Submitted to the Annals of Statistics

THE PENALIZED PROFILE SAMPLER

By Guang Cheng* and Michael R. Kosorok*

Duke University and University of North Carolina at Chapel Hill

The penalized profile sampler for semiparametric inference is an extension of the profile sampler method [8] obtained by profiling a penalized log-likelihood. The idea is to base inference on the posterior distribution obtained by multiplying a profiled penalized log-likelihood by a prior for the parametric component, where the profiling and penalization are applied to the nuisance parameter. Because the prior is not applied to the full likelihood, the method is not strictly Bayesian. A benefit of this approximately Bayesian method is that it circumvents the need to put a prior on the possibly infinite-dimensional nuisance components of the model. We investigate the first and second order frequentist performance of the penalized profile sampler, and demonstrate that the accuracy of the procedure can be adjusted by the size of the assigned smoothing parameter. The theoretical validity of the procedure is illustrated for two examples: a partly linear model with normal error for current status data and a semiparametric logistic regression model. As far as we are aware, there are no other methods of inference in this context known to have second order frequentist validity.

*Supported in part by CA075142

AMS 2000 subject classifications: Primary 62G20, 62F25; secondary 62F15, 62F12

Keywords and phrases: Convergence Rate, Empirical Process, Markov Chain Monte Carlo, Partly Linear Model, Penalized Likelihood, Posterior Distribution, Profile Likelihood, Semiparametric Inference, Semiparametric Logistic Regression, Smoothing Parameter

1. Introduction. Semiparametric models are statistical models indexed by both a finite dimensional parameter of interest θ and an infinite dimensional nuisance parameter η. The profile likelihood is typically defined as

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

where $lik_n(\theta, \eta)$ is the likelihood of the semiparametric model given n observations and $\mathcal{H}$ is the parameter space for η. We also define

$$\hat{\eta}_\theta = \text{argmax}_{\eta \in \mathcal{H}} lik_n(\theta, \eta).$$

The convergence rate of the nuisance parameter η is the order of $d(\hat{\eta}_{\tilde{\theta}_n}, \eta_0)$, where $d(\cdot, \cdot)$ is some metric on η, $\tilde{\theta}_n$ is any sequence satisfying $\tilde{\theta}_n = \theta_0 + o_P(1)$, and η_0 is the true value of η. Typically,

$$d(\hat{\eta}_{\tilde{\theta}_n}, \eta_0) = O_P(\|\tilde{\theta}_n - \theta_0\| + n^{-r}), \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm and $r > 1/4$. Of course, a smaller value of r leads to a slower convergence rate of the nuisance parameter. For instance, the nuisance parameter in the Cox proportional hazards model with right censored data, the cumulative hazard function, has the parametric rate, i.e., $r = 1/2$. If current status data is applied to the Cox model instead, then the convergence rate will be slower, with $r = 1/3$, due to the loss of information provided by this kind of data.

The profile sampler is the procedure of sampling from the posterior of the profile likelihood in order to estimate and draw inference on the parametric component θ in a semiparametric model, where the profiling is done over the possibly infinite-dimensional nuisance parameter η. [8] show that the profile sampler gives a first order correct approximation to the maximum likelihood estimator $\hat{\theta}_n$ and consistent estimation of the efficient Fisher information for θ even when the nuisance parameter is not estimable at the $\sqrt{n}$ rate. Another Bayesian procedure employed to do semiparametric estimation is considered in [16] who study the marginal semiparametric posterior distribution for a parameter of interest. In particular, [16] show that marginal semiparametric posterior distributions are asymptotically normal and centered at the corresponding maximum likelihood estimates or posterior means, with covariance matrix equal to the inverse of the Fisher information. Unfortunately, this fully Bayesian method requires specification of a prior on η, which is quite challenging since for some models there is no direct extension of the concept of a Lebesgue dominating measure for the infinite-dimensional parameter set involved [7]. The advantages of the profile sampler for estimating θ compared to other methods is discussed extensively in [2], [3] and [8].

In many semiparametric models involving a smooth nuisance parameter, it is often convenient and beneficial to perform estimation using penalization. One motivation for this is that, in the absence of any restrictions on the form of the function η, maximum likelihood estimation for some semiparametric models leads to over-fitting. Seminal applications of penalized maximum likelihood estimation include estimation of a probability density function in [17] and nonparametric linear regression in [18]. Note that penalized likelihood is a special case of penalized quasi-likelihood studied in [12]. Under certain reasonable regularity conditions, penalized semiparametric log-likelihood estimation can yield fully efficient estimates for θ (see, for example, [12]). As far as we are aware, the only general procedure for inference for θ in this context known to be theoretically valid is a weighted bootstrap with bounded random weights (see [10]). It is even unclear whether the usual nonparametric bootstrap will work in this context when the nuisance parameter has a convergence rate $r < 1/2$.

In contrast, [2] and [3] have shown that the profile sampler procedure without penalization can essentially yield second order frequentist valid inference for θ in semiparametric models, where the estimation accuracy is dependent on the convergence rate of the nuisance parameter. In other words, a faster convergence rate of the nuisance parameters can yield more precise frequentist inference for θ. These second order results are verified in [2] and [3] for several examples, including the Cox model for both right censored and current status data, the proportional odds model, case-control studies with missing covariates, and the partly linear normal model. The convergence rates for these models range from the parametric to the cubic. The work in [3] has shown clearly that the accuracy of the inference for θ based on the profile sampler method is intrinsically determined by the semiparametric model specifications through its entropy number.

The purpose of this paper is to ask the somewhat natural question: does sampling from a profiled penalized log-likelihood (which process we refer hereafter to as the penalized profile sampler) yield first and even second order accurate frequentist inference? The conclusion of this paper is that the answer is yes and, moreover, the accuracy of the inference depends in a fairly simple way on the size of the smoothing parameter.

The unknown parameters in the semiparametric models we study in this paper includes θ, which we assume belongs to some compact set $\Theta \subset \mathbb{R}^d$, and η, which we assume to be a function in the Sobolev class of functions supported on some compact set on the real line, whose k-th derivative exists and is absolutely continuous with $J(\eta) < \infty$, where

$$J^2(\eta) = \int_{\mathcal{Z}} (\eta^{(k)}(z))^2 dz.$$

Here k is a fixed, positive integer and $\eta^{(j)}$ is the j-th derivative of η with respect to z. Obviously $J^2(\eta)$ is some measurement of complexity of η. We denote $\mathcal{H}_k$ as the Sobolev function class with degree k. The penalized log-likelihood in this context is:

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

where $\log lik(\theta, \eta) \equiv \mathbb{P}_n \ell_{\theta,\eta}(X)$, $\ell_{\theta,\eta}(X)$ is the log-likelihood of the single observation X, and λ_n is a smoothing parameter, possibly dependent on data. In practice, λ_n can be obtained by cross-validation [22] or by inspecting the various curves for different values of λ_n. The penalized maximum likelihood estimators $\hat{\theta}_n$ and $\hat{\eta}_n$ depend on the choice of the smoothing parameter λ_n. Consequently we use the notation $\hat{\theta}_{\lambda_n}$ and $\hat{\eta}_{\lambda_n}$ for the remainder of this paper to denote the estimators obtained from maximizing (2). In particular, a larger smoothing parameter usually leads to a less rough penalized estimator of η_0.

For the purpose of establishing first order accuracy of inference for θ based on the penalized profile sampler, we assume that the bounds for the smoothing parameter are in the form below:

$$\lambda_n = o_P(n^{-1/4}) \quad \text{and} \quad \lambda_n^{-1} = O_P(n^{k/(2k+1)}). \tag{3}$$

The condition (3) is assumed to hold throughout this paper. One way to ensure (3) in practice is simply to set $\lambda_n = n^{-k/(2k+1)}$. Or we can just choose

$\lambda_n = n^{-1/3}$ which is independent of k. It turns out that the upper bound guarantees that $\hat{\theta}_{\lambda_n}$ is $\sqrt{n}$-consistent, while the lower bound controls the penalized nuisance parameter estimator convergence rate. Another approach to controlling estimators is to use sieve estimates with assumptions on the derivatives (see [5]). We will not pursue this further here.

The log-profile penalized likelihood is defined as follows:

$$\log pl_{\lambda_n}(\theta) = \log lik(\theta, \hat{\eta}_{\theta,\lambda_n}) - \lambda_n^2 J^2(\hat{\eta}_{\theta,\lambda_n}), \tag{4}$$

where $\hat{\eta}_{\theta,\lambda_n}$ is $argmax_{\eta\in\mathcal{H}_k} \log lik_{\lambda_n}(\theta,\eta)$ for fixed θ and λ_n. The penalized profile sampler is just the procedure of sampling from the posterior distribution of $pl_{\lambda_n}(\theta)$ by assigning a prior on θ. By analyzing the corresponding MCMC chain from the frequentist's point of view, our paper obtains the following conclusions:

1 *Distribution Approximation*: The posterior distribution with respect to $pl_{\lambda_n}(\theta)$ can be approximated by the normal distribution with mean the maximum penalized likelihood estimator of θ and variance the inverse of the efficient information matrix, with error $O_P(n^{1/2}\lambda_n^2)$;

2 *Moment Approximation*: The maximum penalized likelihood estimator of θ can be approximated by the mean of the MCMC chain with error $O_P(\lambda_n^2)$. The efficient information matrix can be approximated by the inverse of the variance of the MCMC chain with error $O_P(n^{1/2}\lambda_n^2)$;

3 *Confidence Interval Approximation*: An exact frequentist confidence interval of Wald's type for θ can be estimated by the credible set obtained from the MCMC chain with error $O_P(\lambda_n^2)$.

Obviously, given any smoothing parameter satisfying the upper bound in (3), the penalized profile sampler can yield first order frequentist valid inference for θ, similar as to what was shown for the profile sampler in [8]. Moreover, the above conclusions are actually second order frequentist valid results, whose approximation accuracy is directly controlled by the smoothing parameter. Note that the corresponding results for the usual (non-penalized) profile sampler with nuisance parameter convergence rate r in [3] are obtained by replacing in the above $O_P(n^{1/2}\lambda_n^2)$ with $O_P(n^{-1/2} \vee n^{-r+1/2})$ and $O_P(\lambda_n^2)$ with $O_P(n^{-1} \vee n^{-r})$, for all respective occur where r is as defined in (1).

Our results are the first higher order frequentist inference results for penalized semiparametric estimation. The layout of the article is as follows. The next section, section 2, introduces the two main examples we will be using for illustration: partly linear regression for current status data and semiparametric logistic regression. Some background is given in section 3, including the concept of a least favorable submodel as well as some notations and the main model assumptions. In section 4, some preliminary results are developed, including three rather different theorems concerning the convergence rates of the penalized nuisance parameters and the order of the estimated penalty term under different conditions. The corresponding rates for the two featured examples are also calculated in this section. The main results and implications are discussed in section 5, and all remaining model assumptions are verified for the examples in section 6. A brief discussion of future work is given in section 7. We postpone all technical tools and proofs to the last section, section 8.

2. Examples.

2.1. *Partly Linear Normal Model with Current Status Data.* In this example, we study the partly linear regression model with normal residue error. The continuous outcome Y, conditional on the covariates $(U, V) \in \mathbb{R}^d \times \mathbb{R}$, is modeled as

$$Y = \theta^T U + f(V) + \epsilon, \tag{5}$$

where f is an unknown smooth function, and $\epsilon \sim N(0, \sigma^2)$ with finite variance σ^2. For simplicity, we assume for the rest of the paper that $\sigma = 1$. The theory we propose also works when σ is unknown, but the added complexity would detract from the main issues. We also assume that only the current status of response Y is observed at a random censoring time $C \in \mathbb{R}$. In other words, we observe $X = (C, \Delta, U, V)$, where indicator $\Delta = 1\{Y \leq C\}$. Current status data may occur due to study design or measurement limitations. Examples of such data arise in several fields, including demography, epidemiology and econometrics. For simplicity of exposition, θ is assumed to be one dimensional.

Under the model (5) and given that the joint distribution for (C, U, V) does not involve parameters (θ, f), the log-likelihood for a single observation at $X = x \equiv (c, \delta, u, v)$ is

$$\begin{aligned} loglik_{\theta,f}(x) &= \delta \log\left\{\Phi\left(c - \theta u - f(v)\right)\right\} \\ &\quad +(1-\delta)\log\left\{1 - \Phi\left(c - \theta u - f(v)\right)\right\}, \end{aligned} \tag{6}$$

where Φ is the standard normal distribution. The parameter of interest, θ, is assumed to belong to some compact set in $\mathbb{R}^1$. The nuisance parameter is the function f, which belongs to the Sobolev function class of degree k. We further make the following assumptions on this model. We assume that

(Y, C) is independent given (U, V). The covariates (U, V) are assumed to belong to some compact set, and the support for random censoring time C is an interval $[l_c, u_c]$, where $-\infty < l_c < u_c < \infty$. In addition, $EVar(U|V)$ is strictly positive and $Ef(V) = 0$. The first order asymptotic behaviors of the penalized log-likelihood estimates of a slightly more general version of this model have been extensively studied in [9].

2.2. *Semiparametric Logistic Regression.* Let $X_1 = (Y_1, W_1, Z_1)$, $X_2 = (Y_2, W_2, Z_2)$, ... be independent copies of $X = (Y, W, Z)$, where Y is a dichotomous variable with conditional expectation $E(Y|W, Z) = F(\theta^T W + \eta(Z))$. $F(u)$ is the logistic distribution defined as $e^u/(e^u + 1)$. Obviously the likelihood for a single observation is of the following form:

$$p_{\theta,\eta}(x) = F(\theta^T w + \eta(z))^y (1 - F(\theta^T w + \eta(z)))^{1-y} f^{(W,Z)}(w, z). \quad (7)$$

This example is a special case of quasi-likelihood in partly linear models when the conditional variance of response Y is taken to have some quadratic form of the conditional mean of Y. In the absence of any restrictions on the form of the function η, the maximum likelihood of this simple model often leads to over-fitting. Hence [4] propose maximizing instead the penalized likelihood of the form $\log lik(\theta, \eta) - \lambda_n^2 J^2(\eta)$; and [12] studied the asymptotic properties of the maximum penalized likelihood estimators for θ and η. For simplicity, we will restrict ourselves to the case where $\Theta \subset \mathbb{R}^1$ and (W, Z) have bounded support, say $[0, 1]^2$. To ensure the identifiability of the parameters, we assume that $EVar(W|Z)$ is positive and that the support of $\mathcal{Z}$ contains at least k distinct points in $[0, 1]$.

REMARK 1. *Another interesting potential example we may apply the* *penalized profile sampler method to is the classic proportional hazards model with current status data by penalizing the cumulative hazard function with its Sobolev norm. There are two motivations for us to penalize the cumulative hazard function in the Cox model. One is that the estimated step functions from the unpenalized estimation cannot be used easily for other estimation or inference purposes. Another issue with the unpenalized approach is that without making stronger continuity assumptions, we cannot achieve uniform consistency even on a compact set [9]. The asymptotic properties of the corresponding penalized M-estimators have been studied in [11].*

3. Preliminaries.

In this section, we present some necessary preliminary material concerning least favorable submodels, general notational conventions for the paper, and an enumeration of the main assumptions.

3.1. *Least favorable submodels.* In this subsection, we briefly review the concept of a least favorable submodel. A submodel $t \mapsto p_{t,\eta_t}$ is defined to be least favorable at (θ, η) if $\tilde{\ell}_{\theta,\eta} = \partial/\partial t \log p_{t,\eta_t}$, given $t = \theta$, where $\tilde{\ell}_{\theta,\eta}$ is the efficient score function for θ. The efficient score function for θ can be viewed as the projection of the score function for θ onto the tangent space of η. The inverse of its variance is exactly the efficient information matrix $\tilde{I}_{\theta,\eta}$. We abbreviate hereafter $\tilde{\ell}_{\theta_0,\eta_0}$ and $\tilde{I}_{\theta_0,\eta_0}$ with $\tilde{\ell}_0$ and $\tilde{I}_0$, respectively. The "direction" along which η_t approaches η in the least favorable submodel is called the least favorable direction. An insightful review about least favorable submodels and efficient score functions can be found in Chapter 3 of [6]. By the above construction of the least favorable submodel, $\log pl_{\lambda_n}(\theta)$ can be rewritten in the following form:

$$\log pl_{\lambda_n}(\theta) \quad = \quad \ell(\theta,\theta,\hat{\eta}_{\theta,\lambda_n}) - \lambda_n^2 J^2(\eta_\theta(\theta,\hat{\eta}_{\theta,\lambda_n})), \tag{8}$$

where $\ell(t,\theta,\eta)(x) = \log lik(t,\eta_t(\theta,\eta))(x)$, $t \mapsto \eta_t(\theta,\eta)$ is a general map from the neighborhood of θ into the parameter set for η, with $\eta_\theta(\theta,\eta) = \eta$. The concrete forms of (8) will depend on the situation.

3.2. *Notation.* We present in this subsection some notation that will be used throughout the paper. The derivatives of the function $\ell(t,\theta,\eta)$ are with respect to its first argument, t. For the derivatives relative to the other two arguments θ and η, we use the following shortened notation: $\ell_\theta(t,\theta,\eta)$ indicates the first derivative of $\ell(t,\theta,\eta)$ with respect to θ. Similarly, $\ell_{t,\theta}(t,\theta,\eta)$ denotes the derivative of $\dot{\ell}(t,\theta,\eta)$ with respect to θ. Also, $\ell_{t,t}(\theta)$ and $\ell_{t,\theta}(\eta)$ indicate the maps $\theta \mapsto \ddot{\ell}(t,\theta,\eta)$ and $\eta \mapsto \ell_{t,\theta}(t,\theta,\eta)$, respectively. For brevity, we denote $\dot{\ell}_0 = \dot{\ell}(\theta_0,\theta_0,\eta_0)$, $\ddot{\ell}_0 = \ddot{\ell}(\theta_0,\theta_0,\eta_0)$ and $\ell_0^{(3)} = \ell^{(3)}(\theta_0,\theta_0,\eta_0)$, where θ_0, η_0 are the true values of θ and η. Of course, we can write $\tilde{\ell}(X)$ as $\dot{\ell}_0(X)$. $\|\cdot\|$ and $\|\cdot\|_2$ indicate the Euclidean norm and L_2 norm, respectively. The notations $\gtrsim$ and $\lesssim$ mean greater than, or smaller than, up to a universal constant. 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 processes of the observations, respectively.

3.3. *Main Assumptions.* We now make the following three classes of assumptions: Rate assumptions (R1) for the penalized nuisance parameter and the estimated penalty term; Smoothness assumptions (S1-S2) and Empirical processes assumptions (E1) for $\ell(t,\theta,\eta)$ and its related derivatives.

R1 : Assume:

$$d(\hat{\eta}_{\tilde{\theta}_n,\lambda_n},\eta_0) = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|) \tag{9}$$

and

$$\lambda_n J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|). \tag{10}$$

S1 : The maps

$$(t,\theta,\eta) \quad \mapsto \quad \frac{\partial^{l+m}}{\partial t^l \partial \theta^m}\ell(t,\theta,\eta) \tag{11}$$

have integrable envelope functions in $L_1(P)$ in some neighborhood of $(\theta_0,\theta_0,\eta_0)$, for $(l,m) = (0,0),(1,0),(2,0),(3,0),(1,1),(1,2),(2,1)$.

S2 : Assume:

$$P\ddot{\ell}(\theta_0,\theta_0,\eta) - P\ddot{\ell}(\theta_0,\theta_0,\eta_0) = O(d(\eta,\eta_0)), \tag{12}$$

$$P\ell_{t,\theta}(\theta_0,\theta_0,\eta) - P\ell_{t,\theta}(\theta_0,\theta_0,\eta_0) = O(d(\eta,\eta_0)), \tag{13}$$

$$P\dot{\ell}(\theta_0,\theta_0,\eta) = O(d^2(\eta,\eta_0)), \tag{14}$$

for all η in some neighborhood of η_0.

E1 : For all random sequences $\tilde{\theta}_n = \hat{\theta}_n + o_P(1)$ and $\bar{\theta}_n = \theta_0 + o_P(1)$, we have

$$\mathbb{G}_n(\dot{\ell}(\theta_0,\theta_0,\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) - \dot{\ell}_0) = O_P(n^{\frac{1}{4k+2}}(\lambda_n + \|\tilde{\theta}_n - \theta_0\|)), \tag{15}$$

$$\mathbb{G}_n(\ddot{\ell}(\theta_0,\tilde{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n})) = O_P(1), \tag{16}$$

$$\mathbb{G}_n(\ell_{t,\theta}(\theta_0,\bar{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n})) = O_P(1), \tag{17}$$

$$(\mathbb{P}_n - P)\ell^{(3)}(\bar{\theta}_n,\tilde{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) = o_P(1). \tag{18}$$

Assumption R1 implicitly assumes that we have a metric or topology defined on the set of possible values of the nuisance parameter η. The form of

$d(\eta,\eta_0)$ may vary for different situations and does not need to be specified in this subsection beyond the given conditions. (9) implies that $\hat{\eta}_{\tilde{\theta}_n,\lambda_n}$ is consistent for η_0 as $\tilde{\theta}_n \to \theta_0$ in probability. Additionally, from (10) we know that the smoothing parameter λ_n plays a role in determining the complexity degree of the estimated nuisance parameter. (10) implies that $J(\hat{\eta}_{\lambda_n}) = O_P(1)$ if the $\hat{\theta}_{\lambda_n}$ is asymptotically normal, which has been shown in (37). Note that $J(\hat{\eta}_{\tilde{\theta}_n,0}) \geq J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n})$, where $\hat{\eta}_{\theta,0} = \hat{\eta}_\theta \equiv argmax_{\eta\in\mathcal{H}} \log lik(\theta,\eta)$ for a fixed θ, based on the inequality that $\log lik_{\lambda_n}(\tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,0}) \leq \log lik_{\lambda_n}(\tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n})$.

Clearly, the assumptions S1 and S2 are separately the smoothness conditions for the Euclidean parameters (t,θ) and the infinite dimensional nuisance parameter η. The boundedness of the Fréchet derivatives of the maps $\eta \mapsto \ddot{\ell}(\theta_0,\theta_0,\eta)$ and $\eta \mapsto \ell_{t,\theta}(\theta_0,\theta_0,\eta)$ ensures the validity of conditions (12) and (13). Based on the discussions in section 2 of [3], under the given regularity conditions, it suffices to show (14) if the map $\eta \mapsto \dot{\ell}(\theta_0,\theta_0,\eta)$ is Fréchet differentiable and the map $\eta \mapsto lik(\theta_0,\eta)$ is second order Fréchet differentiable.

Condition (15) is concerned with the asymptotic equicontinuity of the empirical process measure of $\dot{\ell}(\theta_0,\theta_0,\eta)$ with η ranging around the neighborhood of η_0. It suffices to show (16) and (17) if $\mathbb{G}_n(\ddot{\ell}(\theta_0,\tilde{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) - \ddot{\ell}_0) = o_P(1)$ and $\mathbb{G}_n(\ell_{t,\theta}(\theta_0,\bar{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) - \ell_{t,\theta}(\theta_0,\theta_0,\eta_0)) = o_P(1)$, provided $\ddot{\ell}_0$ and $\ell_{t,\theta}(\theta_0,\theta_0,\eta_0)$ are square integrable. Thus we will be able to use technical tools T2 and T6 given in the appendix to show (15)–(17). For the verification of (18), we need to make use of a Glivenko-Cantelli theorem for classes of functions that change with n which is a modification of theorem 2.4.3 in [21] and is explained in the appendix.

In principle, assumptions S1, S2 and E1 on the functions of the least favorable submodel directly imply the following empirical no-bias conditions:

$$\mathbb{P}_n\dot{\ell}(\theta_0,\tilde{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) = \mathbb{P}_n\tilde{\ell}_0 + O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|)^2, \tag{19}$$

$$\mathbb{P}_n\ddot{\ell}(\theta_0,\tilde{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) = P\ddot{\ell}_0 + O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|). \tag{20}$$

The derivations of (19) and (20) are simply based on the regular Taylor expansions around the true values. The detailed arguments can be found in the proof of lemmas 1 and 2 in [3]. The two empirical no-bias conditions ensure that the penalized profile likelihood behaves like a penalized likelihood in the parametric model asymptotically and therefore yields a second order asymptotic expansion of the penalized profile log-likelihood.

4. The Penalized Convergence Rate. In the previous section, we have imposed two assumptions about the convergence rates of the estimated nuisance parameter and the order of the estimated penalty term, i.e. (9) and (10). To compute the convergence rates, we present three different theorems below which require different sets of conditions. These theorems can be viewed as extension of general results on M-estimators to penalized M-estimators, and are therefore of independent interest. We first state the classical definitions for the covering number (entropy number) and bracketing number (bracketing entropy number) for a class of functions.

Definition: Let $\mathcal{A}$ be a subset of a (pseudo-) metric space $(\mathcal{L},d)$ of real-valued functions. The δ-covering number $N(\delta,\mathcal{A},d)$ of $\mathcal{A}$ is the smallest N for which there exist functions $a_1,\ldots,a_N$ in $\mathcal{L}$, such that for each $a\in\mathcal{A}$, $d(a,a_j)\le\delta$ for some $j\in\{1,\ldots,N\}$. The δ-bracketing number $N_B(\delta,\mathcal{A},d)$ is the smallest N for which there exist pairs of functions $\{[a_j^L,a_j^U]\}_{j=1}^N\subset\mathcal{L}$, with $d(a_j^L, a_j^U) \leq \delta$, $j = 1, \dots, N$, such that for each $a \in \mathcal{A}$ there is a $j \in \{1, \dots, N\}$ such that $a_j^L \leq a \leq a_j^U$. The δ-entropy number (δ-bracketing entropy number) is defined as $H(\delta, \mathcal{A}, d) = \log N(\delta, \mathcal{A}, d)$ ($H_B(\delta, \mathcal{A}, d) = \log N_B(\delta, \mathcal{A}, d)$).

Before we present the first theorem, define

$$\mathcal{K} = \left\{ \frac{\ell_{\theta,\eta}(X) - \ell_0(X)}{1 + J(\eta)} : \|\theta - \theta_0\| \leq C_1, \|\eta - \eta_0\|_\infty \leq C_1, J(\eta) < \infty \right\},$$

for a known constant $C_1 < \infty$:

THEOREM 1. *Assume conditions (21), (22), (23) and (24) below hold for every $\theta \in \Theta_n$ and $\eta \in \mathcal{V}_n$:*

$$H_B(\epsilon, \mathcal{K}, L_2(P)) \lesssim \epsilon^{-1/k}, \tag{21}$$

$$p_{\theta,\eta}/p_{\theta,\eta_0} \ \textit{is bounded away from zero and infinity}, \tag{22}$$

$$\|\ell_{\theta,\eta} - \ell_0\|_2 \lesssim \|\theta - \theta_0\| + d_\theta(\eta, \eta_0), \tag{23}$$

$$P(\ell_{\theta,\eta} - \ell_{\theta,\eta_0}) \lesssim -d_\theta^2(\eta, \eta_0) + \|\theta - \theta_0\|^2. \tag{24}$$

Then we have

$$d_{\tilde{\theta}_n}(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}, \eta_0) = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|),$$
$$\lambda_n J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|),$$

for $(\tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n})$ satisfying $P(\tilde{\theta}_n \in \Theta_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n} \in \mathcal{V}_n) \to 1$.

Condition (21) determines the order of the increments of the empirical processes indexed by $\ell_{\theta,\eta}$. A detailed discussion about how to compute the increments of the empirical processes can be found in chapter 5 of [19]. Condition (22) is equivalent to the condition that $p_{\theta,\eta}$ is bounded away from zero uniformly in x for (θ, η) ranging over $\Theta_n \times \mathcal{V}_n$. Given that the distance function $d_\theta(\eta, \eta_0)$ in (23) is just $\|p_{\theta,\eta} - p_0\|_2$, (23) trivially holds provided that condition (22) holds. For the verification of (24), we can do an analysis as follows. The natural Taylor expansions of the criterion function $(\theta, \eta) \mapsto P\ell_{\theta,\eta}$ around the maximum point (θ_0, η_0) implies that $P(\ell_{\theta,\eta_0} - \ell_{\theta_0,\eta_0}) \gtrsim -\|\theta - \theta_0\|^2$, and (46) implies that $P(\ell_{\theta,\eta} - \ell_0) \leq -\int(\sqrt{p_{\theta,\eta}} - \sqrt{p_0})^2 d\mu \leq -\|p_{\theta,\eta} - p_0\|_2^2$ given condition (22).

We now apply theorem 1 to derive the related convergence rates in the partly linear model in corollary 1. However, we need to strengthen our previous assumptions to require the existence of a known $M < \infty$ such that $\eta \in \mathcal{H}_k^M$, where $\mathcal{H}_k^M = \mathcal{H}_k \cap \{\|\eta\|_\infty \leq M\}$ and that the density for the joint distribution (U, V, C) is strictly positive and finite. The additional assumptions here guarantee condition (22). The following theorem 2 and theorem 3 can also be employed to derive the convergence rate of the non-penalized estimated nuisance parameter by setting λ_n to zero. However, we would need to assume that $f \in \{g : \|g\|_\infty + J(g) \leq \tilde{M}\}$ for some known $\tilde{M}$ when applying these theorems. Thus we can argue that the the penalized method enables a relaxation of the assumptions needed for the nuisance parameter.

COROLLARY 1. *Under the above set-up for the partly linear normal model with current status data, we have, for* $\tilde{\theta}_n = \theta_0 + o_P(1)$,

$$\|\hat{f}_{\tilde{\theta}_n,\lambda_n} - f_0\|_2 = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|), \tag{25}$$

$$\lambda_n J(\hat{f}_{\tilde{\theta}_n,\lambda_n}) = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|). \tag{26}$$

Moreover, if we also assume that $f \in \{g : \|g\|_\infty + J(g) \leq \tilde{M}\}$ *for some* *known $\tilde{M}$, then*

$$\|\hat{f}_{\tilde{\theta}_n} - f_0\|_2 = O_P(n^{-k/(2k+1)} + \|\tilde{\theta}_n - \theta_0\|), \tag{27}$$

provided condition (3) holds.

REMARK 2. *Corollary 1 implies that the convergence rate of the estimated nuisance parameter is slower than that of the the regular nuisance parameter by comparing (25) and (27). This result is not surprising since the slower rate is the trade off for the smoother nuisance parameter estimator. However the advantage of the penalized profile sampler is that we can control the convergence rate by assigning the smoothing parameter with different rates. Corollary 1 also indicates that $\|\hat{f}_{\lambda_n} - f_0\|_2 = O_P(\lambda_n)$ and $\|\hat{f}_n - f_0\|_2 = O_P(n^{-k/(k+2)})$. Note that the convergence rate of the maximum penalized likelihood estimator, $O_P(\lambda_n)$, is deemed as the optimal rate in [22]. Similar remarks also hold for corollary 2 below.*

The boundedness condition (22) appears hard to achieve in some examples. Hence we propose theorem 2 below to relax this condition by choosing the criterion function $m_{\theta,\eta} = \log[(p_{\theta,\eta} + p_{\theta,\eta_0})/2p_{\theta,\eta_0}]$. Obviously, $m_{\theta,\eta}$ is trivially bounded away from zero. It is also bounded above for (θ, η) around the their true values if $p_{\theta,\eta_0}(x)$ is bounded away from zero uniformly in x and $p_{\theta,\eta}$ is bounded above. The first condition is satisfied if the map $\theta \mapsto p_{\theta,\eta_0}(x)$ is continuous around θ_0 and $p_0(x)$ is uniformly bounded away from zero. The second condition is trivially satisfied in the semiparametric logistic regression model by the given form of the density. The boundedness of $m_{\theta,\eta}$ thus permits the application of lemma 1 below which is used to verify condition (29) in the following theorem:

THEOREM 2. *Assume for any given $\theta \in \Theta_n$, $\hat{\eta}_\theta$ satisfies $\mathbb{P}_n m_{\theta,\hat{\eta}_\theta} \geq \mathbb{P}_n m_{\theta,\eta_0}$ for given measurable functions $x \mapsto m_{\theta,\eta}(x)$. Assume conditions (28) and (29) below hold for every $\theta \in \Theta_n$, every $\eta \in \mathcal{V}_n$ and every $\epsilon > 0$:*

$$P(m_{\theta,\eta} - m_{\theta,\eta_0}) \lesssim -d_\theta^2(\eta,\eta_0) + \|\theta - \theta_0\|^2, \tag{28}$$

$$E^* \sup_{\theta\in\Theta_n, \eta\in\mathcal{V}_n, \|\theta-\theta_0\|<\epsilon, d_\theta(\eta,\eta_0)<\epsilon} |\mathbb{G}_n(m_{\theta,\eta} - m_{\theta,\eta_0})| \lesssim \phi_n(\epsilon). \tag{29}$$

Suppose that (29) is valid for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ and sets $\Theta_n \times \mathcal{V}_n$ such that $P(\tilde{\theta} \in \Theta_n, \hat{\eta}_{\tilde{\theta}} \in \mathcal{V}_n) \to 1$. Then $d_{\tilde{\theta}}(\hat{\eta}_{\tilde{\theta}}, \eta_0) \leq O_P^*(\delta_n + \|\tilde{\theta} - \theta_0\|)$ for any sequence of positive numbers δ_n such that $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$ for every n.

Lemma 1 below is presented to verify the modulus condition for the continuity of the empirical process in (29). Let $\mathcal{S}_\delta = \{x \mapsto m_{\theta,\eta}(x) - m_{\theta,\eta_0}(x) : d_\theta(\eta,\eta_0) < \delta, \|\theta - \theta_0\| < \delta\}$ and write

$$K(\delta, \mathcal{S}_\delta, L_2(P)) = \int_0^\delta \sqrt{1 + H_B(\epsilon, \mathcal{S}_\delta, L_2(P))} d\epsilon : \tag{30}$$

LEMMA 1. *Suppose the functions $(x, \theta, \eta) \mapsto m_{\theta,\eta}(x)$ are uniformly bounded for (θ, η) ranging over a neighborhood of (θ_0, η_0) and that*

$$P(m_{\theta,\eta} - m_{\theta_0,\eta_0})^2 \lesssim d_\theta^2(\eta,\eta_0) + \|\theta - \theta_0\|^2.$$

Then condition (29) is satisfied for any functions ϕ_n such that

$$\phi_n(\delta) \geq K(\delta, \mathcal{S}_\delta, L_2(P)) \left(1 + \frac{K(\delta, \mathcal{S}_\delta, L_2(P))}{\delta^2\sqrt{n}}\right)$$

Consequently, in the conclusion of the above theorem we may use $K(\delta, \mathcal{S}_\delta, L_2(P))$ rather than $\phi_n(\delta)$.

REMARK 3. *Theorem 2 and lemma 1 are theorem 3.2 and lemma 3.3 in [14], respectively. We can apply theorem 2 to the penalized semiparametric logistic regression model by including* λ *in* θ*, i.e.* $m_{\theta,\lambda,\eta} = m_{\theta,\eta} - \frac{1}{2}\lambda^2(J^2(\eta) - J^2(\eta_0))$. *This is accomplished in the following corollary. Note that we assume that the uniform norm and Sobolev norm of* η *are bounded above with known upper bounds when deriving (33) of the corollary, but this assumption is not needed for (31) and (32).*

COROLLARY 2. *Under the above set-up for the semiparametric logistic regression model, we have for* λ_n *satisfying condition (3) and any* $\tilde{\theta}_n \xrightarrow{p} \theta_0$ *that*

$$\|\hat{\eta}_{\tilde{\theta}_n,\lambda_n} - \eta_0\|_2 = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|), \tag{31}$$

$$\lambda_n J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|). \tag{32}$$

If we also assume that $\eta \in \{g : \|g\|_\infty + J(g) \leq \tilde{M}\}$ *for some known* $\tilde{M}$*, then*

$$\|\hat{\eta}_{\tilde{\theta}_n} - \eta_0\|_2 = O_P(n^{-k/(2k+1)} + \|\tilde{\theta}_n - \theta_0\|). \tag{33}$$

REMARK 4. *Corollary 1 and 2 imply that* $J(\hat{\eta}_{\lambda_n}) = O_P(1)$ *and* $J(\hat{f}_{\lambda_n}) = O_P(1)$*, respectively. Thus the maximum likelihood estimators of the nuisance parameters in the two examples of this paper are consistent in the uniform norm, i.e.* $\|\hat{\eta}_{\lambda_n} - \eta_0\|_\infty = o_P(1)$ *and* $\|\hat{f}_{\lambda_n} - f_0\|_\infty = o_P(1)$*, since the sequences* $\hat{\eta}_n$ *and* $\hat{f}_n$ *consist of smooth functions defined on a compact set with asymptotically bounded first-order derivatives.*

The preceding two theorems imply that the convergence rate of the penalized estimated nuisance parameter is affected by the assigned smoothness parameter. However, the next theorem shows that, under different conditions, the above phenomena may not hold. Let

$$\begin{aligned} l^{\lambda_n}_{\theta,\eta,h} &= \frac{\partial}{\partial t}|_{t=0} \log lik_{\lambda_n}(\theta,\eta_t) = A_{\theta,\eta}h - 2\lambda_n^2 \int h^{(k)}\eta^{(k)}dz, \\ V(\theta,\eta)h &= PA_{\theta,\eta}h, \\ \mathbb{V}_n(\theta,\eta)h &= \mathbb{P}_n A_{\theta,\eta}h, \end{aligned}$$

where $\eta_t = \eta + th$ for $h \in \mathcal{H}_k$ and $A_{\theta,\eta}$ is the appropriate score operator for the model. Note that $\eta_t \in \mathcal{H}_k$ for sufficiently small t. Obviously $\mathbb{P}_n l^{\lambda_n}_{\tilde{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n},h} = 0$ and $V(\theta_0,\eta_0)h = 0$. We assume that the maps $h \mapsto V(\theta,\eta)h$ and $h \mapsto \mathbb{V}_n(\theta,\eta)h$ are uniformly bounded such that $\mathbb{V}_n$ and V can be viewed as maps from the parameters set $\Theta \times \mathcal{H}_k$ into $\ell^\infty(\mathcal{H}_k)$. Further we require the following regularity conditions: For some $C_2 > 0$,

$$\{A_{\theta,\eta}h : \|\theta - \theta_0\| < C_2, d_\theta(\eta,\eta_0) < C_2, h \in \mathcal{H}_k\} \text{ is } P\text{-Donsker}, \tag{34}$$

$$\sup_{h\in\mathcal{H}_k} P(A_{\theta,\eta}h - A_{\theta_0,\eta_0}h)^2 \to 0, \text{ as } \theta \to \theta_0 \text{ and } \eta \to \eta_0. \tag{35}$$

THEOREM 3. *Suppose that $V(\cdot,\cdot) : \Theta \times \mathcal{H}_k \mapsto \ell^\infty(\mathcal{H}_k)$ is Fréchet differentiable at (θ_0,η_0) with derivative $\dot{V}(\cdot,\cdot) : \mathbb{R}^d \times lin\mathcal{H}_k \mapsto \ell^\infty(\mathcal{H}_k)$ such that the map $\dot{V}(0,\cdot) : lin\mathcal{H}_k \mapsto \ell^\infty(\mathcal{H}_k)$ is invertible with an inverse that is continuous on its range. Furthermore, we assume that (34) and (35) hold. Then*

$$d_{\tilde{\theta}_n}(\hat{\eta}_{\tilde{\theta}_n,\lambda_n},\eta_0) = O_P(n^{-1/2} + \|\tilde{\theta}_n - \theta_0\| + \lambda_n^2 J^2(\hat{\eta}_{\tilde{\theta}_n,\lambda_n})), \tag{36}$$

for $\tilde{\theta}_n \to \theta_0$ and $\hat{\eta}_{\tilde{\theta}_n,\lambda_n} \to \eta_0$ in probability.

REMARK 5. *The preceding theorem is a variation of theorems used in [13] and [20], among others, to prove the asymptotic normality of the* *maximum likelihood estimator* $(\hat{\theta}_n, \hat{\eta}_n)$. *If we can show that* $\lambda_n J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|)$ *by some other means, then (36) implies that* $d_{\tilde{\theta}_n}(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}, \eta_0) = O_P(n^{-1/2} + \|\tilde{\theta}_n - \theta_0\|)$. *This indicates that the smoothing effect of the penalized method does not occur, which may be due to some very smooth non-penalized estimated nuisance parameter. The high degree of the smoothness of the non-penalized estimated nuisance parameter can be deduced from its fast convergence rate which equals the parametric rate in this instance.*

5. Main Results and Implications. In this section we first present second order asymptotic expansion of the log-profile penalized likelihood which prepare us for deriving the main results about the higher order structure of the penalized profile sampler. The assumptions in section 3 and condition (3) are assumed throughout.

THEOREM 4. *Given* $\tilde{\theta}_n = \hat{\theta}_{\lambda_n} + o_P(1)$, *we have*

$$\sqrt{n}(\hat{\theta}_{\lambda_n} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{I}_0^{-1}\tilde{\ell}_0(X_i) + O_P(n^{1/2}\lambda_n^2), \tag{37}$$

$$\begin{aligned} \log pl_{\lambda_n}(\tilde{\theta}_n) &= \log pl_{\lambda_n}(\hat{\theta}_{\lambda_n}) - \frac{n}{2}(\tilde{\theta}_n - \hat{\theta}_{\lambda_n})^T \tilde{I}_0(\tilde{\theta}_n - \hat{\theta}_{\lambda_n}) \\ &+ O_P(g_{\lambda_n}(\|\tilde{\theta}_n - \hat{\theta}_{\lambda_n}\|)), \end{aligned} \tag{38}$$

where $g_{\lambda_n}(w) = nw^3 + nw^2\lambda_n + nw\lambda_n^2 + n^{1/2}\lambda_n^2$, *provided the efficient information* $\tilde{I}_0$ *is positive definite.*

REMARK 6. *The results in theorem 4 are useful in there own right for inference about* θ. *(37) is a second higher order frequentist result in penalized semiparametric estimation regarding the asymptotic linearity of the maximum penalized likelihood estimator of* θ.

We now state the main results on the penalized posterior profile distribution. A preliminary result, theorem 5 with corollary 3 below, shows that the penalized posterior profile distribution is asymptotically close enough to the distribution of a normal random variable with mean $\hat{\theta}_{\lambda_n}$ and variance $(n\tilde{I}_0)^{-1}$ with second order accuracy, which is controlled by the smoothing parameter. Similar conclusions also hold for the penalized posterior moments. Another main result, theorem 6, shows that the penalized posterior profile log-likelihood can be used to achieve second order accurate frequentist inference for θ.

Let $\tilde{P}^{\lambda_n}_{\theta|\tilde{X}}$ be the penalized posterior profile distribution of θ with respect to the prior $\rho(\theta)$. Define

$$\begin{aligned}\Delta_{\lambda_n}(\theta) &\equiv n^{-1}\{\log pl_{\lambda_n}(\theta) - \log pl_{\lambda_n}(\hat{\theta}_{\lambda_n})\}\\ &= n^{-1}(\ell_n(\theta,\hat{\eta}_{\theta,\lambda_n}) - \ell_n(\hat{\theta}_{\lambda_n},\hat{\eta}_{\lambda_n})) - n^{-1}\lambda_n^2(J^2(\hat{\eta}_{\theta,\lambda_n}) - J^2(\hat{\eta}_{\lambda_n})).\end{aligned}$$

THEOREM 5. *Assume that*

$$\Delta_{\lambda_n}(\tilde{\theta}_n) = o_P(1) \quad \textit{implies} \quad \tilde{\theta}_n = \theta_0 + o_P(1), \tag{39}$$

for every random $\left\{\tilde{\theta}_n\right\} \in \Theta$. *If* $\rho(\theta_0) > 0$ *and* $\rho(\cdot)$ *has continuous and finite first order derivative in some neighborhood of* θ_0, *then we have, for any* $-\infty < \xi < \infty$,

$$\sup_{\xi\in\mathbb{R}^d}\left|\tilde{P}^{\lambda_n}_{\theta|\tilde{X}}(\sqrt{n}\tilde{I}_0^{1/2}(\theta-\hat{\theta}_{\lambda_n}) \le \xi) - \Phi_d(\xi)\right| = O_P(n^{1/2}\lambda_n^2), \tag{40}$$

where $\Phi_d(\cdot)$ *is the distribution of the d-dimensional standard normal random variable.*

COROLLARY 3. *Under the assumptions of theorem 5, we have that if θ has finite second absolute moment, then*

$$\hat{\theta}_{\lambda_n} = E^{\lambda_n}_{\theta|\tilde{X}}(\theta) + O_P(\lambda_n^2), \tag{41}$$

$$\tilde{I}_0 = n^{-1}(Var^{\lambda_n}_{\theta|\tilde{X}}(\theta))^{-1} + O_P(n^{1/2}\lambda_n^2), \tag{42}$$

where $E^{\lambda_n}_{\theta|\tilde{X}}(\theta)$ and $Var^{\lambda_n}_{\theta|\tilde{X}}(\theta)$ are the penalized posterior profile mean and penalized posterior profile covariance matrix, respectively.

We now present another second order asymptotic frequentist property of the penalized profile sampler in terms of quantiles. The α-th quantile of the penalized posterior profile distribution, $\tau_{n\alpha}$, is defined as $\tau_{n\alpha} = inf\{\xi : \tilde{P}^{\lambda_n}_{\theta|\tilde{X}}(\theta \leq \xi) \geq \alpha\}$. Without loss of generality, $\tilde{P}^{\lambda_n}_{\theta|\tilde{X}}(\theta \leq \tau_{n\alpha}) = \alpha$. We can also define $\kappa_{n\alpha} \equiv \sqrt{n}(\tau_{n\alpha} - \hat{\theta}_{\lambda_n})$, i.e., $\tilde{P}^{\lambda_n}_{\theta|\tilde{X}}(\sqrt{n}(\theta - \hat{\theta}_{\lambda_n}) \leq \kappa_{n\alpha}) = \alpha$.

THEOREM 6. *Under the assumptions of theorem 5 and assuming that $\tilde{\ell}_0(X)$ has finite third moment with a nondegenerate distribution, then there exists a $\hat{\kappa}_{n\alpha}$ based on the data such that $P(\sqrt{n}(\hat{\theta}_{\lambda_n} - \theta_0) \leq \hat{\kappa}_{n\alpha}) = \alpha$ and $\hat{\kappa}_{n\alpha} - \kappa_{n\alpha} = O_P(n^{1/2}\lambda_n^2)$ for each choice of $\kappa_{n\alpha}$.*

REMARK 7. *Theorem 6 ensures that there exists a unique α-th quantile for θ up to $O_P(\lambda_n^2)$ in the frequentist set-up for each fixed $\tau_{n\alpha}$. Note that $\tau_{n\alpha}$ is not unique if the dimension of θ is larger than one.*

REMARK 8. *Theorem 5, corollary 3 and theorem 6 above show that the penalized profile sampler generates second order asymptotic frequentist valid results in terms of distributions, moments and quantiles. Moreover, the second order accuracy of this procedure is controlled by the smoothing parameter.*

REMARK 9. *Another interpretation for the role of λ_n in the penalized profile sampler is that we can view λ_n as the prior on $J(\eta)$, or on η to some extent. To see this, we can write $lik_{\lambda_n}(\theta,\eta)$ in the following form:*

$$lik_{\lambda_n}(\theta,\eta) = lik_n(\theta,\eta) \times \exp\left[-\frac{J^2(\eta)}{2(\frac{1}{2\lambda_n^2})}\right]$$

This idea can be traced back to [22]. In other words, the prior on $J(\eta)$ is a normal distribution with mean zero and variance $(2\lambda_n^2)^{-1}$. Hence it is natural to expect λ_n has some effect on the convergence rate of η. Other possible priors on the functional parameter include Dirichlet and Gaussian processes which are more commonly used in nonparametric Bayesian methodology.

6. Examples (Continued). We now illustrate verification of the assumptions in section 3.3 with the two example that were introduced in section 2. Thus this section is a continuation of the earlier examples.

6.1. *Partly Linear Normal Model with Current Status Data.* We will concentrate on the estimation of the regression coefficient θ, considering the infinite dimensional parameter $f \in \mathcal{H}_k^M$ as a nuisance parameter. The score function of θ, $\dot{\ell}_{\theta,f}$, is given as follows:

$$\dot{\ell}_{\theta,f}(x) = uQ(x;\theta,f),$$

where

$$Q(X;\theta,f) = (1-\Delta)\frac{\phi(q_{\theta,f}(X))}{1-\Phi(q_{\theta,f}(X))} - \Delta\frac{\phi(q_{\theta,f}(X))}{\Phi(q_{\theta,f}(X))},$$

$q_{\theta,f}(x) = c - \theta u - f(v)$, and ϕ is the density of a standard normal random variable. The least favorable direction at the true parameter value is:

$$h_0(v) = \frac{E_0(UQ^2(X;\theta,f)|V=v)}{E_0(Q^2(X;\theta,f)|V=v)},$$

where E_0 is the expectation relative to the true parameters. The derivation of $\dot{\ell}_{\theta,f}$ and $h_0(\cdot)$ is given in [3]. Thus, the least favorable submodel can be constructed as follows:

$$\ell(t,\theta,f) = \log lik(t, f_t(\theta,f)), \tag{43}$$

where $f_t(\theta,f) = f + (\theta - t)h_0$. By differentiating (43) with respect to t or θ, we can obtain the maps assessed in assumption S1, $(t,\theta,f) \mapsto (\partial^{l+m}/\partial t^l \partial\theta^m)\ell(t,\theta,f)$. The concrete forms of these maps are given in [3] which considers a more rigid model with a known upper bound on the L_2 norm of the kth derivative. The rate assumptions (9) and (10) have been verified previously in corollary 1. The remaining assumptions are verified in the following two lemmas:

LEMMA 2. *Under the above set-up for the partly linear normal model with current status data, assumptions S1, S2 and E1 are satisfied.*

LEMMA 3. *Under the above set-up for the partly linear normal model with current status data, condition (39) is satisfied.*

6.2. *Semiparametric Logistic Regression.* In the semiparametric logistic regression model, we can obtain the score function for θ and η by similar analysis performed in the first example, i.e. $\dot{\ell}_{\theta,\eta}(x) = (y - F(\theta w + \eta(z)))w$ and $A_{\theta,\eta}h_{\theta,\eta}(x) = (y - F(\theta w + \eta(z)))h_{\theta,\eta}(z)$ for $J(h) < \infty$. And the least favorable direction at the true parameter is given in [14]:

$$h_0(z) = \frac{P_0[W\dot{F}(\theta_0 W + \eta_0(Z))|Z=z]}{P_0[\dot{F}(\theta_0 W + \eta_0(Z))|Z=z]},$$

where $\dot{F}(u) = F(u)(1-F(u))$. The above assumptions plus the requirement that $J(h_0) < \infty$ ensures the identifiability of the parameters. Thus the least favorable submodel can be written as:

$$\ell(t,\theta,\eta) = \log lik(t,\eta_t(\theta,\eta)),$$

where $\eta_t(\theta,\eta) = \eta + (\theta - t)h_0$. By differentiating $\ell(t,\theta,\eta)$ with respect to t or θ, we obtain,

$$\begin{aligned}
\dot{\ell}(t,\theta,\eta) &= (y - F(tw + \eta(z) + (\theta - t)h_0(z)))(w - h_0(z)),\\
\ddot{\ell}(t,\theta,\eta) &= -\dot{F}(tw + \eta(z) + (\theta - t)h_0(z))(w - h_0(z))^2,\\
\ell_{t,\theta}(t,\theta,\eta) &= -\dot{F}(tw + \eta(z) + (\theta - t)h_0(z))(w - h_0(z))h_0(z),\\
\ell^{(3)}(t,\theta,\eta) &= -\ddot{F}(tw + \eta(z) + (\theta - t)h_0(z))(w - h_0(z))^3,\\
\ell_{t,t,\theta}(t,\theta,\eta) &= -\ddot{F}(tw + \eta(z) + (\theta - t)h_0(z))(w - h_0(z))^2 h_0(z),\\
\ell_{t,\theta,\theta}(t,\theta,\eta) &= -\ddot{F}(tw + \eta(z) + (\theta - t)h_0(z))(w - h_0(z))h_0^2(z),
\end{aligned}$$

where $\ddot{F}(\cdot)$ is the second derivative of the function $F(\cdot)$. The rate assumptions have been shown in corollary 2. The remaining assumptions are verified in the following two lemmas:

LEMMA 4. *Under the above set-up for the semiparametric logistic regression model, assumptions S1, S2 and E1 are satisfied.*

LEMMA 5. *Under the above set-up for the semiparametric logistic regression model, condition (39) is satisfied.*

7. Future Work. Our paper evaluates the penalized profile sampler method from the frequentist view and discusses the effect of the smoothing parameter on estimation accuracy. One potential problem of interest is how to select a proper smoothing parameter in applications. A formal study about the higher order comparisons between the profile sampler procedure and fully Bayesian procedure [16], which assign priors to both the finite dimensional parameter and the infinite dimensional nuisance parameter, is also interesting. We expect that the involvement of a suitable prior on the infinite dimensional parameter would at least not decrease the estimation accuracy of the parameter of interest.

Another worthwhile avenue of research is to develop analogs of the profile sampler and penalized profile sampler to likelihood estimation under model misspecification and to general M-estimation. Some first order results for this setting in the case where the nuisance parameter may not be root-n consistent have been developed for a weighted bootstrap procedure in [10].

8. Appendix. We first present some technical tools about the entropy calculations and increments of empirical processes which will be employed in the proofs that follow.

T1. For each $0 < C < \infty$ and $\delta > 0$ we have

$$H_B(\delta, \{\eta : \|\eta\|_\infty \leq C, J(\eta) \leq C\}, \|\cdot\|_\infty) \lesssim (\frac{C}{\delta})^{1/k}, \tag{44}$$

$$H(\delta, \{\eta : \|\eta\|_\infty \leq C, J(\eta) \leq C\}, \|\cdot\|_\infty) \lesssim (\frac{C}{\delta})^{1/k}. \tag{45}$$

T2. Let $\mathcal{F}$ be a class of measurable functions such that $Pf^2 < \delta^2$ and $\|f\|_\infty \leq M$ for every f in $\mathcal{F}$. Then

$$E_P^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim K(\delta, \mathcal{F}, L_2(P)) \left(1 + \frac{K(\delta, \mathcal{F}, L_2(P))}{\delta^2\sqrt{n}} M\right),$$

where $K(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + H_B(\epsilon, \mathcal{F}, \|\cdot\|)} d\epsilon$.

T3. Let $\mathcal{F} = \{f_t : t \in T\}$ be a class of functions satisfying $|f_s(x) - f_t(x)| \leq d(s,t)F(x)$ for every s and t and some fixed function F. Then, for any norm

$\|\cdot\|$,

$$N_{[]}(2\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\epsilon, T, d).$$

T4.

$$- P_{\theta_0} \log \frac{p_\theta}{p_{\theta_0}} \geq \int \left(\sqrt{p_\theta} - \sqrt{p_{\theta_0}}\right)^2 d\mu. \tag{46}$$

T5. Let $\mathcal{F}$ be a class of measurable functions $f : \mathbb{D} \times \mathbb{W} \mapsto \mathbb{R}$ on a product of a finite set and an arbitrary measurable space $(\mathbb{W}, \mathcal{W})$. Let P be a probability measure on $\mathbb{D} \times \mathbb{W}$ and let $P_{\mathcal{W}}$ be its marginal on $\mathbb{W}$. For every $d \in \mathbb{D}$, let $\mathcal{F}_d$ be the set of functions $w \mapsto f(d, w)$ as f ranges over $\mathcal{F}$. If every class $\mathcal{F}_d$ is P-Donsker with $\sup_{f\in\mathcal{F}} |Pf(d, W)| < \infty$ for every d, then $\mathcal{F}$ is P-Donsker.

T6. Let $\mathcal{F}$ be a uniformly bounded class of measurable functions such that for some measurable f_0, $\sup_{f\in\mathcal{F}} \|f - f_0\|_\infty < \infty$. Moreover, assume that $H_B(\epsilon, \mathcal{F}, L_2(P)) \leq K\epsilon^{-\alpha}$ for some $K < \infty$ and $\alpha \in (0, 2)$ and for all $\epsilon > 0$. Then

$$\sup_{f\in\mathcal{F}} \left[\frac{|(\mathbb{P}_n - P)(f - f_0)|}{\|f - f_0\|_2^{1-\alpha/2} \vee n^{(\alpha-2)/[2(2+\alpha)]}}\right] = O_P(n^{-1/2}).$$

T7. For a probability measure P, let $\mathcal{F}_1$ be a class of measurable functions $f_1 : \mathcal{X} \mapsto \mathbb{R}$, and let $\mathcal{F}_2$ denote a class of nondecreasing functions $f_2 : \mathbb{R} \mapsto [0, 1]$ that are measurable for every probability measure. Then,

$$H_B(\epsilon, \mathcal{F}_2(\mathcal{F}_1), L_2(P)) \leq 2H_B(\epsilon/3, \mathcal{F}_1, L_2(P)) + \sup_Q H_B(\epsilon/3, \mathcal{F}_2, L_2(Q)).$$

T8. Let $\mathcal{F}$ and $\mathcal{G}$ be classes of measurable functions. Then for any probability measure Q and any $1 \leq r \leq \infty$,

$$H_B(2\epsilon, \mathcal{F} + \mathcal{G}, L_r(Q)) \leq H_B(\epsilon, \mathcal{F}, L_r(Q)) + H_B(\epsilon, \mathcal{G}, L_r(Q)), \tag{47}$$

and, provided $\mathcal{F}$ and $\mathcal{G}$ are bounded by 1,

$$H_B(2\epsilon, \mathcal{F}\times\mathcal{G}, L_r(Q)) \leq H_B(\epsilon, \mathcal{F}, L_r(Q)) + H_B(\epsilon, \mathcal{G}, L_r(Q)). \tag{48}$$

REMARK 10. *The proof of T1 is found in [1]. T1 implies that the Sobolev class of functions with known bounded Sobolev norm is P-Donsker. T2 and T3 are separately lemma 3.4.2 and theorem 2.7.11 in [21]. (46) in T4 relates the Kullback-Leibler divergence and Hellinger distance. Its proof depends on the inequality that* $\log x \leq 2(\sqrt{x}-1)$ *for every* $x > 0$*. T5 is lemma 9.2 in [15]. T6 is a result presented on page 79 of [19] and is a special case of lemma 5.13 on the same page, the proof of which can be found in pages 79–80. T7 and T8 are separately lemma 15.2 and 9.24 in [6].*

Proof of theorem 1: The definition of $\hat{\eta}_{\tilde{\theta}_n,\lambda_n}$ implies that

$$\begin{aligned} \lambda_n^2 J^2(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) &\leq \lambda_n^2 J^2(\eta_0) + (\mathbb{P}_n - P)\left(\ell_{\tilde{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n}} - \ell_{\tilde{\theta}_n,\eta_0}\right) \\ &+ P\left(\ell_{\tilde{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n}} - \ell_{\tilde{\theta}_n,\eta_0}\right) \\ &\leq \lambda_n^2 J^2(\eta_0) + I + II. \end{aligned}$$

Note that by T6 and assumption (21), we have

$$\begin{aligned} I &\leq (1 + J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}))O_P(n^{-1/2}) \times \left\{ \left\| \frac{\ell_{\tilde{\theta}_n,\hat{\eta}_{\tilde{\theta}_n,\lambda_n}} - \ell_0}{1 + J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n})} \right\|_2^{1-\frac{1}{2k}} \vee n^{-\frac{2k-1}{2(2k+1)}} \right\} \\ &\quad + (1 + J(\eta_0))O_P(n^{-1/2}) \times \left\{ \left\| \frac{\ell_{\tilde{\theta}_n,\eta_0} - \ell_0}{1 + J(\eta_0)} \right\|_2^{1-\frac{1}{2k}} \vee n^{-\frac{2k-1}{2(2k+1)}} \right\}. \end{aligned}$$

By assumption (24), we have

$$II \lesssim -d_{\tilde{\theta}_n}^2(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}, \eta_0) + \|\tilde{\theta}_n - \theta_0\|^2.$$

Combining with the above, we can deduce that

$$\begin{aligned}\hat{d}_n^2+\lambda_n^2\hat{J}_n^2 &\lesssim (1+\hat{J}_n)O_P(n^{-1/2})\times\left\{\left(\frac{\hat{d}_n+\|\tilde{\theta}_n-\theta_0\|}{1+\hat{J}_n}\right)^{1-\frac{1}{2k}}\vee n^{-\frac{2k-1}{2(2k+1)}}\right\}\\ &+ (1+J_0)O_P(n^{-1/2})\times\left\{\left(\frac{\|\tilde{\theta}_n-\theta_0\|}{1+J_0}\right)^{1-\frac{1}{2k}}\vee n^{-\frac{2k-1}{2(2k+1)}}\right\}\\ &+ \lambda_n^2J_0^2+\|\tilde{\theta}_n-\theta_0\|^2, \qquad (49)\end{aligned}$$

where $\hat{d}_n=d_{\tilde{\theta}_n}(\hat{\eta}_{\tilde{\theta}_n,\lambda_n},\eta_0)$, $J(\eta_0)=J_0$ and $\hat{J}_n=J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n})$. The above inequality follows from assumption (23). Combining all of the above inequalities, we can deduce that

$$u_n^2=O_P(1)+O_P(1)u_n^{1-\frac{1}{2k}}, \qquad (50)$$

$$v_n=v_n^{-1}O_P(\|\tilde{\theta}_n-\theta_0\|^2)+u_n^{1-\frac{1}{2k}}O_P(\lambda_n)+O_P(n^{-\frac{1}{2}}\lambda_n^{-1}\|\tilde{\theta}_n-\theta_0\|^{1-\frac{1}{2k}}), \qquad (51)$$

where $u_n=(\hat{d}_n+\|\tilde{\theta}_n-\theta_0\|)/(\lambda_n+\lambda_n\hat{J}_n)$ and $v_n=\lambda_n\hat{J}_n+\lambda_n$. The equation (50) implies that $u_n=O_P(1)$. Inserting $u_n=O_P(1)$ into (51), we can know that $v_n=O_P(\lambda_n+\|\tilde{\theta}_n-\theta_0\|)$, which implies u_n has the desired order. This completes the whole proof. □

Proof of corollary 1: Conditions (22)–(24) can be verified easily in this example based on the arguments in theorem 1 because $\ddot{\ell}_{\theta,f}$ has finite second moment, and $p_{\theta,f}$ is bounded away from zero and infinity uniformly for (θ,f) ranging over the whole parameter space. Note that $d_\theta(f,f_0)=\|p_{\theta,f}-p_0\|_2\gtrsim\|q_{\theta,f}-q_{\theta_0,f_0}\|_2$ by Taylor expansion. Then by the assumption that $EVar(U|V)$ is positive definite, we know that $\|q_{\tilde{\theta}_n,\hat{f}_{\tilde{\theta}_n,\lambda_n}}-q_{\theta_0,f_0}\|_2=O_P(\lambda_n+\|\tilde{\theta}_n-\theta_0\|)$ implies $\|\hat{f}_{\tilde{\theta}_n,\lambda_n}-f_0\|_2=O_P(\lambda_n+\|\tilde{\theta}_n-\theta_0\|)$. Thus we only need to show that the ϵ-bracketing entropy number of the function class $\mathcal{O}$ defined below is of order $\epsilon^{-1/k}$ to complete the proof of (25)–(26):

$$\mathcal{O} \equiv \left\{ \frac{\ell_{\theta,f}(X)}{1+J(f)} : \|\theta-\theta_0\| \leq C_1, \|f-f_0\|_\infty \leq C_1, J(f) < \infty \right\},$$

for some constant C_1. Note that $\ell_{\theta,f}(X)/(1+J(f))$ can be rewritten as:

$$\Delta A^{-1} \log \Phi\left(q_{\bar{\theta},f} A\right) + (1-\Delta) A^{-1} \log\left(1-\Phi\left(q_{\bar{\theta},f} A\right)\right), \tag{52}$$

where $A = 1 + J(f)$ and $\bar{q}_{\theta,f} \in \mathcal{O}_1$, where

$$\mathcal{O}_1 \equiv \left\{ \frac{q_{\theta,f}(X)}{1+J(f)} : \|\theta-\theta_0\| \leq C_1, \|f-f_0\|_\infty \leq C_1, J(f) < \infty \right\},$$

and where we know $H_B(\epsilon, \mathcal{O}_1, L_2(P)) \lesssim \epsilon^{-1/k}$ by T1.

We next calculate the ϵ-bracketing entropy number with L_2 norm for the class of functions $R_1 \equiv \{k_a(t) : t \mapsto a^{-1} \log \Phi(at)$ for $a \geq 1$ and $t \in \mathbb{R}\}$. By some analysis we know that $k_a(t)$ is strictly decreasing in a for $t \in \mathbb{R}$, and $\sup_{t \in \mathbb{R}} |k_a(t) - k_b(t)| \lesssim |a-b|$ because $|\partial/\partial a(k_a(t))|$ is bounded uniformly over $t \in \mathbb{R}$. In addition, we know that $\sup_{a,b \geq A_0, t \in \mathbb{R}} |k_a(t) - k_b(t)| \lesssim A_0^{-1}$ because the function $u \mapsto u \log \Phi(u^{-1}t)$ has bounded derivative for $0 < u \leq 1$ uniformly over $t \in \mathbb{R}$. The above two inequalities imply that the ϵ-bracketing number with uniform norm is of order $O(\epsilon^{-2})$ for $a \in [1, \epsilon^{-1}]$ and is 1 for $a > \epsilon^{-1}$. Thus we know $H_B(\epsilon, R_1, L_2) = O(\log \epsilon^{-2})$. By applying a similar analysis to $R_2 \equiv \{k_a(t) : t \mapsto a^{-1} \log(1-\Phi(at))$ for $a \geq 1$ and $t \in \mathbb{R}\}$, we obtain that $H_B(\epsilon, R_2, L_2) = O(\log \epsilon^{-2})$. Combining this with T7 and T8, we deduce that $H_B(\epsilon, \mathcal{O}, L_2) \lesssim \epsilon^{-1/k}$. This completes the proof of (25)–(26).

For the proof of (27), we apply arguments similar to those used in the proof of theorem 1 but after setting λ_n, J_0 and $\hat{J}_n$ to zero in (49). Then we obtain the following equality: $\hat{d}_n^2 = O_P(n^{-2k/(2k+1)}) + \|\tilde{\theta}_n - \theta_0\|^2 + O_P(n^{-1/2}) \|\tilde{\theta}_n - \theta_0\|^{1-1/2k} + O_P(n^{-1/2})(\|\tilde{\theta}_n - \theta_0\| + \hat{d}_n)^{1-1/2k}$. By treating

$\|\tilde{\theta}_n - \theta_0\| \le n^{-k/(2k+1)}$ and $\|\tilde{\theta}_n - \theta_0\| > n^{-k/(2k+1)}$ differently in the above equality, we obtain (27).□

Proof of corollary 2: Lemma 7.1 in [14] establishes that

$$\left\| p_{\tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n, \lambda_n}} - p_{\theta_0, \eta_0} \right\|_2 + \lambda_n J(\hat{\eta}_{\tilde{\theta}_n, \lambda_n}) = O_P(\lambda_n + \|\tilde{\theta}_n - \theta_0\|) \tag{53}$$

after choosing

$$m_{\theta,\lambda,\eta} = \log \frac{p_{\theta,\eta} + p_{\theta,\eta_0}}{2 p_{\theta,\eta_0}} - \frac{1}{2}\lambda^2 (J^2(\eta) - J^2(\eta_0))$$

in theorem 2. Note that the map $\theta \mapsto p_{\theta,\eta_0}/f^{W,Z}(w,z)$ is uniformly bounded away from zero at $\theta = \theta_0$ and continuous around a neighborhood of θ_0. Hence $m_{\theta,\lambda,\eta}$ is well defined. Moreover, $\mathbb{P}_n m_{\theta,\lambda,\hat{\eta}_{\theta,\lambda}} \ge \mathbb{P}_n m_{\theta,\lambda,\eta_0}$ by the inequality that $((p_{\theta,\eta} + p_{\theta,\eta_0})/2p_{\theta,\eta_0})^2 \ge (p_{\theta,\eta}/p_{\theta,\eta_0})$. (53) now directly implies (32). For the proof of (31), we need to consider the conclusion of lemma 7.4 (i), which states that

$$\|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2 \gtrsim (\|\theta - \theta_0\| \wedge 1 + \| |\eta - \eta_0| \wedge 1\|_2) \wedge 1. \tag{54}$$

Thus we have proved (31). For (33), we just replace the $m_{\theta,\lambda,\eta}$ with $m_{\theta,0,\eta}$ in the proof of lemma 7.1 in [14]. Thus we can show that $d_\theta(\eta, \eta_0) = \|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2$. By combining lemma 1 and (54), we know that $\|\hat{\eta}_{\tilde{\theta}_n} - \eta_0\|_2 = O_P(\delta_n + \|\tilde{\theta}_n - \theta_0\|)$, for δ_n satisfying $K(\delta_n, \mathcal{S}_{\delta_n}, L_2(P)) \le \sqrt{n}\delta_n^2$. Note that $K(\delta, \mathcal{S}_\delta, L_2(P))$ is as defined in (30). By similar analysis as used in the proof of lemma 7.1 in [14] and the strengthened assumption on η, we then find that $K(\delta_n, \mathcal{S}_{\delta_n}, L_2(P)) \lesssim \delta_n^{1-1/2k}$, which leads to the desired convergence rate given in (33). □

Proof of theorem 3. Note that

$$\begin{aligned} Pl^{\lambda_n}_{\tilde\theta_n,\hat\eta_{\tilde\theta_n,\lambda_n},h} &- Pl^{\lambda_n}_{\theta_0,\eta_0,h} \\ &= V(\tilde\theta_n,\hat\eta_{\tilde\theta_n,\lambda_n})h - \mathbb{P}_n l^{\lambda_n}_{\tilde\theta_n,\hat\eta_{\tilde\theta_n,\lambda_n},h} - 2\lambda_n^2\int h^{(k)}(\hat\eta^{(k)}_{\tilde\theta_n,\lambda_n} - \eta_0^{(k)})dz \\ &= -(\mathbb{V}_n - V)(\tilde\theta_n,\hat\eta_{\tilde\theta_n,\lambda_n})h + 2\lambda_n^2\int h^{(k)}\eta_0^{(k)}dz \\ &= -(\mathbb{V}_n - V)(\theta_0,\eta_0)h + o_P^*(n^{-1/2}) + 2\lambda_n^2\int h^{(k)}\eta_0^{(k)}dz \\ &= O_P(n^{-1/2}) + 2\lambda_n^2\int h^{(k)}\eta_0^{(k)}dz. \end{aligned}$$

The last two equalities in the above follow from assumptions (34) and (35). The Fréchet differentiability of $V(\cdot,\cdot)$ at (θ_0,η_0) establishes that

$$\begin{aligned} Pl^{\lambda_n}_{\tilde\theta_n,\hat\eta_{\tilde\theta_n,\lambda_n},h} &- Pl^{\lambda_n}_{\theta_0,\eta_0,h} \\ &= \dot V(\tilde\theta_n - \theta_0, \hat\eta_{\tilde\theta_n,\lambda_n} - \eta_0) + o_P^*(\|\tilde\theta_n - \theta_0\| + d_{\tilde\theta_n}(\hat\eta_{\tilde\theta_n,\lambda_n},\eta_0)) \\ &\quad -2\lambda_n^2\int h^{(k)}(\hat\eta^{(k)}_{\tilde\theta_n,\lambda_n} - \eta_0^{(k)})dz. \end{aligned}$$

Combining the above two sets of equations, we have, by the linearity of $\dot V(\cdot,\cdot)$, established that

$$\dot V(0,\hat\eta_{\tilde\theta_n,\lambda_n}) = O_P(n^{-1/2}) + O_P(\|\tilde\theta_n - \theta_0\|) + 2\lambda_n^2\int_{\mathcal{Z}} h^{(k)}\hat\eta^{(k)}_{\tilde\theta_n,\lambda_n}dz.$$

Now by the invertibility of $\dot V(0,\cdot)$, we can deduce that $d_{\tilde\theta_n}(\hat\eta_{\tilde\theta_n,\lambda_n},\eta_0) = O_P(n^{-1/2} + \|\tilde\theta_n - \theta_0\| + \lambda_n^2 J^2(\hat\eta_{\tilde\theta_n,\lambda_n}))$. □

Proof of theorem 4. We first show (37), and then we need to state one lemma before proceeding to the proof of (38). For the proof of (37), note that

$$0 = \mathbb{P}_n\dot\ell(\hat\theta_{\lambda_n},\hat\theta_{\lambda_n},\hat\eta_{\lambda_n}) + 2\lambda_n^2\int_{\mathcal{Z}}\hat\eta^{(k)}_{\lambda_n}(z)h_0^{(k)}(z)dz.$$

Combining the third order Taylor expansion of $\hat\theta_{\lambda_n}\mapsto\mathbb{P}_n\dot\ell(\hat\theta_{\lambda_n},\theta,\eta)$ around θ_0, where $\theta=\hat\theta_{\lambda_n}$ and $\eta=\hat\eta_{\lambda_n}$, with conditions (19) and (20), the first term in the right-hand-side of the above displayed equality equals $\mathbb{P}_n\tilde{\ell}_0 - \tilde{I}_0(\hat{\theta}_{\lambda_n} - \theta_0) + O_P(\lambda_n + \|\hat{\theta}_{\lambda_n} - \theta_0\|)^2$. By the inequality $2\lambda_n^2 \int_{\mathcal{Z}} \hat{\eta}_{\lambda_n}^{(k)}(z) h_0^{(k)}(z) dz \leq \lambda_n^2 (J^2(\hat{\eta}_{\lambda_n}) + J^2(h_0))$ and assumption (10), the second term in the right-hand-side of the above equality is equal to $O_P(\lambda_n + \|\hat{\theta}_{\lambda_n} - \theta_0\|)^2$. Combining everything, we obtain the following:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}_0^{-1} \tilde{\ell}_0(X_i) = \sqrt{n}(\hat{\theta}_{\lambda_n} - \theta_0) + O_P(n^{1/2}(\lambda_n + \|\hat{\theta}_{\lambda_n} - \theta_0\|)^2). \tag{55}$$

The right-hand-side of (55) is of the order $O_P(\sqrt{n}\lambda_n^2 + \sqrt{n} w_n (1 + w_n + \lambda_n))$, where w_n represents $\|\hat{\theta}_{\lambda_n} - \theta_0\|$. However, its left-hand-side is trivially $O_P(1)$. Considering the fact that $\sqrt{n}\lambda_n^2 = o_P(1)$, we can deduce that $\hat{\theta}_{\lambda_n} - \theta_0 = O_P(n^{-1/2})$. Inserting this into the previous display completes the proof of (37).

We next prove (38). Note that $\hat{\theta}_{\lambda_n} - \theta_0 = O_P(n^{-1/2})$. Hence the order of the remainder terms in (19) and (20) becomes $O_P(\lambda_n + \|\tilde{\theta}_n - \hat{\theta}_{\lambda_n}\|)^2$ and $O_P(\lambda_n + \|\tilde{\theta}_n - \hat{\theta}_{\lambda_n}\|)$, respectively. Expression (61) in lemma 6 below implies that

$$\begin{aligned} logpl_{\lambda_n}(\hat{\theta}_{\lambda_n}) &= \log pl_{\lambda_n}(\theta_0) + n(\hat{\theta}_{\lambda_n} - \theta_0)^T \mathbb{P}_n \tilde{\ell}_0 \\ &- \frac{n}{2}(\hat{\theta}_{\lambda_n} - \theta_0)^T \tilde{I}_0 (\hat{\theta}_{\lambda_n} - \theta_0) + O_P(n^{1/2}\lambda_n^2). \end{aligned} \tag{56}$$

The difference between (56) and (61) generates

$$\begin{aligned} \log pl_{\lambda_n}(\tilde{\theta}_n) &= \log pl_{\lambda_n}(\hat{\theta}_{\lambda_n}) + n(\tilde{\theta}_n - \hat{\theta}_{\lambda_n})^T \left(\mathbb{P}_n \tilde{\ell}_0 - \tilde{I}_0(\hat{\theta}_{\lambda_n} - \theta_0)\right) \\ &- \frac{n}{2}(\tilde{\theta}_n - \hat{\theta}_{\lambda_n})^T \tilde{I}_0 (\tilde{\theta}_n - \hat{\theta}_{\lambda_n}) + O_P(g_{\lambda_n}(\|\tilde{\theta}_n - \hat{\theta}_{\lambda_n}\|)). \end{aligned}$$

(38) is now immediately obtained after considering (37). □

Proof of theorem 5. Suppose that $F_{\lambda_n}(\cdot)$ is the penalized posterior profile distribution of $\sqrt{n}\varrho_n$ with respect to the prior $\rho(\theta)$, where the vector ϱ_n

is defined as $\tilde{I}_0^{1/2}(\theta - \hat{\theta}_n)$. The parameter set for ϱ_n is Ξ_n. $F_{\lambda_n}(\cdot)$ can be expressed as:

$$(57) \quad F_{\lambda_n}(\xi) = \frac{\int_{\varrho_n \in (-\infty, n^{-1/2}\xi] \cap \Xi_n} \rho(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n) \frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n)}{pl_{\lambda_n}(\hat{\theta}_{\lambda_n})} d\varrho_n}{\int_{\varrho_n \in \Xi_n} \rho(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n) \frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n)}{pl_{\lambda_n}(\hat{\theta}_{\lambda_n})} d\varrho_n}.$$

Note that $d\varrho_n$ in the above is the short notation for $d\varrho_{n1} \times \ldots \times d\varrho_{nd}$. To prove theorem 5, we first partition the parameter set Ξ_n as $\{\Xi_n \cap \{\|\varrho_n\|_2 > r_n\}\} \cup \{\Xi_n \cap \{\|\varrho_n\|_2 \leq r_n\}\}$. By choosing the proper order of r_n, we find the posterior mass in the first partition region is of arbitrarily small order, as verified in lemma 5.1 immediately below, and the mass inside the second partition region can be approximated by a stochastic polynomial in powers of $n^{-1/2}$ with error of order dependent on the smoothing parameter, as verified in lemma 5.2 below. This basic technique applies to both the denominator and the numerator, yielding the quotient series, which gives the desired result.

lemma 5.1. Choose $r_n = o(n^{-1/3})$ and $\sqrt{n} r_n \to \infty$. Under the conditions of theorem 5, we have

$$(58) \quad \int_{\|\varrho_n\| > r_n} \rho(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n) \frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n)}{pl_{\lambda_n}(\hat{\theta}_{\lambda_n})} d\varrho_n = O_P(n^{-M}),$$

for any positive number M.

Proof: Fix $r > 0$. We then have

$$\int_{\|\varrho_n\| > r} \rho(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n) \frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n)}{pl_{\lambda_n}(\hat{\theta}_{\lambda_n})} d\varrho_n$$
$$\leq \quad I\{\Delta_{\lambda_n}^r < -n^{-\frac{1}{2}}\} \exp(-\sqrt{n}) \int_{\Theta} \rho(\theta) d\theta + I\{\Delta_{\lambda_n}^r \geq -n^{-\frac{1}{2}}\},$$

where $\Delta_{\lambda_n}^r = sup_{\|\varrho_n\| > r} \Delta_{\lambda_n}(\hat{\theta}_{\lambda_n} + \varrho_n \tilde{I}_0^{-1/2})$. Then by lemma 3.2 in [2], $I\{\Delta_{\lambda_n}^r \geq -n^{-\frac{1}{2}}\} = O_P(n^{-M})$ for any fixed $r > 0$. This implies that there exists a positive decreasing sequence $r_n = o(n^{-1/3})$ with $\sqrt{n} r_n \to \infty$ such that (58) holds. □

lemma 5.2. Choose $r_n = o(n^{-1/3})$ and $\sqrt{n} r_n \to \infty$. Under the conditions of theorem 5, we have

$$\int_{\|\varrho_n\| \le r_n} \left| \frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n)}{pl_{\lambda_n}(\hat{\theta})} \rho(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n) - \exp\left(-\frac{n}{2} \varrho_n^T \varrho_n\right) \rho(\hat{\theta}_{\lambda_n}) \right| \times d\varrho_n = O_P(\lambda_n^2). \tag{59}$$

Proof: The posterior mass over the region $\|\varrho_n\|_2 \le r_n$ is bounded by

$$\int_{\|\varrho_n\|_2 \le r_n} \left| \frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n)}{pl_{\lambda_n}(\hat{\theta}_{\lambda_n})} \rho(\hat{\theta}_{\lambda_n}) - \exp\left(-\frac{n}{2} \varrho_n^T \varrho_n\right) \rho(\hat{\theta}_{\lambda_n}) \right| d\varrho_n \quad (*)$$

$$+ \int_{\|\varrho_n\|_2 \le r_n} \left| \frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n)}{pl_{\lambda_n}(\hat{\theta}_{\lambda_n})} \rho(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n) - \frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n} + \tilde{I}_0^{-\frac{1}{2}} \varrho_n)}{pl_{\lambda_n}(\hat{\theta}_{\lambda_n})} \rho(\hat{\theta}_{\lambda_n}) \right| d\varrho_n. \quad (**)$$

By (38), we obtain

$$(*) \quad = \quad \int_{\|\varrho_n\|_2 \le r_n} \left[\rho(\hat{\theta}_{\lambda_n}) \exp\left(-\frac{n \varrho_n^T \varrho_n}{2}\right) |\exp(O_P(g_{\lambda_n}(\|\varrho_n\|))) - 1| \right] d\varrho_n.$$

Obviously the order of $(*)$ depends on that of $|\exp(O_P(g_{\lambda_n}(\|\varrho_n\|))) - 1|$ for λ_n satisfying (3) and $\|\varrho_n\| \le r_n$. In order to analyze its order, we partition the set $\{\lambda_n = o_P(n^{-1/4})$ and $\lambda_n^{-1} = O_P(n^{k/(2k+1)})\}$ with the set $\{\lambda_n = O_P(n^{-1/3})\}$, i.e. $U_n = \{\lambda_n = o_P(n^{-1/4})$ and $\lambda_n^{-1} = O_P(n^{k/(2k+1)})\} \cap \{\lambda_n = O_P(n^{-1/3})\}$ and $L_n = \{\lambda_n = o_P(n^{-1/4})$ and $\lambda_n^{-1} = O_P(n^{k/(2k+1)})\} \cap \{\lambda_n = O_P(n^{-1/3})\}^C$. For the set U_n, we have $|\exp(O_P(g_{\lambda_n}(\|\varrho_n\|))) - 1| = g_{\lambda_n}(\|\varrho_n\|) \times O_P(1)$. For the set L_n, we have $O_P(g_{\lambda_n}(\|\varrho_n\|)) = O_P(n\|\varrho_n\|\lambda_n^2 + n^{1/2}\lambda_n^2)$. We can take $r_n = n^{-1-\delta}\lambda_n^{-2}$ for some $\delta > 0$ such that $\sqrt{n} r_n \to \infty$ and $r_n = o(n^{-1/3})$. Then $|\exp(O_P(g_{\lambda_n}(\|\varrho_n\|))) - 1| = (n\|\varrho_n\|\lambda_n^2 + n^{1/2}\lambda_n^2) \times O_P(1)$. Combining with the above, we know that $(*) = O_P(\lambda_n^2)$. By similar analysis, we can also show that $(**)$ has the same order. This completes the proof of lemma 5.2. □

We next start the formal proof of theorem 5. By considering both lemma 5.1 and lemma 5.2, we know the denominator of (57) equals

$$\int_{\{\|\varrho_n\|_2\leq r_n\}\cap\Xi_n}\left[\exp\left(-\frac{n}{2}\varrho_n^T\varrho_n\right)\rho(\hat{\theta}_{\lambda_n})\right]d\varrho_n+O_P(\lambda_n^2).$$

The first term in the above display equals

$$\begin{aligned} n^{-1/2}\rho(\hat{\theta}_{\lambda_n})\int_{\{\|u_n\|_2\leq\sqrt{n}r_n\}\cap\sqrt{n}\Xi_n}e^{-u_n^Tu_n/2}du_n &= n^{-1/2}\rho(\hat{\theta}_{\lambda_n})\int_{\mathbb{R}^d}e^{-u_n^Tu_n/2}du_n \\ &+ O(\lambda_n^2), \end{aligned}$$

where $u_n=\sqrt{n}\varrho_n$. The above equality follows from the inequality that $\int_x^\infty e^{-y^2/2}dy\leq x^{-1}e^{-x^2/2}$ for any $x>0$. Consolidating the above analyses, we deduce that the denominator of (57) equals $n^{-\frac{1}{2}}\rho(\hat{\theta}_{\lambda_n})(2\pi)^{d/2}+O_P(\lambda_n^2)$. The same analysis also applies to the numerator, thus completing the whole proof. □

Proof of corollary 3: We only show (41) in what follows. (42) can be verified similarly. Showing (41) is equivalent to establishing $\tilde{E}_{\theta|x}^{\lambda_n}(\varrho_n)=O_P(\lambda_n^2)$. Note that $\tilde{E}_{\theta|x}^{\lambda_n}(\varrho_n)$ can be written as:

$$\tilde{E}_{\theta|x}^{\lambda_n}(\varrho_n)=\frac{\int_{\varrho_n\in\Xi_n}\varrho_n\rho(\hat{\theta}_{\lambda_n}+\tilde{I}_0^{-\frac{1}{2}}\varrho_n)\frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n}+\tilde{I}_0^{-\frac{1}{2}}\varrho_n)}{pl_{\lambda_n}(\hat{\theta}_{\lambda_n})}d\varrho_n}{\int_{\varrho_n\in\Xi_n}\rho(\hat{\theta}_{\lambda_n}+\tilde{I}_0^{-\frac{1}{2}}\varrho_n)\frac{pl_{\lambda_n}(\hat{\theta}_{\lambda_n}+\tilde{I}_0^{-\frac{1}{2}}\varrho_n)}{pl_{\lambda_n}(\hat{\theta}_{\lambda_n})}d\varrho_n}.$$

By analysis similar to that applied in the proof of theorem 5, we know the denominator in the above display is $n^{-1/2}(2\pi)^{d/2}\rho(\hat{\theta}_{\lambda_n})+O_P(\lambda_n^2)$ and the numerator is a random vector of order $O_P(n^{-1/2}\lambda_n^2)$. This yields the conclusion.□

Proof of theorem 6. Note that (40) implies $\kappa_{n\alpha} = \tilde{I}_0^{-1/2} z_\alpha + O_P(n^{1/2}\lambda_n^2)$, for any $\xi < \alpha < 1-\xi$, where $\xi \in (0, \frac{1}{2})$. Note also that the α-th quantile of a d dimensional standard normal distribution, z_α, is not unique if $d > 1$. The classical Edgeworth expansion implies that $P(n^{-1/2}\sum_{i=1}^n \tilde{I}_0^{-1/2}\tilde{\ell}_0(X_i) \leq z_\alpha + a_n(\alpha)) = \alpha$, where $a_n(\alpha) = O(n^{-1/2})$, for $\xi < \alpha < 1-\xi$. Note that $a_n(\alpha)$ is uniquely determined for each fixed z_α since $\tilde{\ell}_0(X_i)$ has at least one absolutely continuous component. Let $\hat{\kappa}_{n\alpha} = \tilde{I}_0^{-1/2} z_\alpha + (\sqrt{n}(\hat{\theta}_{\lambda_n} - \theta_0) - n^{-1/2}\sum_{i=1}^n \tilde{I}_0^{-1}\tilde{\ell}_0(X_i)) + \tilde{I}_0^{-1/2} a_n(\alpha)$. Then $P(\sqrt{n}(\hat{\theta}_{\lambda_n} - \theta_0) \leq \hat{\kappa}_{n\alpha}) = \alpha$. Combining with (37), we obtain $\hat{\kappa}_{n\alpha} = \kappa_{n\alpha} + O_P(n^{1/2}\lambda_n^2)$. The uniqueness of $\hat{\kappa}_{n\alpha}$ up to order $O_P(n^{1/2}\lambda_n^2)$ follows from that of $a_n(\alpha)$ for each chosen z_α.$\Box$

Proof of lemma 2. Assumptions S1 and S2 are verified in lemma 5 of [3]. For the verifications of the assumption E1, we first show the asymptotic equicontinuity condition (15). Without loss of generality, we assume that λ_n is bounded below by a multiple of $n^{-k/(2k+1)}$ and bounded above by $n^{-1/4}$ in view of (3). Thus

$$P\left(\frac{\dot{\ell}(\theta_0, \theta_0, \hat{f}_{\tilde{\theta}_n,\lambda_n}) - \dot{\ell}_0}{n^{\frac{1}{4k+2}}(\lambda_n + \|\tilde{\theta}_n - \theta_0\|)}\right)^2 \lesssim \frac{\|\hat{f}_{\tilde{\theta}_n,\lambda_n} - f_0\|_2^2}{n^{\frac{1}{2k+1}}(\lambda_n + \|\tilde{\theta}_n - \theta_0\|)^2} = O_P\left(n^{-\frac{1}{2k+1}}\right),$$

where (25) implies the equality in the above expression.

By (26), we know that $J(\hat{f}_{\tilde{\theta}_n,\lambda_n}) = O_P(1 + \|\tilde{\theta}_n - \theta_0\|/\lambda_n)$ and $\|\hat{f}_{\tilde{\theta}_n,\lambda_n}\|_\infty$ is bounded by some constant, since $f \in \mathcal{H}_k^M$. We then define the set $\mathcal{Q}_n$ as follows:

$$\left\{\frac{\dot{\ell}(\theta_0, \theta_0, f) - \dot{\ell}_0}{n^{\frac{1}{4k+2}}(\lambda_n + \|\theta - \theta_0\|)} : J(f) \leq C_n(1 + \frac{\|\theta - \theta_0\|}{\lambda_n}), \|f\|_\infty \leq M, \|\theta - \theta_0\| \leq \delta\right\}$$
$$\cap \left\{g \in L_2(P) : Pg^2 \leq C_n n^{-\frac{1}{2k+1}}\right\},$$

for some $\delta > 0$. Obviously the function $n^{-1/(4k+2)}(\dot{\ell}(\theta_0, \theta_0, \hat{f}_{\tilde{\theta}_n,\lambda_n}) - \dot{\ell}_0)/(\lambda_n + \|\tilde{\theta}_n - \theta_0\|)) \in \mathcal{Q}_n$ on a set of probability arbitrarily close to one, as $C_n \to \infty$. If we can show $\lim_{n\to\infty} E^*\|\mathbb{G}_n\|_{\mathcal{Q}_n} < \infty$ by T2, then assumption (15) is verified. Note that $\dot{\ell}(\theta_0, \theta_0, f)$ depends on f in a Lipschitz manner. Consequently we can bound $H_B(\epsilon, \mathcal{Q}_n, L_2(P))$ by the product of some constant and $H(\epsilon, \mathcal{R}_n, L_2(P))$ in view of T3. $\mathcal{R}_n$ is defined as

$$\{H_n(f) : J(H_n(f)) \lesssim \lambda_n^{-1} n^{-1/(4k+2)}, \|H_n(f)\|_\infty \lesssim \lambda_n^{-1} n^{-1/(4k+2)}\},$$

where $H_n(f) = f/(n^{1/(4k+2)}(\lambda_n + \|\theta - \theta_0\|))$. By [1], we know that

$$H(\epsilon, \mathcal{R}_n, L_2(P)) \lesssim (\lambda_n^{-1} n^{\frac{-1}{(4k+2)}})/\epsilon)^{1/k}.$$

Note that $\delta_n = n^{-1/(4k+2)}$ and $M_n = n^{(2k-1)/(4k+2)}$ in T2. Thus by calculation we know that $K(\delta_n, \mathcal{Q}_n, L_2(P)) \lesssim \lambda_n^{-1/2k} n^{-1/(4k+2)}$. Then by T2 we can show that $\lim_{n\to\infty} E^*\|\mathbb{G}_n\|_{\mathcal{Q}_n} < \infty$.

We next show (18). It suffices to verify that the sequence of classes of functions $\mathcal{V}_n$ is P-Glivenko-Cantelli, where $\mathcal{V}_n \equiv \{\ell^{(3)}(\bar{\theta}_n, \tilde{\theta}_n, \hat{f}_{\tilde{\theta}_n,\lambda_n})(x)\}$, for every random sequence $\bar{\theta}_n \to \theta_0$ and $\tilde{\theta}_n \to \theta_0$ in probability. A Glivenko-Cantelli theorem for classes of functions that change with n is needed. By revising theorem 2.4.3 in [21] with minor notational changes, we obtain the following suitable extension of the uniform entropy Glivenko-Cantelli theorem: Let $\mathcal{F}_n$ be suitably measurable classes of functions with uniformly integrable functions and $H(\epsilon, \mathcal{F}_n, L_1(\mathbb{P}_n)) = o_P^*(n)$ for any $\epsilon > 0$. Then $\|\mathbb{P}_n - P\|_{\mathcal{F}_n} \to 0$ in probability for every $\epsilon > 0$. We then apply this revised theorem to the set $\mathcal{F}_n$ of functions $\ell^{(3)}(t, \theta, f)$ with t and θ ranging over a neighborhood of θ_0 and $\lambda_n J(f)$ bounded by a constant. By the form of

$\ell^{(3)}(t,\theta,f)$, the entropy number for $\mathcal{V}_n$ is equal to that of

$$\tilde{\mathcal{F}}_n \equiv \{\phi(q_{t,f_t(\theta,f)}(x))R(q_{t,f_t(\theta,f)}(x)) : (t,\theta)\in V_{\theta_0}, \lambda_n J(f)\le C, \|f\|_\infty \le M\}.$$

By arguments similar to those used in lemma 7.2 of [14], we know that $\sup_Q H(\epsilon, \tilde{\mathcal{F}}_n, L_1(Q)) \lesssim (1+\lambda_n^{-1}/\epsilon)^{1/k} = o_P(n)$. Moreover, the $\tilde{\mathcal{F}}_n$ are uniformly bounded since $f\in\mathcal{H}_k^M$. Considering the fact that the probability that $\mathcal{V}_n$ is contained in $\tilde{\mathcal{F}}_n$ tends to 1, we have completed the proof of (18).

For the proof of (16), we only need to show that $\mathbb{G}_n(\ddot{\ell}(\theta_0,\tilde{\theta}_n,\hat{f}_{\tilde{\theta}_n,\lambda_n})-\ddot{\ell}_0) = o_P(1)$ since $\ddot{\ell}_0(x)$ is uniformly bounded in x. Note that we only need to show (16) holds for $\tilde{\theta}_n=\hat{\theta}_n+o(n^{-1/3})$ based on the arguments in lemma 5.2. We next show that $\mathbb{G}_n(\ddot{\ell}(\theta_0,\tilde{\theta}_n,\hat{f}_{\tilde{\theta}_n,\lambda_n})-\ddot{\ell}_0) = o_P(1+n^{1/3}\|\tilde{\theta}_n-\theta_0\|) = o_P(1)$. By the rate assumptions R1, we have

$$P\left(\frac{\ddot{\ell}(\theta_0,\tilde{\theta}_n,\hat{f}_{\tilde{\theta}_n,\lambda_n})-\ddot{\ell}_0}{1+n^{1/3}\|\tilde{\theta}_n-\theta_0\|}\right)^2 \lesssim \frac{\|\tilde{\theta}_n-\theta_0\|^2+\|\hat{f}_{\tilde{\theta}_n,\lambda_n}-f_0\|_2^2}{(1+n^{1/3}\|\tilde{\theta}_n-\theta_0\|)^2} = O_P(n^{-1/2}).$$

We next define $\bar{\mathcal{Q}}_n$ as follows:

$$\left\{\frac{\ddot{\ell}(\theta_0,\theta,f)-\ddot{\ell}_0}{1+n^{1/3}\|\theta-\theta_0\|} : J(f)\le C_n(1+\frac{\|\theta-\theta_0\|}{\lambda_n}), \|f\|_\infty\le M, \|\theta-\theta_0\|<\delta\right\}$$
$$\cap\left\{g\in L_2(P): Pg^2\le C_n n^{-\frac{1}{2}}\right\}.$$

Obviously the function $(\ddot{\ell}(\theta_0,\tilde{\theta}_n,\hat{f}_{\tilde{\theta}_n,\lambda_n})-\ddot{\ell}_0)/(1+n^{1/3}\|\tilde{\theta}_n-\theta_0\|)\in\bar{\mathcal{Q}}_n$ on a set of probability arbitrarily close to one, as $C_n\to\infty$. If we can show $\lim_{n\to\infty}E^*\|\mathbb{G}_n\|_{\bar{\mathcal{Q}}_n}\to 0$ by T2, then the proof of (16) is completed. Accordingly, note that $\ddot{\ell}(\theta_0,\theta,f)$ depends on (θ,f) in a Lipschitz manner. Consequently we can bound $H_B(\epsilon,\bar{\mathcal{Q}}_n,L_2(P))$ by the product of some constant and $(H(\epsilon,\bar{\mathcal{R}}_n,L_2(P))+\log(1/\epsilon))$ in view of T3. $\bar{\mathcal{R}}_n$ is defined as

$$\{H_n(f): J(H_n(f))\lesssim 1+(n^{1/3}\lambda_n)^{-1}, \|H_n(f)\|_\infty\lesssim 1+(n^{1/3}\lambda_n)^{-1}\},$$

where $H_n(f) = f/(1 + n^{1/3}\|\theta - \theta_0\|)$. By [1], we know that

$$H(\epsilon, \bar{\mathcal{R}}_n, L_2(P)) \lesssim ((1 + n^{-1/3}\lambda_n^{-1})/\epsilon)^{1/k}.$$

Then by analysis similar to that used in the proof of (15), we can show that $\lim_{n\to\infty} E^*\|\mathbb{G}_n\|_{\bar{\mathcal{Q}}_n} \to 0$ in view of T2. This completes the proof of (16).

For the proof of (17), it suffices to show that $\mathbb{G}_n(\ell_{t,\theta}(\theta_0, \bar{\theta}_n, \hat{f}_{\tilde{\theta}_n,\lambda_n}) - \ell_{t,\theta}(\theta_0, \theta_0, f_0)) = o_P(1)$ for $\tilde{\theta}_n = \hat{\theta}_n + o(n^{-1/3})$ and for $\bar{\theta}_n$ between $\tilde{\theta}_n$ and θ_0, in view of lemma 5.2. Then we can show that $\mathbb{G}_n(\ell_{t,\theta}(\theta_0, \bar{\theta}_n, \hat{f}_{\tilde{\theta}_n,\lambda_n}) - \ell_{t,\theta}(\theta_0, \theta_0, f_0)) = o_P(1 + n^{1/3}\|\tilde{\theta}_n - \theta_0\|) = o_P(1)$ by similar analysis as used in the proof of (16).□

Proof of lemma 3. By the assumption that $\Delta_{\lambda_n}(\tilde{\theta}_n) = o_P(1)$, we have $\Delta_{\lambda_n}(\tilde{\theta}_n) - \Delta_{\lambda_n}(\theta_0) \geq o_P(1)$. Thus the following inequality holds:

$$n^{-1}\sum_{i=1}^{n} \log\left[\frac{lik(\tilde{\theta}_n, \hat{f}_{\tilde{\theta}_n,\lambda_n}, X_i)}{lik(\theta_0, \hat{f}_{\theta_0,\lambda_n}, X_i)}\right] - n^{-1}\lambda_n^2[J^2(\hat{f}_{\tilde{\theta}_n,\lambda_n}) - J^2(\hat{f}_{\theta_0,\lambda_n})] \geq o_P(1)$$

By considering assumption (10), the above inequality simplifies to

$$n^{-1}\sum_{i=1}^{n} \log\left[\frac{H(\tilde{\theta}_n, \hat{f}_{\tilde{\theta}_n,\lambda_n}; X_i)}{H(\theta_0, \hat{f}_{\theta_0,\lambda_n}; X_i)}\right] \geq o_P(1),$$

where $H(\theta, f; X) = \Delta\Phi(C - \theta U - f(V)) + (1 - \Delta)(1 - \Phi(C - \theta U - f(V)))$. By arguments similar to those used in lemma 2.2 and by T5, we know $H(\tilde{\theta}_n, \hat{f}_{\tilde{\theta}_n,\lambda_n}; X_i)$ belongs to some P-Donsker class. Combining the above conclusion and the inequality $\alpha \log x \leq \log(1 + \alpha\{x - 1\})$ for some $\alpha \in (0, 1)$ and any $x > 0$, we can show that

$$P \log\left[1 + \alpha\left(\frac{H(\tilde{\theta}_n, \hat{f}_{\tilde{\theta}_n,\lambda_n}; X_i)}{H(\theta_0, \hat{f}_{\theta_0,\lambda_n}; X_i)} - 1\right)\right] \geq o_P(1). \tag{60}$$

The remainder of the proof follows the proof of lemma 6 in [3].□

Proof of lemma 4. The maps (11) are uniformly bounded since $F(\cdot)$, $\dot{F}(\cdot)$ and $\ddot{F}(\cdot)$ are all uniformly bounded in $(-\infty,+\infty)$. This completes the verifications of $S1$. Note that (W,Z) are in $[0,1]^2$ and $h_0(\cdot)$ is intrinsically bounded over $[0,1]$. Hence we can show that the Fréchet derivatives of $\eta\mapsto\ddot{\ell}(\theta_0,\theta_0,\eta)$ and $\eta\mapsto\ell_{t,\theta}(\theta_0,\theta_0,\eta)$ for any $\eta\in\mathcal{H}_k$ are bounded operators, from which we can deduce that $|\ddot{\ell}(\theta_0,\theta_0,\eta)(X)-\ddot{\ell}_0(X)|$ is bounded by the product of some integrable function and $|\eta-\eta_0|(Z)$. This ensures (12) and (13). For (14), $P\dot{\ell}(\theta_0,\theta_0,\eta)$ can be written as $P(F(\theta_0 w+\eta_0)-F(\theta_0 w+\eta(z)))(w-h_0(z))$ since $P\dot{\ell}_0=0$. Note that $P(w-h_0(z))\dot{F}(\theta_0 w+\eta_0(z))(\eta-\eta_0)(z)=0$. This implies that $P\dot{\ell}(\theta_0,\theta_0,\eta) = P(F(\theta_0 w+\eta_0)-F(\theta_0 w+\eta(z))+\dot{F}(\theta_0 w+\eta_0(z))(\eta-\eta_0)(z))(w-h_0(z))$. However, by the common Taylor expansion, we have $|F(\theta_0 w+\eta)-F(\theta_0 w+\eta_0)-\dot{F}(\theta_0 w+\eta_0)(\eta-\eta_0)|\leq\|\ddot{F}\|_\infty|\eta-\eta_0|^2$. This proves (14).

We next verify assumption E1. For the asymptotic equicontinuity condition (15), we first apply analysis similar to that used in the proof of lemma 2 to obtain

$$P\left(\frac{\dot{\ell}(\theta_0,\theta_0,\hat{\eta}_{\tilde{\theta}_n,\lambda_n})-\dot{\ell}_0}{n^{\frac{1}{4k+2}}(\lambda_n+\|\tilde{\theta}_n-\theta_0\|)}\right)^2\lesssim O_P\left(n^{-\frac{1}{2k+1}}\right).$$

By lemma 7.1 in [14], we know that $J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n})=O_P(1+\|\tilde{\theta}_n-\theta_0\|/\lambda_n)$ and $\|\hat{\eta}_{\tilde{\theta}_n,\lambda_n}\|_\infty$ is bounded in probability by a multiple of $J(\hat{\eta}_{\tilde{\theta}_n,\lambda_n})+1$. Now we construct the set $\tilde{\mathcal{Q}}_n$ as follows:

$$\left\{\frac{\dot{\ell}(\theta_0,\theta_0,\eta)-\dot{\ell}_0}{n^{\frac{1}{4k+2}}(\lambda_n+\|\theta-\theta_0\|)}:J(\eta)\leq C_n(1+\frac{\|\theta-\theta_0\|}{\lambda_n}),\|\eta\|_\infty\leq C_n(1+J(\eta)),\right.$$
$$\left.\|\theta-\theta_0\|<\delta\right\}\cap\left\{g\in L_2(P):Pg^2\leq C_n n^{-\frac{1}{2k+1}}\right\}.$$

Clearly, the probability that the function $n^{-1/(4k+2)}(\dot{\ell}(\theta_0,\theta_0,\hat{\eta}_{\tilde{\theta}_n,\lambda_n})-\dot{\ell}_0)/(\lambda_n+\|\tilde{\theta}_n-\theta_0\|))\in\tilde{\mathcal{Q}}_n$ approaches 1 as $C_n\to\infty$. We next show that $\lim_{n\to\infty}E^*\|\mathbb{G}_n\|_{\tilde{\mathcal{Q}}_n}<$

∞ by T2. Note that $\dot{\ell}(\theta_0, \theta_0, \eta)$ depends on η in a Lipschitz manner. Consequently, we can bound $H_B(\epsilon, \tilde{\mathcal{Q}}_n, L_2(P))$ by the product of some constant and $H(\epsilon, \mathcal{R}_n, L_2(P))$ in view of T3, where $\mathcal{R}_n$ is as defined in the proof of lemma 2. By similar calculations as those performed in lemma 2, we can obtain $K(\delta_n, \tilde{\mathcal{Q}}_n, L_2(P)) \lesssim \lambda_n^{-1/2k} n^{-1/(4k+2)}$. Thus $\lim_{n\to\infty} E^* \|\mathbb{G}_n\|_{\tilde{\mathcal{Q}}_n} < \infty$, and (15) follows.

Next we define $\bar{\mathcal{V}}_n \equiv \{\ell^{(3)}(\bar{\theta}_n, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n})(x)\}$. Similar arguments as those used in the proof of lemma 2 can be directly applied to the verification of (18) in this second model. By the form of $\ell^{(3)}(t, \theta, \eta)$, the entropy number for $\bar{\mathcal{V}}_n$ is bounded above by that of $\bar{\mathcal{F}}_n \equiv \{\ddot{F}(tw + \eta(z) + (\theta - t)h_0(z)) : (t, \theta) \in V_{\theta_0}, \lambda_n J(\eta) \leq C_n, \|\eta\|_\infty \leq C_n(1 + J(\eta))\}$. Similarly, we know $\sup_Q H(\epsilon, \bar{\mathcal{V}}_n, L_1(Q)) \leq \sup_Q H(\epsilon, \bar{\mathcal{F}}_n, L_1(Q)) \lesssim ((1 + \lambda_n^{-1})/\epsilon)^{1/k} = o_P(n)$. Moreover, the $\bar{\mathcal{F}}_n$ are uniformly bounded. This completes the proof for (18).

The proof of (16) and (17) follows arguments quite similar to those used in the proof of lemma 2. In other words, we can show that $\mathbb{G}_n(\ddot{\ell}(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n}) - \ddot{\ell}_0) = o_P(1+n^{1/3}\|\tilde{\theta}_n - \theta_0\|) = o_P(1)$ and $\mathbb{G}_n(\ell_{t,\theta}(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n}) - \ell_{t,\theta}(\theta_0, \theta_0, \eta_0)) = o_P(1 + n^{1/3}\|\tilde{\theta}_n - \theta_0\|)$. This concludes the proof.□

Proof of lemma 5: The proof of lemma 5 is analogous to that of lemma 3.□

LEMMA 6. *Assuming the assumptions in theorem 4, we have*

$$
\begin{aligned}
(61)\quad logpl_{\lambda_n}(\tilde{\theta}_n) \;&=\; \log pl_{\lambda_n}(\theta_0) + n(\tilde{\theta}_n - \theta_0)^T \mathbb{P}_n \tilde{\ell}_0 \\
&-\; \frac{n}{2}(\tilde{\theta}_n - \theta_0)^T \tilde{I}_0 (\tilde{\theta}_n - \theta_0) + O_P(g_{\lambda_n}(\|\tilde{\theta}_n - \hat{\theta}_{\lambda_n}\|)),
\end{aligned}
$$

for any $\tilde{\theta}_n = \theta_0 + o_P(1)$.

Proof. $n^{-1}(\log pl_{\lambda_n}(\tilde{\theta}_n) - \log pl_{\lambda_n}(\theta_0))$ is bounded above and below by

$$\mathbb{P}_n(\ell(\tilde{\theta}_n, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n}) - \ell(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n})) - \frac{1}{n}\lambda_n^2(J^2(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) - J^2(\eta_{\theta_0}(\tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n})))$$

and

$$\mathbb{P}_n(\ell(\tilde{\theta}_n, \theta_0, \hat{\eta}_{\theta_0,\lambda_n}) - \ell(\theta_0, \theta_0, \hat{\eta}_{\theta_0,\lambda_n})) - \frac{1}{n}\lambda_n^2(J^2(\eta_{\tilde{\theta}_n}(\theta_0, \hat{\eta}_{\theta_0,\lambda_n})) - J^2(\hat{\eta}_{\theta_0,\lambda_n})),$$

respectively. By the third order Taylor expansion of $\tilde{\theta}_n \mapsto \mathbb{P}_n\ell(\tilde{\theta}_n, \theta, \eta)$ around θ_0, for $\theta = \tilde{\theta}_n$ and $\eta = \hat{\eta}_{\tilde{\theta}_n,\lambda_n}$, and the above empirical no-bias conditions (19) and (20), we can find that the order of the difference between $\mathbb{P}_n(\ell(\tilde{\theta}_n, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n}) - \ell(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n}))$ and $(\tilde{\theta}_n - \theta_0)^T\mathbb{P}_n\tilde{\ell}_0 - (\tilde{\theta}_n - \theta_0)^T(\tilde{I}_0/2)(\tilde{\theta}_n - \theta_0)$ is $O_P(n^{-1}g_{\lambda_n}(\|\tilde{\theta}_n - \hat{\theta}_{\lambda_n}\|))$. By the inequality $J^2(\eta_t(\theta,\eta)) \leq 2J^2(\eta) + 2(\theta - t)^2J^2(h_0)$, we know that $\lambda_n^2(J^2(\hat{\eta}_{\tilde{\theta}_n,\lambda_n}) - J^2(\eta_{\theta_0}(\tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n,\lambda_n}))) = O_P(\|\tilde{\theta}_n - \hat{\theta}_{\lambda_n}\| + \lambda_n)^2$ provided assumptions (3) and (10) hold. Similar analysis also applies to the lower bound. This proves (61).□

Acknowledgments. The authors thank Dr. Joseph Kadane for several insightful discussions.

REFERENCES

[1] Birman, M.S. and Solomjak, M.J. (1967). Piece-wise polynomial approximations of functions of the classes W_p^α. *Mat. Sbornik* **3** 295–317.

[2] Cheng, G. and Kosorok, M.R. (2006). Higher order semiparametric frequentist inference with the profile sampler. *Annals of Statistics*, Tentatively Accepted.

[3] Cheng, G. and Kosorok, M.R. (2006). General Frequentist Properties of the Posterior Profile Distribution. Submitted.

[4] Good, I.J. and Gaskins, R.A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

[5] Huang, J. (1999). Efficient estimation of the partly linear Cox model. *Annals of Statistics* **27** 1536–1563.

[6] Kosorok, M. R. (To appear). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

[7] Kuo, H. H. (1975). *Gaussian Measure on Banach Spaces. Lecture Notes in Mathematics* **463** Berlin: Springer.

[8] Lee, B. L., Kosorok, M. R. and Fine, J. P. (2005). The profile sampler. *Journal of the American Statistical Association* **100** 960–969.

[9] Ma, S. and Kosorok, M.R. (2005). Penalized Log-likelihood Estimation for Partly Linear Transformation Models with Current Status Data. *Annals of Statistics* **33** 2256-2290.

[10] Ma, S. and Kosorok, M.R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis* **96** 190-217.

[11] Ma, S. and Kosorok, M.R. (2006). Adaptive penalized M-estimation with current status data. *Annals of the Institute of Statistical Mathematics* **58** 511-526.

[12] Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Annals of Statistics* **25** 1014-1035.

[13] Murphy, S. A. (1995). Asymptotic Theory for the Frailty Model. *Annals of Statistics* **23** 182–198.

[14] Murphy, S. A. and Van der Vaart, A. W. (1999). Observed information in semi-parametric models. *Bernoulli* **5** 381–412.

[15] Murphy, S. A. and Van der Vaart, A. W. (2001). Semiparametric mixtures in case-control studies. *Journal of Multivariate Analysis* **79** 1–32.

[16] Shen, X. (2002). Asymptotic normality in semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association* **97** 222–235.

[17] Silverman, B.W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics* **10** 795–810.

[18] Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society. Series B* **47** 1–52.

[19] van de Geer, S. (2000). *Empirical Processes in M-estimation* Cambridge University Press, Cambridge.

[20] van der Vaart, A. W. (1994). *Maximum Likelihood Estimation with Partially Censored Observations.*. *Annals of Statistics* **22** 1896–1916.

[21] van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York.

[22] Wahba, G. (1998). *Spline Models for Observational Data.* SIAM, Philadelphia.

Guang Cheng
Institute of Statistics and
Decision Sciences
Duke University
214 Old Chemistry Building
Durham, NC 27708
USA
Email: chengg@stat.duke.edu

Michael R. Kosorok
Department of Biostatistics
School of Public Health
University of North Carolina
at Chapel Hill
3101 McGavran-Greenberg Hall
Chapel Hill, NC 27599
USA
Email: kosorok@unc.edu