^{(1)}(\theta+n^{-1/2}t_2v_j)]_i-[\hat{S}_n^{(1)}(\theta+n^{-1/2}t_1v_j)]_i}{n^{-1/2}t_2-n^{-1/2}t_1}, \tag{21}$$

where t_1 and t_2 ($t_1<t_2$) are arbitrarily fixed real numbers.

Recall that $S_n(\theta)=\sum_{i=1}^n \log\ lik(X_i;\theta,\eta_*(\theta))$ and define $S_n^{(j)}(\cdot)$ as the jth derivative of $S_n(\cdot)$. In view of the discussions in section 2, that is (4) and (6), we expect that $\hat\theta_n^{(k)}$ converges to $\hat\theta_n$ if $\hat S_n^{(j)}(\cdot)$ approximates $S_n^{(j)}(\cdot)$ well enough round θ_0 for $j=1,2,3$. Therefore, we assume the following general condition G.

G. Assume that

$$\frac{1}{n}\hat S_n^{(1)}(\theta_0)-\frac{1}{n}S_n^{(1)}(\theta_0)=O_P(n^{-2g}), \tag{22}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\hat S_n^{(2)}(\theta)-\frac{1}{n}S_n^{(2)}(\theta)\right|=O_P(n^{-g}), \tag{23}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\hat S_n^{(3)}(\theta)\right|=O_P(1), \tag{24}$$

where $1/4<g\le 1/2$.

In the kernel (penalized) estimation, the value of g is determined by the bandwidth order of the kernel function (the order of the smoothing parameter). We may verify (24) by showing

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\hat S_n^{(3)}(\theta)-\frac{1}{n}S_n^{(3)}(\theta)\right|=o_P(1), \tag{25}$$

and that the class of functions $\{(\partial^3/\partial\theta^3)\log\ lik(x;\theta,\eta_*(\theta)):\theta\in\mathcal{N}(\theta_0)\}$ is P-Glivenko-Cantelli and

$$\sup_{\theta\in\mathcal{N}(\theta_0)} E\left|(\partial^3/\partial\theta^3)\log\ lik(X;\theta,\eta_*(\theta))\right|<\infty.$$

Now we present our third main theorem, that is, theorem 3. Define

$$R(\psi,g,k)=\begin{cases} R_1(\psi,g,k) & k\le L_1(\psi,g)\\ R_2(R_1(\psi,g,L_1(\psi,g)),g,k-L_1(\psi,g)) & k>L_1(\psi,g)\end{cases} \tag{26}$$

where $R_1(\psi,g,k)=(1/2-g)+2^k(\psi+g-1/2)$, $L_1(\psi,g)=\mathrm{int}[\log\ (g/(g+\psi-1/2))/\log 2]$, $\tilde L_1(\psi,g)=\widetilde{\mathrm{int}}[\log\ (g/(g+\psi-1/2))/\log 2]$ and $R_2(\psi,g,k)=kg+\psi$.

Theorem 3. *Suppose that condition G holds, $\hat\theta_n$ defined in* (7) *is consistent and $\tilde I_0$ is non-singular. We have the following asymptotic linear expansion of $\hat\theta$:*

$$\sqrt{n}(\hat\theta_n-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde I_0^{-1}\tilde\ell_0(X_i)+O_P(n^{1/2-2g}). \tag{27}$$

Let $\hat\theta_n^{(k)}$ be the k-step estimator defined in (19) *and $\hat\theta_n^{(0)}$ be n^ψ-consistent for* $(1/2-g)<\psi\le 1/2$. *Define $\|\hat\theta_n^{(k)}-\hat\theta_n\|=O_P(n^{-r_k})$. We show that r_k increases from ψ to ∞ as $k\to\infty$. Specifically, we show*

$$\|\hat\theta_n^{(k)}-\hat\theta_n\|=O_P(n^{-2^k\psi})\qquad \textit{if } \hat I_n(\cdot) \textit{ is defined in } (20), \tag{28}$$

$$\|\hat\theta_n^{(k)}-\hat\theta_n\|=O_P(n^{-R(\psi,g,k)})\qquad \textit{if } \hat I_n(\cdot) \textit{ is defined in } (21). \tag{29}$$

This implies that $\|\hat\theta_n^{(k^*)}-\hat\theta_n\|=o_P(n^{-1/2})$, where $k^*=\widetilde{int}[\log\ (1/2\psi)/\log 2]$ for (28) *and $k^*=\tilde L_1(\psi,g)$ for* (29).

Note that (28) is a statistical counterpart to the well known quadratic convergence of the NR algorithm (see Ortega & Rheinboldt, 1970, p. 312). Theorems 2 and 3 imply that

(i) $\hat{\theta}_n^{(k^*)}$ shares the same limit distribution as $\hat{\theta}_n$ and gains more asymptotic efficiency, that is, smaller error term in its asymptotic linear expansion, if more iterations are implemented; (ii) the higher order asymptotic efficiency of $\hat{\theta}_n^{(k)}$ is determined by how accurately η is estimated, that is, the values of r or g; (iii) $\hat{\theta}_n^{(k)}$ converges to $\hat{\theta}_n$ faster when $\hat{I}_n$ is constructed as an analytical derivative no matter whether the regularization is used or not.

A by-product of theorem 3 is its application to the parametric models, that is, η is known, where $\hat{S}_n(\theta)$ becomes the parametric log-likelihood. We skip the relevant discussions due to the similarity.

Remark 1. Given that the initial estimate is $\sqrt{n}$ consistent, we have

$$\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(n^{-2^{k-1}}) \qquad \text{if } \hat{I}_n(\cdot)\text{is constructed as in (20),}$$
$$\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(n^{-(1/2+kg)}) \quad \text{if } \hat{I}_n(\cdot) \text{ is constructed as in (21)}$$

based on theorem 3. This implies $k^* = 1$.

Remark 2. Theorems 2–3 together with the previous theorem 1 offer rigorous statistical analysis for the iterative semiparametric estimation approach. Those theorems also indicate a tradeoff between the computational cost of searching for an initial estimate, that is, *card*$(\mathcal{D}_n)$ or *card*$(\mathcal{S}_n)$, and that of generating an efficient estimate, that is, k^*.

5.1. Kernel estimation in semiparametric models

In this subsection, we consider the kernel stimation in semiparametric models (see Speckman, 1988; Andrews, 1995). In particular, the kernel approach is proven to be a powerful inferential tool for the class of conditionally parametric models (CPM) (see Severini & Wong, 1992; Severini & Stainswalis, 1994). The practical performance of the iterative estimation procedure (I)–(IV) for the CPM is extensively studied in Severini & Stainswalis, (1994). Thus, we will focus on the class of CPM although our conclusions can be extended to more general class of semiparametric models by incorporating the results in Andrews (1995). Under kernel estimation, k^* is shown to depend on the order of bandwidth used in the kernel function.

The class of CPM was first introduced by Severini & Wong (1992) and further generalized to the quasi-likelihood framework by Severini & Staniswalis (1994). Specifically, we observe $X = (Y, W, Z)$ such that the distribution of Y conditional on partitioned covariates $W = w$ and $Z = z$ is parameterized by a finite dimensional parameter $\phi = (\theta, \lambda_z)$, where $\lambda_z \in H \subset \mathbb{R}$ depends on the value of z as a function $\eta(z)$. The joint distribution of (W, Z) is assumed to be independent of ϕ.

We assume that $\eta \in C^2(\mathcal{Z})$, where $\mathcal{Z}$ is the support of z. An important feature of CPM is that its least favourable curve can be expressed as (see Severini & Wong, 1992 for details)

$$\eta_*(z;\theta) = \arg \sup_{\eta \in C^2(\mathcal{Z})} E[\log\, lik(X;\theta,\eta) \,|\, Z = z], \tag{30}$$

and thus its kernel estimate is written as

$$\hat{\eta}(z;\theta) = \arg \sup_{\eta \in C^2(\mathcal{Z})} \sum_{i=1}^{n} \log\, lik(X_i;\theta,\eta(Z_i)) K\left(\frac{z - Z_i}{b_n}\right), \tag{31}$$

where $K(\cdot)$ is a kernel with the bandwidth $b_n \to 0$. For example, if $(Y \mid w = W, Z = z) \sim N(\theta' w, \eta(z))$, then we have

$$\begin{aligned} \hat\eta(z;\theta) &= \frac{\sum_{i=1}^n (Y_i - \theta' W_i)^2 K((z - Z_i)/b_n)}{\sum_{i=1}^n K((z - Z_i)/b_n)}, \\ \hat S_n(\theta) &= -\frac{1}{2}\sum_{i=1}^n \frac{(Y_i - \theta' W_i)}{\hat\eta(Z_i;\theta)} - \frac{1}{2}\sum_{i=1}^n \log\, \hat\eta(Z_i;\theta). \end{aligned} \tag{32}$$

In general, $\hat\eta(\theta)$ and $\hat S_n(\theta)$ have no explicit form. However, based on (31), we can control the asymptotic behaviours of $\hat\eta(\theta)$ (and thus $\hat S_n(\theta)$) through proper kernel conditions.

By exploiting the parametric structure of CPM, we will show that $\hat S_n(\theta)$ satisfies the general condition G under the below conditions K1–K2 and C1–C2.

K1. For arbitrary $\theta_1 \in \Theta$ and $\lambda_1 \in H$, $E_{\theta_1,\lambda_1} \log\, lik(X;\theta,\lambda) < E_{\theta_1,\lambda_1} \log\, lik(X;\theta_1,\lambda_1)$ if $\theta \neq \theta_1$.

K2. Assume that

$$E\left\{ \sup_{(\theta,\lambda)\in\Theta\times H} \left| \frac{\partial^{r+s} \log\, lik(X;\theta,\lambda)}{\partial\theta^r \partial\lambda^s} \right|^2 \right\} < \infty \tag{33}$$

for all $r, s = 0, \ldots, 4$ and $r + s \leq 4$.

Conditions C1–C2 below are about the smoothness and convergence rate of $\eta_*(\theta)$ and $\hat\eta(\theta)$. Denote the sth derivative of $\eta_*(\theta)$ ($\hat\eta(\theta)$) w.r.t. θ as $\eta_*^{(s)}(\theta)$ ($\hat\eta^{(s)}(\theta)$), and their values at θ_0 as $\eta_{*0}^{(s)}$ ($\hat\eta_0^{(s)}$).

C1. Assume that, for all $r, s = 0, 1, 2, 3$ and $r + s \leq 3$, $(\partial^{r+s}/\partial z^r \partial\theta^s)\eta_*(z;\theta)$ and $(\partial^{r+s}/\partial z^r \partial\theta^s)$ $\hat\eta(z;\theta)$ exist and $\sup_{\theta\in\mathcal{N}(\theta_0)} \|\eta_*^{(s)}(\theta)\|_\infty < \infty$.

C2. Assume that

$$\sup_{\theta\in\mathcal{N}(\theta_0)} \|\hat\eta^{(s)}(\theta) - \eta_*^{(s)}(\theta)\|_\infty = O_P(n^{-g}) \quad \text{for } s = 0, 1, 2, \tag{34}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)} \|\hat\eta^{(3)}(\theta) - \eta_*^{(3)}(\theta)\|_\infty = o_P(1), \tag{35}$$

$$\left\| \frac{\partial}{\partial z}\hat\eta_0(z) - \frac{\partial}{\partial z}\eta_{*0}(z) \right\|_\infty = o_P(n^{-\delta}), \tag{36}$$

$$\left\| \frac{\partial}{\partial z}\hat\eta_0^{(1)}(z) - \frac{\partial}{\partial z}\eta_{*0}^{(1)}(z) \right\|_\infty = o_P(n^{-\delta}), \tag{37}$$

for some $g \in (1/4, 1/2]$ and $(2g - 1/2) \leq \delta \leq g$.

In view of (30)–(31), we can verify C2 by applying the kernel theories under some proper kernel conditions and K1–K2; see lemma 2 below. Note that condition C2 implies (12) assumed for the NPMLE since $\|\hat\eta(\tilde\theta_n) - \eta_0\| \leq \|\hat\eta(\tilde\theta_n) - \hat\eta(\theta_0)\|_\infty + \|\hat\eta(\theta_0) - \eta_*(\theta_0)\|_\infty \leq O_P(\|\tilde\theta_n - \theta_0\| \vee n^{-g})$ by the construction that $\eta_*(\theta_0) = \eta_0$, C1–C2 and (34). Thus, our conditions K1–K2 and C1–C2 for the class of CPM are generally stronger than M1–M4 and (12).

Theorem 4. *Assuming that conditions K1–K2 and C1–C2 hold, then the condition G required in theorem 3 is satisfied for the kernel estimation in CPM.*

The consistency of $\hat\theta_n$ required in theorem 3 can be established if we further require the global condition $\sup_{\theta\in\Theta} \|\hat\eta(\theta) - \eta_*(\theta)\|_\infty \to 0$, see proposition 1 of Severini & Wong (1992). In the third example, we apply theorems 1 and 4 to a subclass of CPM, called conditionally exponential models (CEM), in which $\hat{\eta}(\theta)$ has a closed-form. This makes the verifications of C1–C2 much easier. The relation between k^* and the order of b_n in (31) is specified in lemma 2.

Example 3. Conditionally exponential models. In CEM, there exists a function $\psi_\theta(\cdot)$ such that the conditional distribution of $\psi_\theta(Y, W)$ given $Z=z$ does not depend on θ and forms an exponential family. And its log-likelihood is expressed as $\log\, lik(X;\theta,\eta)=\psi_\theta(Y,W)T(\eta(Z))-A(\eta(Z))+S(\psi_\theta(Y,W))$ for some functions T, A and S. Some simple algebra gives that

$$\hat{\eta}(z;\theta)=\rho\left(\frac{\sum_{i=1}^{n}\psi_\theta(Y_i,W_i)K((z-Z_i)/b_n)}{\sum_{i=1}^{n}K((z-Z_i)/b_n)}\right), \tag{38}$$

where $\eta=\rho\{E_{\theta,\eta}(\psi_\theta(Y,W))\}$. In the previous conditional normal model, we have $\psi_\theta(Y,W)=(Y-\theta' W)^2$ and $\rho(t)=t$. Another example is that $(Y\mid W=w, Z=z)\sim$ Exp$(0,\exp\,(\theta' w+\eta(z)))$ in which $\psi_\theta(Y,W)=Y\exp(-\theta' W)$ and $\rho(t)=\log\, t$.

We first apply theorem 1 to obtain $\hat{\theta}_n^{(0)}$. Condition I1 can be verified by adapting the consistency proof of $\hat{\theta}_n$ in Severini and Wong (1992), see than proposition 1. Condition I2 just follows from condition G with v being g given in (39) as discussed in section 3. We next discuss how to verify K1–K2 and C1–C2 in theorem 4. Conditions K1–K2 are easily verified when $\Theta\times H$ is assumed to be compact. However, we need the following lemma to verify conditions C1–C2. Let $\psi_\theta^{(j)}(\cdot)$ be $(\partial^j/\partial\theta^j)\psi_\theta(\cdot)$ and $f_{\theta j}(\cdot\mid z)$ be its conditional density. Denote $f(z)$ as the marginal density of Z. Let M be a compact set so that $m_\theta(z)\equiv E[\psi_\theta(Y,W)\mid Z=z]\in \mathrm{int}(M)$ for all z, θ.

Lemma 2. *Assume the following conditions hold:*

(a) $E\{\sup_\theta|\psi_\theta^{(j)}|\}<\infty$ *for* $j=0,1,2,3$.
(b) *For some even integer* $q\geq 10$, $\sup_\theta E\{|\psi_\theta^{(j)}|^q\}<\infty$ *for* $j=0,1,2,3$.
(c) $\sup_\theta\sup_x|f_{\theta j}^{(r)}(y,w|z)|<\infty$ *for* $j=0,1,2$ *and* $r=0,\ldots,4$.
(d) $\sup_z|f^{(r)}(z)|<\infty$ *for* $r=0,\ldots,4$.
(e) $0<\inf_z f(z)\leq\sup_z f(z)<\infty$.
(f) $\sup_{m\in M}|\rho^{(j)}(m)|<\infty$ *for* $j=0,\ldots,4$.

Suppose that the kernel function $K(\cdot)$ *in* (38) *satisfies*

$$\int K(u)\,\mathrm{d}u=1,\ \int uK(u)\,\mathrm{d}u=0,\ \int u^2K(u)\,\mathrm{d}u<\infty \text{ and } \sup_u|K^{(r)}(u)|<\infty \text{ for } r=0,\ldots,4.$$

Condition C1 holds under the above conditions. If we choose $b_n\asymp n^{-\alpha}$ for $1/8<\alpha<(q-2)/(4q+16)$*, then condition C2 is satisfied with, for any* $\epsilon>0$,

$$g=2\alpha\wedge\left(\frac{q}{2q+4}-\frac{\alpha(q+4)}{q+2}-\epsilon\right), \tag{39}$$

$$\delta=\frac{q}{2q+4}-\frac{\alpha(2q+6)}{q+2}-2\epsilon. \tag{40}$$

The above lemma specifies the relation between the bandwidth order α in (38) and k^* in theorem 3. By some algebra, we can verify that $g\in(1/4,1/2]$ and $(2g-1/2)\leq\delta\leq g$ given the above range of α and q. We want to point out that the convergence rates of $\hat{\eta}(\theta)$ (and its derivatives) may be improved, that is, larger value of g, under more restrictive kernel conditions, (see Stainswalis, 1989; Andrews, 1995).

Table 3. *Conditional normal (exponential) model* ($g=151/600$)

	$\psi=1/2$	$\psi=1/3$	$\psi=1/4$
Construction I	$r_1=1$ $k^*=1$	$r_1=2/3$ $k^*=1$	$r_1=1/2, r_2=1$ $k^*=2$
Construction II	$r_1=451/600$ $k^*=1$	$r_1=251/600, r_2=353/600$ $k^*=2$	$r_1=151/600, r_2=153/600,$ $r_3=157/600, r_4=165/600$ $r_5=181/600, r_6=213/600,$ $r_7=277/600, r_8=405/600$ $k^*=8$

Remark: Define $\|\hat\theta_n^{(0)}-\hat\theta_n\|=O_P(n^{-\psi})$ and $\|\hat\theta_n^{(k)}-\hat\theta_n\|=O_P(n^{-r_k})$; construction I: $\hat I_n$ is constructed by (20); construction II: $\hat I_n$ is constructed by (21).

Now we are ready to apply theorem 1 and lemma 2 to the previous conditional normal (exponential) example, in which q is shown to be arbitrarily large and M is chosen as a sufficiently large compact subset of $(0,\infty)$. In Table 3, we assume that $q=28$, $b_n \asymp n^{-1/5}$, $\epsilon=1/600$ such that $g=151/600>1/4$ and $\delta=1/20$ according to (39)–(40).

In the end of this section, we empirically verify our theoretical results via some simulations. For simplicity, we consider the conditional exponential model $(Y\mid W=w, Z=z)\sim \mathrm{Exp}(0,\exp(\theta w+\eta(z)))$, where $\theta_0=1$ and $\eta_0(z)=-z^2$. The covariates Z and W were generated from Unif $[0,1]$ independently. According to (38), the non-parametric estimate is calculated as

$$\hat\eta(z;\theta)=\log\left(\frac{\sum_{i=1}^n Y_i\exp(-\theta W_i)K((z-Z_i)/b_n)}{\sum_{i=1}^n K((z-Z_i)/b_n)}\right),$$

where $K(\cdot)$ is Gaussian kernel with the smoothing bandwidth b_n selected by R function 'density' (see Silverman, 1986). The initial estimate was identified by the deterministic grid search with the grid size approximately $n^{-1/4}$, for example, approximately 0.25 for $n=250$. Hence, the initial estimate $\hat\theta^{(0)}$ has the convergence rate $n^{-1/4}$, that is, $\psi=1/4$. The sample size was taken to be 150, 200, 250. In Figs 1 and 2, we plot the mean and standard deviation of $\hat\theta_n^{(k)}$ v.s. k for $k=1,2,\ldots,8$ out of 100 simulated data sets under constructions I and II (with $t_1=1$ and $t_2=2$), respectively. Under construction II, that is (21), we find that our simulation results are insensitive to the choice of (t_1,t_2) due to the differentiability of $\hat S_n^{(1)}$. It is clear to see that the mean (SD) of $\hat\theta_n^{(k)}$ converges to approximately 1 (4.2) after two steps

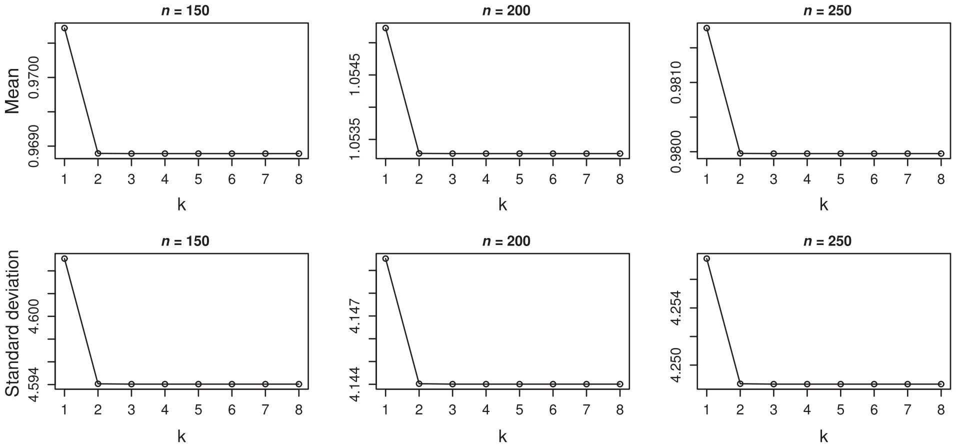

Fig. 1. Mean and standard deviation of $\hat\theta_n^{(k)}$ under construction I (100 simulated data sets).

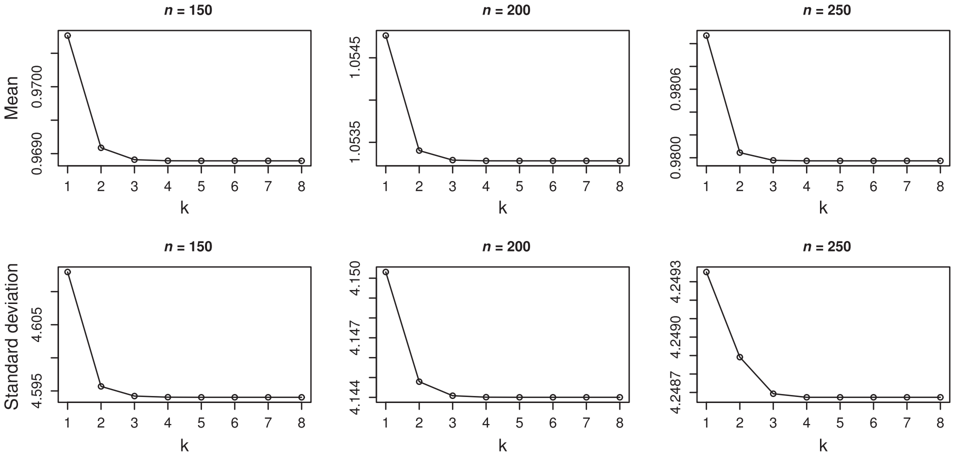

Fig. 2. Mean and standard deviation of $\hat{\theta}_n^{(k)}$ under construction II (100 simulated data sets).

in construction I and four steps in construction II. In view of Table 3, we empirically confirm our theoretical finding that the iterative estimate needs less steps to achieve convergence in construction I. Note that the derived convergence rate $n^{-151/600}$ in example 3 may not be sharp; see discussions after lemma 2. This may explain why four rather than eight iterations are needed in practice under construction II.

5.2. Penalized estimation in semiparametric models

In many semiparametric models, it is also common to perform estimation using penalization which also yields fully efficient estimates for θ (e.g. Mammen & van der Geer, 1997). In this case, we will show that the value of k^* relates to the order of the smoothing parameter λ_n. A surprising result we find is that k^* iterations are also sufficient for recovering the estimation sparsity in high dimensional data, see the partly linear example below.

In this subsection, we assume that η belongs to the Sobolev class of functions $\mathcal{H}_k \equiv \{\eta : J^2(\eta) = \int_{\mathcal{Z}} (\eta^{(k)}(z))^2 \,\mathrm{d}z < \infty\}$, where $\eta^{(j)}$ is the jth derivative of η and $\mathcal{Z}$ is some compact set on the real line. The penalized log-likelihood in this context is defined as

$$\log lik_{\lambda_n}(\theta, \eta) = \log lik_n(\theta, \eta) - n\lambda_n^2 J^2(\eta), \tag{41}$$

where λ_n is a smoothing parameter. We assume the following bounds for λ_n:

$$\lambda_n = o_P(n^{-1/4}) \quad \text{and} \quad \lambda_n^{-1} = O_P(n^{k/(2k+1)}). \tag{42}$$

In practice, λ_n can be obtained by cross-validation (Wahba, 1998). Here, $\hat{S}_n(\theta)$ becomes the log-profile penalized likelihood $\hat{S}_{\lambda_n}(\theta)$: $\hat{S}_{\lambda_n}(\theta) = \log lik_{\lambda_n}(\theta, \hat{\eta}_{\lambda_n}(\theta))$, where $\hat{\eta}_{\lambda_n}(\theta) = \arg\sup_{\eta \in \mathcal{H}_k} \log lik_{\lambda_n}(\theta, \eta)$ for any fixed θ and λ_n. We define the penalized estimate as $\hat{\theta}_{\lambda_n}$. The construction of the k-step penalized estimate $\hat{\theta}_{\lambda_n}^{(k)}$ follows from (19) just with the change of $\hat{S}_n(\cdot)$ to $\hat{S}_{\lambda_n}(\cdot)$. For the penalized estimation, we need to slightly modify condition G as follows:

G'. Assume that, for some constant $c \neq 0$,

$$\frac{1}{n}\hat{S}_{\lambda_n}^{(1)}(\theta_0) - \frac{c}{n}\sum_{i=1}^{n} \tilde{\ell}_0(X_i) = O_P(\lambda_n^2), \tag{43}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\hat{S}^{(2)}_{\lambda_n}(\theta)+c\tilde{I}_0\right|=O_P(\lambda_n\vee\|\theta-\theta_0\|), \tag{44}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\hat{S}^{(3)}_{\lambda_n}(\theta)\right|=O_P(1). \tag{45}$$

It is easy to verify condition G' if $\hat{\eta}_{\lambda_n}(\theta)$ has an explicit expression and log $lik_{\lambda_n}(\theta,\eta)$ is smooth w.r.t. (θ,η), see the example 4 below. We also want to point out that condition G' is relaxable to a large extent, see remark 3.

In view of (4) and (6), we can prove theorem 5 similarly as theorem 3. Theorem 5 implies that k^* depends on the order of the smoothing parameter λ_n, that is, the value of g, see (29).

Theorem 5. *Suppose condition G' holds, the penalized MLE $\hat{\theta}_{\lambda_n}$ is consistent and $\tilde{I}_0$ is non-singular. We have*

$$\sqrt{n}(\hat{\theta}_{\lambda_n}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{I}_0^{-1}\tilde{\ell}_0(X_i)+O_P(\sqrt{n}\lambda_n^2). \tag{46}$$

Define $g=\max\{g':\lambda_n=O_P(n^{-g'})\}$, and thus $1/4<g\le k/(2k+1)$ based on condition (42). Construct $\hat{\theta}^{(k)}_{\lambda_n}$ as in (19) with the change of $\hat{S}_n(\cdot)$ to $\hat{S}_{\lambda_n}(\cdot)$. Then all the conclusions for $\hat{\theta}^{(k)}_n$ in theorem 3 also hold for $\hat{\theta}^{(k)}_{\lambda_n}$.

Remark 3. We want to mention that the penalized profile log-likelihood may not be differentiable in some semiparametric models, for example, the partly linear models under current status data studied in Cheng & Kosorok (2009). In such cases, we can take the discretization approach to construct $\hat{\theta}^{(k)}_n$ as in the profile likelihood framework, that is (16), and obtain similar results as in theorem 2 by assuming the condition that $\|\hat{\eta}_{\lambda_n}(\tilde{\theta}_n)-\eta_0\|=O_P(\|\tilde{\theta}_n-\theta_0\|\vee\lambda_n)$ for any consistent $\tilde{\theta}_n$. This requires the use of the higher order quadratic expansion of the penalized profile likelihood derived in Cheng & Kosorok (2009).

We next apply theorem 5 to the following partly linear models under sparse assumption. Interestingly, we discover that one-step iteration is sufficient for achieving the semiparametric estimation efficiency and recovering the estimation sparsity simultaneously.

Example 4. Sparse and efficient estimation of partial spline model. In the partial smoothing spline model, we consider $Y=W'\theta+\eta(Z)+\epsilon$, where $\eta\in\mathcal{H}_k$ and $0\le Z\le 1$. For simplicity, we assume that $\epsilon\overset{\text{i.i.d}}{\sim}N(0,\sigma^2)$ and is independent of (W,Z). The normality of ϵ can be relaxed to the sub-exponential tail condition. In this example, we assume that some components of θ_0 are exactly zero which is common for high dimensional data. To achieve the estimation efficiency and recover the sparsity of θ, Cheng & Zhang (2012) proposed the following double penalty regularization:

$$(\hat{\theta}_{\lambda_n},\hat{\eta}_{\lambda_n})=\arg\min_{\Theta\times\mathcal{H}_k}\left\{\sum_{i=1}^{n}(Y_i-W_i'\theta-\eta(Z_i))^2+n\lambda_n^2J^2(\eta)+n\tau_n^2\sum_{j=1}^{d}\frac{|\theta_j|}{|\tilde{\theta}_j|^\gamma}\right\}, \tag{47}$$

where γ is a fixed positive constant, $\tilde{\theta}=(\tilde{\theta}_1,\ldots,\tilde{\theta}_d)'$ is the consistent initial estimate and τ_n is the smoothing parameter for the purpose of sparsity.

The standard smoothing spline theory suggests that

$$\hat{\eta}_{\lambda_n}(\mathbf{z};\theta)=A(\lambda_n)(\mathbf{y}-\mathbf{w}\theta), \tag{48}$$

where $\hat{\eta}_{\lambda_n}(\mathbf{z};\theta)=(\hat{\eta}_{\lambda_n}(z_1;\theta),\ldots,\hat{\eta}_{\lambda_n}(z_n;\theta))'$, $\mathbf{y}=(y_1,\ldots,y_n)'$ and $\mathbf{w}=(w_1',\ldots,w_n')'$. The expression of the $n\times n$ influence matrix $A(\lambda_n)$ can be found in Heckman (1986). Therefore, $\hat{\eta}_{\lambda_n}(\theta)$ is a natural spline of order $(2k-1)$ with knots on z_i' for any fixed θ. Plugging (48) back to (47), we have

$$\hat{S}_{\lambda_n}(\theta)=\tilde{S}_{\lambda_n}(\theta)+n\tau_n^2\sum_{j=1}^{d}\frac{|\theta_j|}{|\tilde{\theta}_j|^{\gamma}}, \tag{49}$$

where

$$\tilde{S}_{\lambda_n}(\theta)=(\mathbf{y}-\mathbf{w}\theta)'[I-A(\lambda_n)](\mathbf{y}-\mathbf{w}\theta) \tag{50}$$

and I is the identity matrix of size n. Note that $\hat{\theta}_{\lambda_n}$ does not have an explicit expression, and has to be iteratively computed using software like Quadratic Programming or LARS (Efron *et al.*, 2004). Specifically, based on (19)–(20), we construct $\hat{\theta}_{\lambda_n}^{(1)}$ as follows:

$$\hat{\theta}_{\lambda_n}^{(1)}=\hat{\theta}_{\lambda_n}^{(0)}+\left[\frac{\mathbf{w}'(I-A(\lambda_n))\mathbf{w}}{n}\right]^{-1}\left[\frac{\mathbf{w}'(I-A(\lambda_n))(\mathbf{y}-\mathbf{w}\hat{\theta}_{\lambda_n}^{(0)})}{n}-\frac{\tau_n^2}{2}\delta_n(\hat{\theta}_{\lambda_n}^{(0)})\right],$$

where $\delta_n(\theta)=(\mathrm{sign}(\theta_1)/|\tilde{\theta}_1|^{\gamma},\ldots,\mathrm{sign}(\theta_d)/|\tilde{\theta}_d|^{\gamma})'$. The partial smoothing spline estimate or the difference based estimate (Yatchew, 1997), which are both $\sqrt{n}$ consistent, can serve as $\tilde{\theta}$ or $\hat{\theta}_{\lambda_n}^{(0)}$.

We will show that $\hat{\theta}_{\lambda_n}^{(1)}$ possesses the same *semiparametric oracle property*, whose definition is given below, as $\hat{\theta}_{\lambda_n}$. Without loss of generality, we write $\theta_0=(\theta_1',\theta_2')'$, where θ_1 consists of all q non-zero components and θ_2 consists of the rest $(d-q)$ zero elements, and define $\hat{\theta}_{\lambda_n}=(\hat{\theta}_{\lambda_n,1}'\hat{\theta}_{\lambda_n,2}')'$ accordingly. We assume that W has zero mean, strictly positive definite covariance matrix Σ and finite fourth moment. The observations z_i's (real numbers) are sorted and satisfy $\int_0^{z_i}u(w)\mathrm{d}w=(i/n)z$ for $i=1,2,\ldots,n$, where $u(\cdot)$ is a continuous and strictly positive function. The above regularity conditions are commonly used in the literature, (e.g. Craven & Wahba, 1979; Heckman, 1986). In this example, we say $\hat{\theta}_{\lambda_n}$ satisfies the *semiparametric oracle property* if

O1. $\sqrt{n}(\hat{\theta}_{\lambda_n,1}-\theta_1)\xrightarrow{d}N(0,\sigma^2\Sigma_{11}^{-1})$, where Σ_{11} is the $q\times q$ upper-left submatrix of Σ;
O2. $\hat{\theta}_{\lambda_n,2}=0$ with probability tending to one.

Note that $\sigma^2\Sigma_{11}^{-1}$ in O1 is the semiparametric efficiency bound for θ_1 due to the fixed z.

Corollary 1. *If* $n^{k/(2k+1)}\lambda_n\rightarrow\lambda_0>0$ *and* $n^{k/(2k+1)}\tau_n\rightarrow\tau_0>0$, *then* $\hat{\theta}_{\lambda_n}$ *is* $\sqrt{n}$*-consistent and satisfies the semiparametric oracle property. Given that* $\hat{\theta}_{\lambda_n}^{(0)}$ is $\sqrt{n}$*-consistent and* $\gamma=1$, *then* $\|\hat{\theta}_{\lambda_n}^{(1)}-\hat{\theta}_{\lambda_n}\|=O_P(n^{-1})$ *and* $\hat{\theta}_{\lambda_n}^{(1)}$ *also enjoys the semiparametric oracle property.*

The above corollary is a simple but interesting application of theorem 5. We can definitely relax its conditions to the general γ and non-$\sqrt{n}$ consistent $\hat{\theta}_{\lambda_n}^{(0)}$ in which we may require more than one iteration. It is also possible to extend the conclusions of corollary 1 to the semiparametric quasi-likelihood framework proposed in Mammen & van der Geer (1997).

Acknowledgements

Guang Cheng is Assistant Professor, Department of Statistics, Purdue University. I would like to thank PhD student Zhuqing Yu for her research assistance on the simulation studies. Thanks are also due to an associate editor and two referees for helpful comments that led to numerous improvements. The work was supported in part by National Science Foundation of the United States.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Proofs of lemmas A.1–A.5 and lemmas 1–2.

References

Andrews, D. (1995). Nonparametric Kernel estimation for semiparametric models. *Economet. Theory* **11**, 560–596.

Bickel, P. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647–671.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York.

Carroll, R. J., Fan, J., Gijbels, I. & Wand, M. P. (1997). Generalized partially linear single index models, *J. Am. Statist. Ass.* **92**, 477–489.

Cheng, G. (2009). Semiparametric additive isotonic regression, *J. Statist. Plann. Inference* **100**, 345–362.

Cheng, G., Yu, Z. & Huang, J. Z. (2013). The cluster bootstrap consistency in generalized estimating equations. *J. Multivar. Anal.* **115**, 33–47.

Cheng, G. & Kosorok, M. R. (2008a). Higher order semiparametric frequentist inference with the profile sampler. *Ann. Statist.* **36**, 1786–1818.

Cheng, G. & Kosorok, M. R. (2008b). General frequentist properties of the posterior profile distribution. *Ann. Statist.* **36**, 1819–1853.

Cheng, G. & Kosorok, M. R. (2009). The penalized profile sampler. *J. Multivar. Anal.* **100**, 345–362.

Cheng, G. & Zhang, H. H. (2012). Sparse and efficient estimation for partial spline models with increasing dimension. *Ann. Inst. Statist. Math.* In Revision.

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

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerishe Mathematik* **31**, 377–403.

Dominitz, J. & Sherman, R. P. (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Economet. Theory* **21**, 838–863.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–451.

Fan, J., Heckman, N. E. & Wand, W. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *JASA* **90**, 141–150, 196–216.

Heckman, N. (1986). Spline smoothing in a partly linear models. *JRSS-B* **48**, 244–248.

Hoffmann, B. (1966). *About vectors*. Prentice Hall, New Jersey.

Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *Ann. Statist.* **24**, 540–568.

Huang, J. & Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis Lecture Notes in Statistics*, (eds D. Y. Lin *et al.*), 123–169. Springer, New York.

Jiang, J., Luan, Y. & Wang, Y. G. (2007). Iterative estimating equations: linear convergence and asymptotic properties. *Ann. Statist.***35**, 2233–2260.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer, New York.

Lee, B. L., Kosorok, M. R. & Fine, J. P. (2005). The profile sampler. *JASA* **100**, 960–969.

Lipsitz, S. R., Laird, N. M. & Harrington, D. P. (1990). Using the jackknife to estimate variance of regression estimators from repeated measures studies. *Commun. Statist. – Theory Meth.* **19**, 821– 845.

Mammen, E. & van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25**, 1014–1035.

Murphy, S. A. & van der Vaart, A. W. (1999). Observed information in semiparametric models. *Bernoulli* **5**, 381–412.

Murphy, S. A. & van der Vaart, A. W. (2000). On profile likelihood. *JASA* **93**, 1461–1474.

Ortega, J. M. & Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*. Academic Press, New York.

Robinson, P. M. (1988). The stochastic difference between econometric statistics. *Econometrica* **56**, 531–548.

Roeder, K., Carroll, R. J. & Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *JASA* **91**, 722–732.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14**, 1139–1151.

Severini, T. A. & Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768–1802.

Severini, T. A. & Staniswalis, J. G. (1994). Quasi-likelihood estimatio in semiparmetric models. *J. Amer. Statist. Assoc.* **89**, 501–511.

Silverman, B. W. (1986). *Density estimation*. Chapman and Hall, London.

Speckman, P. (1988). Kernel smoothing in partial linear models. *JRSS-B* **50**, 413–436.

Staniswalis, J. G. (1989). On the kernel estimate of a regression function in likelihood based models. *JASA* **84**, 276–283.

Swann, W. H. (1972). Discrete search methods. In *Numerical methods for unconstrained optimization* (ed. W. Murray), 13–28. Academic Press, New York.

Wahba, G. (1998). *Spline models for observational data*. SIAM, Philadelphia.

Yatchew, A. (1997). An elementary estimator of the partial linear model. *Econ. Lett.* **57**, 135–143.

Received November 2011, in final form October 2012

Guang Cheng, Department of Statistics, Purdue University, 250 N. Univ. St., West Lafayette, IN 47906, USA.
E-mail: chengg@purdue.edu

Appendix

A.1. Useful lemmas

The first two lemmas are used in the proof of lemma 1. The lemmas A3, A4 and A5 are used in the proofs of theorem 3, theorem 4 and corollary 1, respectively.

Lemma A1. *Suppose that conditions M1–M4 and* (12) *hold. If $\tilde{\theta}_n$ is n^{ψ}-consistent, then we have*

$$\hat{\ell}_n(\tilde{\theta}_n, s_n) = \mathbb{P}_n\tilde{\ell}_0 + O_P\left(n^{-\psi} \vee |s_n| \vee \frac{g_r(n^{-\psi} \vee |s_n|)}{n|s_n|}\right), \tag{51}$$

$$\hat{\ell}_n(\hat{\theta}_n + U_n, s_n) = \hat{\ell}_n(\hat{\theta}_n, s_n) - \tilde{I}_0 U_n + O_P\left(\frac{g_r(|s_n| \vee \|U_n\|) \vee n^{1/2-2r}}{n|s_n|}\right), \tag{52}$$

$$\hat{I}_n(\tilde{\theta}_n, t_n) = \tilde{I}_0 + O_P\left(\frac{g_r(\|\tilde{\theta}_n - \hat{\theta}_n\| \vee |t_n|) \vee nt_n\|\tilde{\theta}_n - \hat{\theta}_n\| \vee n^{1/2-2r}}{nt_n^2}\right), \tag{53}$$

where $g_r(t) = nt^3 \vee n^{1-2r}t$ and $U_n = o_P(1)$.

Lemma A2. *Suppose that conditions M1–M4 and* (12) *hold. If*

$$\hat{I}_n(\hat{\theta}_n^{(k-1)}, t_n) - \tilde{I}_0 = O_P(r_n^{(k-1)}), \tag{54}$$

then we have

$$\|\hat{\theta}_n^{(k)}-\hat{\theta}_n\|=O_P\left(|s_n|\vee\|\hat{\theta}_n^{(k-1)}-\hat{\theta}_n\|r_n^{(k-1)}\vee\frac{g_r(|s_n|\vee\|\hat{\theta}_n^{(k-1)}-\hat{\theta}_n\|)\vee n^{1/2-2r}}{n|s_n|}\right) \tag{55}$$

for $k=1,2,\ldots.$

Lemma A3. *Suppose that condition G holds. If* $\tilde{\theta}_n$ *is a* n^{ψ}*-consistent estimator for* $0<\psi\le 1/2$, *then we have*

$$n^{-1}[\hat{S}_n^{(1)}(\tilde{\theta}_n)-\hat{S}_n^{(1)}(\theta_0)]=-\tilde{I}_0(\tilde{\theta}_n-\theta_0)+O_P((n^{-g}\vee\|\tilde{\theta}_n-\theta_0\|)\|\tilde{\theta}_n-\theta_0\|), \tag{56}$$

$$n^{-1}[\hat{S}_n^{(1)}(\tilde{\theta}_n+U_n)-\hat{S}_n^{(1)}(\tilde{\theta}_n)]=-\tilde{I}_0U_n+O_P((n^{-g}\vee\|\tilde{\theta}_n-\theta_0\|)\|U_n\|), \tag{57}$$

where U_n *is a statistic of the order* $O_P(n^{-s})$ *for some* $s\ge\psi$.

Lemma A4. *Suppose conditions K1–K2 & C1–C2 hold. Then we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}|_{\theta=\theta_0}A_{\theta,\,\eta_*(\theta)}\right)[\hat{\eta}_0-\eta_{*0}](X_i)=O_P(n^{-\delta}), \tag{58}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_{\theta_0,\eta_0}[\hat{\eta}_0^{(1)}-\eta_{*0}^{(1)}](X_i)=O_P(n^{-\delta}), \tag{59}$$

$$\frac{1}{\sqrt{n}}\dot{r}_n(\theta_0)=O_P(n^{1/2-2g}), \tag{60}$$

where $r_n(\theta)\equiv\hat{S}_n(\theta)-S_n(\theta)-\sum_{i=1}^{n}A_{\theta,\eta_*(\theta)}[\hat{\eta}(\theta)-\eta_*(\theta)]$.

Lemma A5. *Let* $\eta_0(\mathbf{z})=(\eta_0(z_1),\ldots,\eta_0(z_n))'$ *and* $\epsilon=(\epsilon_1,\ldots,\epsilon_n)'$. *If* $\lambda_n\to 0$, *then we have*

$$\mathbf{w}'A(\lambda_n)\epsilon=O_P(\lambda_n^{-1/(2k)}), \tag{61}$$

$$\mathbf{w}'[I-A(\lambda_n)]\eta_0(\mathbf{z})=O_P(n^{1/2}\lambda_n), \tag{62}$$

$$\mathbf{w}'(I-A(\lambda_n))\mathbf{w}/n=\Sigma+O_P(n^{-1/2}\vee n^{-1}\lambda_n^{-1/k}). \tag{63}$$

A.2. Proof of theorem 1

Define $\mathcal{N}_n=\{\theta:\|\theta-\theta_0\|\le Mn^{-\psi}\}$ and $\mathcal{N}_n^c$ as its complement for any $0<M<\infty$. Note that $\mathcal{D}_n\cap\mathcal{N}_n\neq\emptyset$ for large enough M and $\mathcal{D}_n\cap\mathcal{N}_n^c\neq\emptyset$ for large enough n. We first consider (10). For sufficiently large M and any $C_1>0$, we have

$$\begin{aligned}
P(\theta_n^D\in\mathcal{N}_n^c)&=P\left(\theta_n^D\in\mathcal{N}_n^c\text{ and }\theta_{iD}\in\mathcal{N}_n\text{ for some }i\right)\\
&\le P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n}\hat{S}_n(\theta)\le\max_{\mathcal{D}_n\cap\mathcal{N}_n^c}\hat{S}_n(\theta)\right)\\
&\le P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n}\hat{S}_n(\theta)<\hat{S}_n(\theta_0)-C_1n^{1-2\psi}\right)\\
&\quad+P\left(\left\{\max_{\mathcal{D}_n\cap\mathcal{N}_n}\hat{S}_n(\theta)\le\max_{\mathcal{D}_n\cap\mathcal{N}_n^c}\hat{S}_n(\theta)\right\}\cap\left\{\max_{\mathcal{D}_n\cap\mathcal{N}_n}\hat{S}_n(\theta)\ge\hat{S}_n(\theta_0)-C_1n^{1-2\psi}\right\}\right)\\
&\le P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n}n^{-1/2}(\hat{S}_n(\theta)-\hat{S}_n(\theta_0))<-C_1n^{1/2-2\psi}\cap\{\theta_n^o\text{ is consistent}\}\right)
\end{aligned}$$

$$
\begin{aligned}
&\quad + P\left(\max_{\mathcal{N}_n^c} n^{-1/2}(\hat{S}_n(\theta) - \hat{S}_n(\theta_0)) \geq -C_1 n^{1/2-2\psi}\right) \\
&\quad + P\left(\theta_n^o \text{ is inconsistent}\right) \\
&\leq I + II + III,
\end{aligned}
$$

where $\theta_n^o = \arg\max_{\mathcal{D}_n \cap \mathcal{N}_n} \hat{S}_n(\theta)$.

The definition of $\mathcal{N}_n$ implies $III \to 0$ for any M as $n \to \infty$. We next analyse the term I as follows. In view of (9) and the definition of $\mathcal{N}_n$, we have that

$$
\begin{aligned}
I &= P\left(\sqrt{n}(\theta_n^o - \theta_0)' \mathbb{P}_n \tilde{\ell}_0 - \frac{\sqrt{n}}{2}(\theta_n^o - \theta_0)' \tilde{I}_0 (\theta_n^o - \theta_0) + n^{-1/2} \Delta_n(\theta_n^o) < -C_1 n^{1/2-2\psi}\right) \\
&\leq P\left(\|\sqrt{n}\mathbb{P}_n \tilde{\ell}_0\| \|\theta_n^o - \theta_0\| + (\delta_{\max}\sqrt{n}/2)\|\theta_n^o - \theta_0\|^2 + \|n^{-1/2}\Delta_n(\theta_n^o)\| > C_1 n^{1/2-2\psi}\right) \\
&\leq P\left(\|\sqrt{n}\mathbb{P}_n \tilde{\ell}_0\| > \frac{C_1 - \delta_{\max} M^2/2}{M} n^{1/2-\psi} + o_P(n^{1/2-\psi})\right) \\
&\leq \bar{I},
\end{aligned}
$$

where $\delta_{\max}$ is the largest eigenvalue of $\tilde{I}_0$, and the second inequality follows from the definitions of $\mathcal{N}_n$ and Δ_n, and the range that $2v > 1/2 \geq \psi > 0$. Denote $\theta_n^* = \arg\max_{\mathcal{N}_n^c} \hat{S}_n(\theta)$. We will show $II \to 0$ by first decomposing it as $II_1 + II_2$, where

$$
\begin{aligned}
II_1 &= P\left(n^{-1/2}(\hat{S}_n(\theta_n^*) - \hat{S}_n(\theta_0)) \geq -C_1 n^{1/2-2\psi} \cap \{\theta_n^* \text{ is consistent }\}\right), \\
II_2 &= P\left((\hat{S}_n(\theta_n^*) - \hat{S}_n(\theta_0)) \geq -C_1 n^{1-2\psi} \cap \{\theta_n^* \text{ is inconsistent }\}\right).
\end{aligned}
$$

Note that we can write $n^{-1/2}\Delta_n(\theta_n^*)$ as $\sqrt{n}\|\theta_n^* - \theta_0\|^2 \epsilon_{1n} + \sqrt{n}\|\theta_n^* - \theta_0\|\epsilon_{2n}$, where $\epsilon_{1n} = o_P(1)$ and $\epsilon_{2n} = o_P(n^{-1/2})$, in the event that $\{\theta_n^*$ is consistent$\}$. Thus, according to (9), we can write II_1 as

$$
\begin{aligned}
&P\left((\theta_n^* - \theta_0)'\sqrt{n}\mathbb{P}_n \tilde{\ell}_0 + \sqrt{n}\|\theta_n^* - \theta_0\|\epsilon_{2n} \geq \frac{\sqrt{n}}{2}(\theta_n^* - \theta_0)' \tilde{I}_0 (\theta_n^* - \theta_0) - \sqrt{n}\|\theta_n^* - \theta_0\|^2 \epsilon_{1n} - C_1 n^{1/2-2\psi}\right) \\
&\quad \leq P\left(\|\theta_n^* - \theta_0\| \left[\|\sqrt{n}\mathbb{P}_n \tilde{\ell}_0\| + \sqrt{n}\epsilon_{2n}\right] \geq \frac{\sqrt{n}}{2}\|\theta_n^* - \theta_0\|^2 \delta_{\min} - \sqrt{n}\|\theta_n^* - \theta_0\|^2 \epsilon_{1n} - C_1 n^{1/2-2\psi}\right) \\
&\quad \leq P\left(\left[\|\sqrt{n}\mathbb{P}_n \tilde{\ell}_0\| + \sqrt{n}\epsilon_{2n}\right] \geq \sqrt{n}\|\theta_n^* - \theta_0\|(\delta_{\min}/2 - \epsilon_{1n}) - \frac{C_1 n^{1/2-\psi}}{K}\right) \\
&\quad \leq P\left(\left[\|\sqrt{n}\mathbb{P}_n \tilde{\ell}_0\| + \sqrt{n}\epsilon_{2n}\right] \geq \frac{\delta_{\min} K^2/2 - C_1}{K} n^{1/2-\psi} + o_P(n^{1/2-\psi})\right) \\
&\quad \leq \bar{II}_1,
\end{aligned}
$$

where $\delta_{\min} > 0$ is the smalest eigenvalue of $\tilde{I}_0$. All the above inequalities follow from the fact that $\|\theta_n^* - \theta_0\| \geq K n^{-\psi}$ for some $K > M$ and $\epsilon_{1n} = o_P(1)$. The term II_2 is shown to converge to zero by the following contradiction arguments. By assuming that the event $\{(\hat{S}_n(\theta_n^D) - \hat{S}_n(\theta_0)) \geq -C_1 n^{1-2\psi}\}$ holds, we have $|\hat{S}_n(\theta_n^D) - \hat{S}_n(\hat{\theta}_n)| = \hat{S}_n(\hat{\theta}_n) - \hat{S}_n(\theta_n^D) \leq \hat{S}_n(\hat{\theta}_n) - \hat{S}_n(\theta_0) + C_1 n^{1-2\psi}$. Note that (9) and the consistency of $\hat{\theta}_n$ imply $\hat{S}_n(\theta_0) - \hat{S}_n(\hat{\theta}_n) = o_P(n)$. Then, we can show that $|\hat{S}_n(\theta_n^D) - \hat{S}_n(\hat{\theta}_n)|/n = o_P(1)$ which implies that θ_n^D is consistent by (8). This implication contradicts with another event in II_2, that is, $\{\theta_n^D$ is inconsistent$\}$. Therefore we can claim that $II_2 \to 0$.

In view of the above discussions, it remains to show that $\bar{I}$ and $\bar{II}_1$ converge to zero. Note that $\|\sqrt{n}\mathbb{P}_n \tilde{\ell}_0\|$ in $\bar{I}$ is $O_P(1)$, and so is $(\|\sqrt{n}\mathbb{P}_n \tilde{\ell}_0\| + \sqrt{n}\epsilon_{2n})$ in $\bar{II}_1$. Therefore, by choosing sufficiently large C_1 and $K > M$, meanwhile keeping the inequality $\delta_{\max} M^2 < 2C_1 < \delta_{\min} K^2$ valid, we show that $\bar{I}$ and $\bar{II}_1$ can be arbitrarily close to zero. For example, we can take

$K = M + B$ and $C_1 = (\delta_{\max} M^2 + \delta_{\min}(M + B)^2)/4$ for some fixed $B > 0$ and sufficiently large M. This completes the proof of (10).

Our proof of (11) is similar as that of (10). Denote θ_{iS} as an element in $\mathcal{S}_n$. Similarly, we have

$$\begin{aligned}
P(\theta_n^S \in \mathcal{N}_n^c) \le & E\left\{P\left(\theta_n^S \in \mathcal{N}_n^c \text{ and } \theta_{iS} \in \mathcal{N}_n \text{ for some } i \,|\, \mathcal{S}_n\right)\right\} + E\left\{P\left(\theta_{iS} \in \mathcal{N}_n^c \text{ for all } i \,|\, \mathcal{S}_n\right)\right\} \\
\le & P\left(\max_{\mathcal{S}_n \cap \mathcal{N}_n} \hat{S}_n(\theta) \le \max_{\mathcal{S}_n \cap \mathcal{N}_n^c} \hat{S}_n(\theta)\right) + P\left(\theta_{iS} \in \mathcal{N}_n^c \text{ for all } i\right) \\
\le & P\left(\max_{\mathcal{S}_n \cap \mathcal{N}_n} n^{-1/2}(\hat{S}_n(\theta) - \hat{S}_n(\theta_0)) < -C_2 n^{1/2-2\psi}\right) + P\left(\theta_{iS} \in \mathcal{N}_n^c \text{ for all } i\right) \\
& + P\left(\max_{\mathcal{S}_n \cap \mathcal{N}_n^c} n^{-1/2}(\hat{S}_n(\theta) - \hat{S}_n(\theta_0)) \ge -C_2 n^{1/2-2\psi}\right) \\
\le & P\left(\max_{\mathcal{S}_n \cap \mathcal{N}_n} n^{-1/2}(\hat{S}_n(\theta) - \hat{S}_n(\theta_0)) < -C_2 n^{1/2-2\psi} \cap \{\theta_n^\dagger \text{ is consistent}\}\right) \\
& + P\left(\max_{\mathcal{N}_n^c} n^{-1/2}(\hat{S}_n(\theta) - \hat{S}_n(\theta_0)) \ge -C_2 n^{1/2-2\psi}\right) \\
& + P(\theta_n^\dagger \text{ is inconsistent}) + P\left(\theta_{iS} \in \mathcal{N}_n^c \text{ for all } i\right) \\
\le & I' + II' + III' + IV',
\end{aligned}$$

where C_2 is an arbitrary positive constant and $\theta_n^\dagger = \arg\max_{\mathcal{S}_n \cap \mathcal{N}_n} \hat{S}_n(\theta)$.

We first consider the terms III' and IV'. Since $\theta_n^\dagger \in \mathcal{N}_n$, we have $III' \to 0$ for any M as $n \to \infty$. The term IV' is computed as

$$(1 - P(\bar{\theta} \in \mathcal{N}_n))^{\mathrm{card}(\mathcal{S}_n)}. \tag{64}$$

Since the density of $\bar{\theta}$ is assumed to be bounded away from zero around θ_0 and $card(\mathcal{S}_n) \ge \tilde{C} n^{\psi}$, (64) is bounded above by

$$(1 - \rho n^{-\psi} M)^{\mathrm{card}(\mathcal{S}_\mathrm{n})} \le (1 - \rho M \tilde{C}/\mathrm{card}(\mathcal{S}_n))^{\mathrm{card}(\mathcal{S}_\mathrm{n})} \longrightarrow \exp(-\rho M \tilde{C}), \tag{65}$$

for some $\rho > 0$.

We next consider I'. According to (9), we can show

$$\begin{aligned}
& n^{-1/2}(\hat{S}_n(\theta_n^\dagger) - \hat{S}_n(\theta_0)) \\
& \quad \ge \max_{\mathcal{S}_n \cap \mathcal{N}_n} \left\{-\frac{\sqrt{n}}{2}(\theta - \theta_0)' \tilde{I}_0 (\theta - \theta_0)\right\} - \max_{\mathcal{S}_n \cap \mathcal{N}_n} \{-\sqrt{n}(\theta - \theta_0)' \mathbb{P}_n \tilde{\ell}_0 - \Delta_n(\theta)/\sqrt{n}\} \\
& \quad \ge -\min_{\mathcal{S}_n \cap \mathcal{N}_n} \left\{\frac{\sqrt{n}}{2}(\theta - \theta_0)' \tilde{I}_0 (\theta - \theta_0)\right\} - \max_{\mathcal{S}_n \cap \mathcal{N}_n} \{-\sqrt{n}(\theta - \theta_0)' \mathbb{P}_n \tilde{\ell}_0 - \Delta_n(\theta)/\sqrt{n}\}.
\end{aligned}$$

Therefore, we can bound I' by $I_1' + I_2'$, where

$$\begin{aligned}
I_1' &= P\left(\max_{\mathcal{S}_n \cap \mathcal{N}_n} \{-\sqrt{n}(\theta - \theta_0)' \mathbb{P}_n \tilde{\ell}_0 - n^{-1/2} \Delta_n(\theta)\} > (C_2/2) n^{1/2-2\psi}\right), \\
I_2' &= P\left(\min_{\mathcal{S}_n \cap \mathcal{N}_n} \{\sqrt{n}(\theta - \theta_0)' \tilde{I}_0 (\theta - \theta_0)\} > C_2 n^{1/2-2\psi}\right).
\end{aligned}$$

Given sufficiently large C_2/M, I_1' can be arbitrarily close to zero since

$$\begin{aligned}
I_1' &\le P\left(\|\sqrt{n} \mathbb{P}_n \tilde{\ell}_0\| > \frac{C_2}{2M} n^{1/2-\psi} + O_P(n^{1/2-2\psi} \vee n^{1/2-2v})\right) \\
&\le P\left(\|\sqrt{n} \mathbb{P}_n \tilde{\ell}_0\| > \frac{C_2}{2M} n^{1/2-\psi} + o_P(n^{1/2-\psi})\right),
\end{aligned} \tag{66}$$

where the last inequality follows from the assumption that $2v > 1/2 \geq \psi$. Since $\min_{\mathcal{N}_n^c}\{\sqrt{n}(\theta - \theta_0)'\tilde{I}_0(\theta - \theta_0)\} > C_2 n^{1/2-2\psi}$ by choosing $\delta_{\min} M^2 > C_2$, I_2' is bounded above by

$$
\begin{aligned}
&P\left(\min_{\mathcal{S}_n}\{\sqrt{n}(\theta - \theta_0)'\tilde{I}_0(\theta - \theta_0)\} > C_2 n^{1/2-2\psi}\right)\\
&\quad \leq \left[P(\sqrt{n}(\bar{\theta} - \theta_0)'\tilde{I}_0(\bar{\theta} - \theta_0) > C_2 n^{1/2-2\psi})\right]^{\mathrm{card}(\mathcal{S}_n)}\\
&\quad \leq \left[1 - P(\|\bar{\theta} - \theta_0\| \leq (C_2/\delta_{\max})^{1/2} n^{-\psi})\right]^{\mathrm{card}(\mathcal{S}_n)}\\
&\quad \leq \left[1 - P\left(\|\bar{\theta} - \theta_0\| \leq (C_2/\delta_{\max})^{1/2}\tilde{C}/\mathrm{card}(\mathcal{S}_n)\right)\right]^{\mathrm{card}(\mathcal{S}_n)}\\
&\quad \leq (1 - \rho\tilde{C}(C_2/\delta_{\max})^{1/2}/\mathrm{card}(\mathcal{S}_n))^{\mathrm{card}(\mathcal{S}_n)} \longrightarrow \exp(-\rho\tilde{C}\sqrt{C_2/\delta_{\max}})
\end{aligned}
\tag{67}
$$

for some $\rho > 0$. In the above, the third and fourth inequality follows from the assumptions that $\mathrm{card}(\mathcal{S}_n) \geq \tilde{C} n^{\psi}$ and the density for $\bar{\theta}$ is bounded away from zero around θ_0, respectively. By assuming that $2C_2 < K^2\delta_{\min}$ for some $K > M$, we can prove that $II' \to 0$ in the same manner as we show $II \to 0$.

Let $L = \min\{K^2/2, M^2\}$. In view of (65), (66), (67) and the above discussions on II', by choosing sufficiently large C_2, $K > M$ and C_2/M, meanwhile keeping the inequality $C_2 < L\delta_{\min}$ valid, we can make $P(\theta_n^S \in \mathcal{N}_n^c)$ arbitrarily small. For example, we can take $C_2 = M^{3/2}\delta_{\min}$ and $K = M + B$, for some fixed $B > 0$ and sufficiently large M. This completes the whole proof.

A.3. Proof of theorem 2

According to the proof in lemma 1, we also need to consider the stochastic order of $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\|$ in terms of three stages: (i) $r_{k-1} < r$, (ii) $r \leq r_{k-1} < 1/2$, and (iii) $r_{k-1} \geq 1/2$. In stage (i), we have $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{3/2}) = O_P(n^{-S_1(\psi,k)})$ if $k \leq K_1(\psi, r)$. In stage (ii), we have $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{1/2} n^{-r})$, which implies that $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(n^{-S_2(\psi,\, r,\, k)})$ if $r \leq \psi < 1/2$. It is easy to show that $S_2(\psi, r, k) \geq 1/2$ if $k \geq K_2(\psi, r, 1/2)$. In stage (iii), we obtain the the smallest order of $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\|$, that is, $O_P(n^{-r-1/4})$. Combining the above analysis of (i)–(iii), we can conclude that the stochastic order of $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\|$ is continuously improving till the optimal bound $O_P(n^{-r-1/4})$ and can be expressed as $O_P(n^{-S(\psi,r,k)})$. Equation (18) also follows from the above analysis.

A.4. Proof of theorem 3

We first show (27) by applying lemma A3. In (56), we replace $\tilde{\theta}_n$ by $\hat{\theta}_n$. Since $\hat{\theta}_n$ is assumed to be consistent and θ_0 is an interior point of Θ, we have $\hat{S}_n^{(1)}(\hat{\theta}_n) = 0$. By (4) and (22), we have

$$
\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}\tilde{I}_0^{-1}\mathbb{P}_n\tilde{\ell}_0 + O_P(n^{1/2-2g} \vee n^{1/2}\|\hat{\theta}_n - \theta_0\|^2)
\tag{68}
$$

given that $\hat{\theta}_n$ is consistent and $\tilde{I}_0$ is non-singular. Considering the range of g, that is, $1/4 < g \leq 1/2$, we can show $\hat{\theta}_n$ is actually $\sqrt{n}$-consistent, and thus simplify (68) to (27).

We next show (28). By (19), we can write $\sqrt{n}\hat{I}_n(\hat{\theta}_n^{(0)})(\hat{\theta}_n^{(1)} - \hat{\theta}_n)$ as

$$
\begin{aligned}
&\sqrt{n}\,\hat{I}_n(\hat{\theta}_n^{(0)})(\hat{\theta}_n^{(0)} - \hat{\theta}_n) + n^{1/2}(\hat{\ell}_n(\hat{\theta}_n^{(0)}) - \hat{\ell}_n(\hat{\theta}_n))\\
&\quad = \sqrt{n}\,\hat{I}_n(\hat{\theta}_n^{(0)})(\hat{\theta}_n^{(0)} - \hat{\theta}_n) + n^{-1/2}\hat{S}_n^{(2)}(\hat{\theta}_n^{(0)})(\hat{\theta}_n^{(0)} - \hat{\theta}_n) + O_P(\sqrt{n}\|\hat{\theta}_n - \hat{\theta}_n^{(0)}\|^2)\\
&\quad = O_P(\sqrt{n}\|\hat{\theta}_n - \hat{\theta}_n^{(0)}\|^2)
\end{aligned}
$$

under condition G. Further, by (23) and (24), we have the invertibility of $\hat{I}_n(\hat{\theta}_n^{(0)})$ based on that of $\tilde{I}_0$. This implies $\hat{\theta}_n^{(1)} - \hat{\theta}_n = O_P(\|\hat{\theta}_n^{(0)} - \hat{\theta}_n\|^2)$. By the induction principal, we can thus show

$$\hat{\theta}_n^{(k)} - \hat{\theta}_n = O_P(\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^2) \quad \text{for any } k \geq 1. \tag{69}$$

Equation (28) follows from (69) trivially.

To show (29), we first prove

$$\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P\left(n^{1/2-g}\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^2 \vee n^{-g}\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|\right). \tag{70}$$

By replacing $\tilde{\theta}_n$ and U_n with $\hat{\theta}_n$ and $(\hat{\theta}_n^{(k-1)} - \hat{\theta}_n)$ in (57), respectively, we establish that

$$n^{-1/2}[\hat{S}_n^{(1)}(\hat{\theta}_n^{(k-1)}) - \hat{S}_n^{(1)}(\hat{\theta}_n)] = -\sqrt{n}\tilde{I}_0(\hat{\theta}_n^{(k-1)} - \hat{\theta}_n) + O_P(n^{1/2-g}\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|). \tag{71}$$

Similarly, by setting $\tilde{\theta}_n$ as $\hat{\theta}_n$, and then setting U_n as $(\hat{\theta}_n^{(k-1)} - \hat{\theta}_n + n^{-1/2}t_1v_j)$ and $(\hat{\theta}_n^{(k-1)} - \hat{\theta}_n + n^{-1/2}t_2v_j)$ in (57), respectively, we have that

$$[\hat{I}_n(\hat{\theta}_n^{(k-1)})]_{ij} = [\tilde{I}_0]_{ij} + O_P(n^{1/2-g}\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\| \vee n^{-g}) \tag{72}$$

when $\hat{I}_n^{(k-1)}$ is defined in (21). Following similar logic in analyzing (69), we can obtain (70) by considering (71)–(72). Next we will show that (70) implies (29) by the following analysis. Based on (70) we have

$$\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = \begin{cases} O_P(n^{-g}\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|) & \text{if } \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\| = O_P(n^{-1/2}), \\ O_P(n^{1/2-g}\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^2) & \text{if } \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{-1} = O_P(n^{1/2}). \end{cases} \tag{73}$$

It is easy to show that $\|\hat{\theta}_n^{(L_1(\psi,g))} - \hat{\theta}_n\| = O_P(n^{-1/2})$ and $\|\hat{\theta}_n^{(L_1(\psi,g)-1)} - \hat{\theta}_n\|^{-1} = O_P(n^{1/2})$. In other words, if $k \leq L_1(\psi, g)$, then we have the relation that $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(n^{1/2-g}\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^2)$ based on (73). This implies the form of $R_1(\psi, g, k)$ in (26). Note that $R_1(\psi, g, k)$ is an increasing function of k under the condition that $\psi + g > 1/2$. After $L_1(\psi, g)$ iterations, we have

$$\|\hat{\theta}_n^{(L_1(\psi,g))} - \hat{\theta}_n\| = O_P(n^{-R_1(\psi,g,L_1(\psi,g))}) = O_P(n^{-1/2}). \tag{74}$$

Thus, we have the relation that $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\| = O_P(n^{-g}\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|)$ for $k \geq (L_1(\psi, g) + 1)$ based on (73). Combining this relation with (74), we can show the form of $R_2(\psi, g, k)$ when $k > L_1(\psi, g)$. Since $R(\psi, g, k)$ is an increasing function of k given that $1/2 - g < \psi \leq 1/2$, the stochastic order of $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\|$ is continuously decreasing as $k \to \infty$. The calculation of k^* also follows from the above analysis.

A.5. Proof of theorem 4

We first consider (22) by rewriting its LHS as

$$\frac{1}{n}\frac{\partial}{\partial\theta}|_{\theta=\theta_0}\left[\sum_{i=1}^{n} A_{\theta,\eta_*(\theta)}[\hat{\eta}(\theta) - \eta_*(\theta)](X_i) + r_n(\theta)\right],$$

where $r_n(\theta)$ is defined in lemma A4. Therefore, we have

$$\begin{aligned} n^{-1}[\hat{S}_n^{(1)}(\theta_0) - S_n^{(1)}(\theta_0)] &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}|_{\theta=\theta_0}A_{\theta,\eta_*(\theta)}\right)(\hat{\eta}_0 - \eta_{*0}) + \frac{1}{n}\sum_{i=1}^{n}A_{\theta_0,\eta_0}(\hat{\eta}_0^{(1)} - \eta_{*0}^{(1)}) + \frac{1}{n}\dot{r}_n(\theta_0) \\ &= O_P(n^{-2g}) \end{aligned}$$

by lemma A4 and the condition that $\delta \geq (2g-1/2)$. As discussed previously, we will show (23) together with (25). By Taylor expansion, we have

$$\begin{aligned}\hat{S}_n(\theta)-S_n(\theta) &= \sum_{i=1}^{n}\int_0^1 \frac{\partial \log\, lik(X_i;\theta,\eta_t(Z_i;\theta)x)}{\partial\lambda}\, \mathrm{d}t[\hat{\eta}(Z_i;\theta)-\eta_*(Z_i;\theta)] \\ &\equiv \sum_{i=1}^{n} R_\theta(X_i)[\hat{\eta}(Z_i;\theta); -\eta_*(Z_i;\theta)],\end{aligned}$$

where $\eta_t(\theta)=\eta_*(\theta)+t(\hat{\eta}(\theta)-\eta_*(\theta))$. Hence, to prove (23) and (25), it suffices to show that

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\sup_{z\in\mathcal{Z}}\left| n^{-1}\sum_{i=1}^{n}\frac{\partial^j}{\partial\theta^j}R_\theta(X_i)\right| = O_P(1) \text{ for } j=0,1,2,3 \tag{75}$$

in view of (34) and (35). Considering the smoothness Condition K2, we can prove (75) using the same approach as in the proof of (S.4) in the supplementary materials.

In the end, it remains to show that the class of functions $\{(\partial^3/\partial\theta^3)\log\, lik(x;\theta,\eta_*(\theta)) : \theta\in\mathcal{N}(\theta_0)\}$ is P-Glivenko-Cantelli and that

$$\sup_{\theta\in\mathcal{N}(\theta_0)} E\left|(\partial^3/\partial\theta^3)\log\, lik(X;\theta,\eta_*(\theta))\right| < \infty. \tag{76}$$

Let $\ell^{(3)}(\theta,\eta(\theta))=(\partial^3/\partial\theta^3)\log\, lik(x;\theta,\eta_*(\theta))$. For any $\theta_1,\theta_2\in\mathcal{N}(\theta_0)$, we have

$$\begin{aligned}&|\ell^{(3)}(\theta_1,\eta_*(\theta_1))-\ell^{(3)}(\theta_2,\eta_*(\theta_2))| \\ &\quad\leq \sup_{\theta,\lambda}\left|\frac{\partial\ell^{(3)}}{\partial\theta}(\theta,\lambda)\right| \|\theta_1-\theta_2\| + \sup_{\theta,\lambda}\left|\frac{\partial\ell^{(3)}}{\partial\lambda}(\theta,\lambda)\right| \|\eta_*(\theta_1)-\eta_*(\theta_2)\|_\infty \\ &\quad\leq \sup_{\theta,\lambda}\left|\frac{\partial\ell^{(3)}}{\partial\theta}(\theta,\lambda)\right| \|\theta_1-\theta_2\| + \sup_{\theta,\lambda}\left|\frac{\partial\ell^{(3)}}{\partial\lambda}(\theta,\lambda)\right| \sup_{\theta\in\mathcal{N}(\theta_0)}\|\eta_*^{(1)}(\theta)\|_\infty \times \|\theta_1-\theta_2\| \\ &\quad\leq A\|\theta_1-\theta_2\|.\end{aligned}$$

By condition K2 and $\sup_{\theta\in\mathcal{N}(\theta_0)}\|\eta_*^{(1)}(\theta)\|_\infty < \infty$ in condition C1, we know that $EA^2<\infty$. Thus, by the P-G-C preservation theorem 9.23 of Kosorok (2008) and compactness of $\mathcal{N}(\theta_0)$, we know that $\{(\partial^3/\partial\theta^3)\log\, lik(x;\theta,\eta_*(\theta)) : \theta\in\mathcal{N}(\theta_0)\}$ is P-Glivenko-Cantelli. The last condition (76) follows from the conditions K2 and C1 by some algebra.

A.6. Proof of corollary 1

For the $\sqrt{n}$ consistency of $\hat{\theta}_{\lambda_n}$, it suffices to show that, for any given $\epsilon>0$, there exists a large constant M such that

$$P\left\{\inf_{\|s\|=M}\Delta_n(s)>0\right\}\geq 1-\epsilon, \tag{77}$$

where $\Delta_n(s)\equiv[\hat{S}_{\lambda_n}(\theta_0+n^{-1/2}s)-\hat{S}_{\lambda_n}(\theta_0)]$. According to (49), we have

$$\Delta_n(s)\geq\tilde{S}_{\lambda_n}(\theta_0+n^{-1/2}s)-\tilde{S}_{\lambda_n}(\theta_0)+n\tau_n^2\sum_{j=1}^{q}\frac{|\theta_{0j}+n^{-1/2}s_j|-|\theta_{0j}|}{|\tilde{\theta}_j|},$$

where s_j is the jth element of s. The Taylor expansion further gives

$$\Delta_n(s)\geq n^{-1/2}s'\tilde{S}^{(1)}_{\lambda_n}(\theta_0)+\frac{1}{2}s'[\tilde{S}^{(2)}_{\lambda_n}(\theta_0)/n]s+n\tau_n^2\sum_{j=1}^{q}\frac{|\theta_{0j}+n^{-1/2}s_j|-|\theta_{0j}|}{|\tilde{\theta}_j|}, \tag{78}$$

where $\tilde{S}^{(j)}_{\lambda_n}(\theta_0)$ represents the jth derivative of $\tilde{S}_{\lambda_n}(\theta)$ at θ_0. Based on (50), we have

$$\tilde{S}^{(1)}_{\lambda_n}(\theta_0) = -2\mathbf{w}'[I - A(\lambda_n)](\mathbf{y} - \mathbf{w}\theta_0), \tag{79}$$

$$\tilde{S}^{(2)}_{\lambda_n}(\theta_0) = 2\mathbf{w}'[I - A(\lambda_n)]\mathbf{w}. \tag{80}$$

Lemma A5 implies that

$$\tilde{S}^{(1)}_{\lambda_n}(\theta_0) = O_P(n^{1/2}), \tag{81}$$

$$\tilde{S}^{(2)}_{\lambda_n}(\theta_0) = O_P(n) \tag{82}$$

since λ_n is required to converge to zero. Hence, we know the first two terms in the right-hand side of (78) have the same order, that is, $O_P(1)$. And the second term, which converges to some positive constant, dominates the first one by choosing sufficiently large M. The third term is bounded by $n^{1/2}\tau_n^2 M_0$ for some positive constant M_0 since $\tilde{\beta}_j$ is the consistent estimate for the non-zero coefficient. Considering that $\sqrt{n}\tau_n^2 \to 0$, we have shown the $\sqrt{n}$-consistency of $\hat{\theta}_{\lambda_n}$.

To complete the proof of other parts, we first need to show

$$\|\hat{\theta}^{(1)}_{\lambda_n} - \hat{\theta}_{\lambda_n}\| = O_P(n^{-1}) \tag{83}$$

based on theorem 5. And then we will verify condition G' for the case $c = -2$. It is easy to show that $\mathbb{P}_n\tilde{\ell}_0 = \mathbf{w}'\epsilon/n$ and $\tilde{I}_0 = \Sigma$ in this example. To verify (43), we have

$$\begin{aligned}\frac{1}{n}\hat{S}^{(1)}_{\lambda_n}(\theta_0) + 2\mathbb{P}_n\tilde{\ell}_0 &= -\frac{2}{n}\mathbf{w}'(I - A(\lambda_n))\eta_0(\mathbf{z}) + \frac{2}{n}\mathbf{w}'A(\lambda_n)\epsilon + \tau_n^2\delta_n(\theta_0)\\ &= O_P(n^{-1/2}\lambda_n \vee n^{-1}\lambda_n^{-1/(2k)} \vee \tau_n^2),\end{aligned}$$

where the second equality follows from lemma A5 and the fact that $\delta_n(\theta_0) = O_P(1)$. Considering the conditions on τ_n and λ_n, we have proved (43). Equation (44) follows from (80) and (63), and (45) trivially holds. Having shown the consistency of $\hat{\theta}_{\lambda_n}$ and verified G', we are able to show (83).

For any sequence of estimate θ_n, the below arguments show that $\theta_n = 0$ with probability tending to one if it is $\sqrt{n}$-consistent. For any $\sqrt{n}$-consistent estimator, it suffices to show that

$$\hat{S}_{\lambda_n}(\bar{\theta}_1, 0) = \min_{\|\bar{\theta}_2\| \le Cn^{-1/2}} \hat{S}_{\lambda_n}(\bar{\theta}_1, \bar{\theta}_2) \tag{84}$$

for any $\bar{\theta}_1$ satisfying $\|\bar{\theta}_1 - \theta_1\| = O_P(n^{-1/2})$ with probability approaching to 1. In order to show (84), we need to show that $\partial\hat{S}_{\lambda_n}(\theta)/\partial\theta_j < 0$ for $\theta_j \in (-Cn^{-1/2}, 0)$ and $\partial\hat{S}_{\lambda_n}(\theta)/\partial\theta_j > 0$ for $\theta_j \in (0, Cn^{-1/2})$ holds when $j = q+1, \ldots, d$ with probability tending to 1. By two-term Taylor expansion of $\tilde{S}_{\lambda_n}(\theta)$ at θ_0, $\partial\hat{S}_{\lambda_n}(\theta)/\partial\theta_j$ can be expressed in the following form:

$$\frac{\partial\hat{S}_{\lambda_n}(\theta)}{\partial\theta_j} = \frac{\partial\tilde{S}_{\lambda_n}(\theta_0)}{\partial\theta_j} + \sum_{k=1}^{d}\frac{\partial^2\tilde{S}_{\lambda_n}(\theta_0)}{\partial\theta_j\partial\theta_k}(\theta_k - \theta_{0k}) + n\tau_n^2\frac{1 \times \mathrm{sign}(\theta_j)}{|\tilde{\theta}_j|},$$

for $j = q+1, \ldots, d$. Note that $\|\bar{\theta} - \theta_0\| = O_P(n^{-1/2})$ by the above construction. Hence, we have

$$\frac{\partial\hat{S}_{\lambda_n}(\theta)}{\partial\theta_j} = O_P(n^{1/2}) + \mathrm{sign}(\theta_j)\frac{n\tau_n^2}{|\tilde{\theta}_j|}$$

by (81) and (82). We assume that $n^{k/(2k+1)}\tau_n \to \tau_0 > 0$ which implies that $\sqrt{n}\tau_n^2/|\tilde{\theta}_j| \to \infty$ for $\sqrt{n}$ consistent $\tilde{\theta}_j$ and $j = q+1, \ldots, d$. Thus, we show that the sign of θ_j determines that of $\partial\hat{S}_{\lambda_n}(\theta)/\partial\theta_j$. The above arguments apply to $\hat{\theta}_{\lambda_n,2}$ and $\hat{\theta}^{(1)}_{\lambda_n,2}$ since both of them are proven to be $\sqrt{n}$ consistent in view of the previous discussions, that is, (83).

Now it remains to show the semiparametric efficiency of $\hat{\theta}_{\lambda_n,1}$, which immediately implies that of $\hat{\theta}^{(1)}_{\lambda_n,1}$ based on (83). Since we have shown $\hat{\theta}_{\lambda_n,2}=0$, we can establish that

$$\frac{\partial \hat{S}_{\lambda_n}(\theta)}{\partial \theta_j}|_{\theta=(\hat{\theta}_{\lambda_n,1},0)}=0 \quad \text{for any } j=1,\ldots,q \tag{85}$$

with probability tending to one. Let $\mathbf{w}_1$ denote the first q columns of $\mathbf{w}$. Applying Taylor expansion to (85) around θ_0, we obtain

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{\lambda_n,1}-\theta_1) &= \sqrt{n}\left\{\frac{1}{n}\mathbf{w}_1'[I-A(\lambda_n)]\mathbf{w}_1\right\}^{-1}\frac{1}{n}\mathbf{w}_1'[I-A(\lambda_n)](\eta_0(\mathbf{z})+\epsilon)+O_P(\sqrt{n}\tau_n^2)\\ &= \left\{\Sigma_{11}+O_P(n^{-1/2}\vee n^{-1}\lambda_n^{-1/k})\right\}^{-1}\frac{1}{\sqrt{n}}\mathbf{w}_1'\epsilon+O_P(\sqrt{n}\tau_n^2\vee n^{-1/2}\lambda_n^{-1/(2k)}\vee\lambda_n)\\ &= \frac{1}{\sqrt{n}}\Sigma_{11}^{-1}\sum_{i=1}^{n}W_{1i}\epsilon_i+O_P(\sqrt{n}\lambda_n^2\vee\sqrt{n}\tau_n^2)\end{aligned}$$

based on (79) and (80). This completes the whole proof.