The Annals of Statistics
1997, Vol. 25, No. 6, 2555–2591

ON METHODS OF SIEVES AND PENALIZATION[1]

BY XIAOTONG SHEN

Ohio State University

We develop a general theory which provides a unified treatment for the asymptotic normality and efficiency of the maximum likelihood estimates (MLE's) in parametric, semiparametric and nonparametric models. We find that the asymptotic behavior of substitution estimates for estimating smooth functionals are essentially governed by two indices: the degree of smoothness of the functional and the local size of the underlying parameter space. We show that when the local size of the parameter space is not very large, the substitution standard (nonsieve), substitution sieve and substitution penalized MLE's are asymptotically efficient in the Fisher sense, under certain stochastic equicontinuity conditions of the log-likelihood. Moreover, when the convergence rate of the estimate is slow, the degree of smoothness of the functional needs to compensate for the slowness of the rate in order to achieve efficiency. When the size of the parameter space is very large, the standard and penalized maximum likelihood procedures may be inefficient, whereas the method of sieves may be able to overcome this difficulty. This phenomenon is particularly manifested when the functional of interest is very smooth, especially in the semiparametric case.

1. Introduction. Let $Y_1, \ldots, Y_n$ be independently and identically distributed according to density $p_0(y) = p(\theta_0, y)$, where θ_0 is the true parameter value in Θ, the space of all possible parameters θ. We estimate a real functional of θ, denoted as $f(\theta)$. Such functionals characterize many interesting problems. In semiparametric models in which the interest is in estimation of the parametric component, $f(\theta)$ takes the form β, which is the parametric component, where the parameter $\theta = (\beta, \eta)$ and η is a nonparametric nuisance parameter. In a nonparametric model in which the interest is in estimation of the Shannon information entropy, $f(\theta)$ can take the form of $-2\int \theta^2 \log \theta$, where θ is the square root of the density.

In this paper, we are interested in efficient estimation of functionals $f(\theta)$. The asymptotic normality and the efficiency of the variants of maximum likelihood estimates (MLE's) is established in general parameter spaces that permit treatment of nonparametric and semiparametric situations.

To estimate $f(\theta)$, first consider standard maximum likelihood type of estimation. Let the empirical criterion $L_n(\theta)$ be $n^{-1}\sum_{i=1}^n l(\theta, Y_i)$, where $l(\theta, y)$ is the criterion based on a single observation. Here $l(\theta, y)$ may be chosen as a log-likelihood (ML estimation) or some criterion other than the log-likelihood,

Received June 1993; revised March 1997.

[1]Supported by a seed grant of the research foundation at Ohio State University.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62A10.

Key words and phrases. Asymptotic normality, efficiency, maximum likelihood estimation, methods of sieves and penalization, constraints, substitution, nonparametric and semiparametric models.

such as $-(y-\theta)^2$ in the least-squares regression. A maximizer of $L_n(\theta)$ over $\theta \in \Theta$, denoted by $\hat{\theta}_n$, is called a MLE type of estimate. With $\hat{\theta}_n$ as defined, $f(\theta)$ is estimated by a substitution estimate $f(\hat{\theta}_n)$.

In parametric models, it is well known [e.g., Bahadur (1967)] that under general conditions $n^{1/2}(f(\hat{\theta}_n) - f(\theta_0))$ converges to a normal distribution with mean zero and variance $(f'(\theta_0))^T I(\theta_0)^{-1} f'(\theta_0)$ and the MLE $f(\hat{\theta}_n)$ is asymptotically efficient in the Fisher sense, where $f'(\theta_0)$ is the usual derivative at θ_0 and $I(\theta_0)$ is the Fisher information matrix. Cramér (1946) and Pollard (1984) gave fairly general results for asymptotic normality based, respectively, on score equations and log-likelihood ratios. In the infinite-dimensional case, establishing such a theory, however, is difficult [see Wong and Severini (1991)]. Little is known about the properties of the MLE and related estimates even though various generalizations of the information lower bounds evaluating the performance of estimates are available. The issues that govern the finite-dimensional case do not readily extend to the infinite-dimensional case. The difficulties are that, unlike the parametric case, (1) the corresponding score equation evaluated at the maximizer $\hat{\theta}_n$ may not even be close to zero, especially in sieve estimation problems, (2) $\hat{\theta}_n$ is often on the boundary of the parameter space and (3) θ_0 is constrained, such as in density estimation problems. Furthermore, the remainders in local expansions depend on the convergence rate of the estimate, which may be much slower than $n^{-1/2}$ when the parameter space is large.

Estimating functionals in general parameter spaces clearly is important. In the semiparametric setting, the efficient score method leads to an efficient estimate for the parametric component [see Bickel (1982), Ritov and Bickel (1990) and the monograph by Bickel, Klassen, Ritov and Wellner (1994) for a comprehensive survey]. This method requires constructing a $n^{-1/2}$-consistent estimate and estimating the efficient score function. In the nonparametric setting, von Mises (1947) considered the problem of estimating a functional based on a class of distribution functions. For smooth functionals, Pfanzagl (1982) and Ibragimov and Has'minskii (1981, 1991) constructed $n^{-1/2}$-optimal estimates in the minimax sense for specific models. Unfortunately, a general theory is lacking. There is also an enormous amount of literature on the related topics of convergence rates for general functions [see, e.g., Bickel and Ritov (1988) for more details]. Wong and Severini (1991) showed that the standard MLE is asymptotically efficient in a compact space where the size of the parameter space is not large. Severini and Wong (1992) studied the efficiency of the profile MLE in semiparametric models. In many situations, especially when the parameter space Θ is large, variants of the MLE type of estimates such as sieve and penalized estimates are often used to overcome the difficulty of optimization and certain undesirable properties of the standard MLE; see Sections 2 and 3 for detailed discussions. Unfortunately, a general theory of asymptotic normality for the sieve and penalized estimates has not been available.

The fundamental questions are, of course, in general parameter spaces (1) whether there exists any estimate which can achieve the Fisher information lower bound typically used to evaluate the asymptotic performance of

estimates and (2) whether the variants of MLE's such as the standard, sieve and penalized MLE's can achieve the information lower bound in general, especially when the parameter space is very large and if so, to what extent.

In this paper, we introduce a general theory for establishing the asymptotic normality and efficiency for $f(\hat{\theta}_n)$. We (1) explore the relationship between the size of the parameter space and the performance of the substitution estimates, (2) try to understand the extent to which the variants of MLE's are efficient in the Fisher sense, (3) give a unified framework to the asymptotic normality for the finite- and the infinite-dimensional problems and (4) provide some insight into the structure of estimation problems.

We will show for the sieve, the standard (nonsieve) and the penalized estimates that $n^{1/2}(f(\hat{\theta}_n) - f(\theta_0))$ has an asymptotic normal distribution under general conditions: (1) the degree of smoothness of f can compensate for the slowness of the convergence rate of the estimate; (2) the empirical criterion satisfies certain stochastic equicontinuity conditions. See Section 4 for formal definitions. On this basis, we show that the standard (nonsieve), sieve and penalized MLE's are asymptotically efficient in the Fisher sense. Furthermore, there indeed is a cutoff point at a certain stage corresponding to the local metric entropy index (used to measure the size of the parameter space, as defined in Section 4) being equal to 2. When the local metric entropy index of the parameter space is less than 2, the above results are expected to hold for the standard, sieve and penalized MLE's. However, as shown in Example 3, when the local metric entropy index of the parameter space is at least 2, the above results may not hold for the standard and penalized MLE's. In contrast, the sieve with an orthogonality property to be specified in Section 4 may not have this difficulty when the functional of interest is very smooth [ω in (4.1) or (4.4) is large]. This phenomenon is illustrated in Examples 2 and 3. Moreover, as illustrated in Example 3, if f is not smooth enough, then the above result does not hold. This is because the behavior of $f(\hat{\theta}_n) - f(\theta_0)$ is determined by that of $\|\hat{\theta}_n - \theta_0\|$. This aspect of the theory offers additional insights into the structure of the problem.

The theory developed here is general, allowing for a general criterion function (with a penalty) with constrained optimization over a general sieve. The present theory encompasses, for instance, the existing results for the standard MLE in a compact space when the size of parameter space is not large, and the classical results on the asymptotic normality and efficiency of the standard MLE in the finite-dimensional case. Thus, it provides a unified treatment for most problems of estimation of a smooth functional using the ML method with independent observations. Moreover, the theory can also apply to the situation in which the functional of interest is multivariate; see Example 4 for an illustration.

The present theory is formulated based on stochastic equicontinuity related to log-likelihood ratios and convergence rates. Here the convergence rate of the estimate plays an important role. In conjunction with the convergence rate results on the sieve estimates [e.g., Shen and Wong (1994), Wong and Shen (1995) and Birgé and Massart (1994)] and on the penalized estimates [e.g.,

Shen (1997)], we make the verification of regularity conditions easier. In the semiparametric case, the present theory provides an alternative approach for constructing efficient estimates using the variants of MLE's. In this way, one bypasses the requirements of the efficient score method, that is, constructing an $n^{-1/2}$-consistent estimate and obtaining a suitable estimate for the score function are not necessary. The theory also provides some insight into the phenomenon that, in semiparametric models, the parametric component can be estimated efficiently in the Fisher sense at exactly the rate of $n^{-1/2}$, even if the estimated nonparametric nuisance component converges to the true parameter at a rate much slower than $n^{-1/2}$.

The organization of the paper is as follows. In Sections 2.1 and 2.2, we discuss the methods of sieves and penalization. In Section 3, we provide some examples and illustrative conclusions from the general theory. In Section 4, we develop the general theory on the asymptotic normality for the corresponding substitution sieve and penalized estimates. In Section 5, we address the issue of asymptotic efficiency in the Fisher sense. In Section 6, we obtain some results on constrained estimation. In Section 7, we illustrate the main results by several examples, including nonparametric regression, semiparametric regression, the proportional odds model and the density estimation problem. As particular applications of the general theory, we obtain the asymptotic normality and efficiency of the variants of MLE's in estimating moments of the regression function, the parametric component in a semiparametric model and the Shannon information entropy. In Section 8, we compare various estimation procedures and discuss implications of the present theory. In Section 9, we provide the technical proofs.

2. Estimation methods.

2.1. *The method of sieves.* Often, optimization over a large parameter space leads to undesirable properties of the estimates, such as inconsistency and roughness. Moreover, such an optimization procedure is difficult to implement and certain approximations need to be made in order to carry out computation in practice. For the above reasons, the optimization is usually carried out within a subset which is dense in the original parameter space, and the size of the subset grows as the sample size increases. More specifically, let Θ_n be a sequence of approximating spaces to the parameter space Θ (not necessarily a subset of Θ), denoted as a sieve, in the sense that for any $\theta \in \Theta$ there exists $\pi_n\theta \in \Theta_n$ such that $\|\pi_n\theta - \theta\| \to 0$ as $n \to \infty$. An approximate sieve estimate, denoted by $\hat{\theta}_n$, is defined as an approximate maximizer of $L_n(\theta)$ over Θ_n, that is,

$$L_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta_n} L_n(\theta) - O(\varepsilon_n^2), \tag{2.1}$$

where $\varepsilon_n \to 0$ as $n \to \infty$. For the exact estimate, $\varepsilon_n = 0$. The substitution sieve estimate for $f(\theta)$, by definition, is $f(\hat{\theta}_n)$. The above procedure is called the method of sieves [Grenander (1981)], which may be regarded as a gener-

alization of the standard ML estimation based on optimization over the whole parameter space Θ since Θ_n can be taken to be Θ for all n.

2.2. *The method of penalization.* In some cases, to overcome the difficulties in optimization and the undesirable properties associated with estimates based on a large parameter space, a penalty assessing the physical plausibility of each parameter value is attached to the empirical criterion to be optimized. To be more specific, let $\tilde{l}(\theta, y)$ be $l(\theta, y) - \lambda_n J(\theta)$ and $\tilde{L}_n(\theta) = L_n(\theta) - \lambda_n J(\theta)$, where $J(\theta)$ is a nonnegative penalty function and λ_n is the penalization coefficient. An approximate penalized estimate is defined as an approximate maximizer $\hat{\theta}_n$ of $\tilde{L}_n(\theta)$ over Θ, that is,

$$\tilde{L}_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} \tilde{L}_n(\theta) - O(\varepsilon_n^2), \tag{2.2}$$

where $\varepsilon_n \to 0$ as $n \to \infty$. The approximate substitution penalized estimate then is $f(\hat{\theta}_n)$. This procedure is called the method of penalization.

The use of penalty has a long history, which may trace back to Whittaker (1923) and Tikhonov (1963); see Wahba (1990) for a review. In a regression context, the penalty $J(\cdot)$ is often chosen to penalize the undesired properties such as "roughness." In this situation, the method of penalization leads to splines. Actually, the penalty plays a role of forcing the optimization in (2.2) to be carried out within compact sets depending on the sample size. Indeed, the optimization is done within a finite-dimensional space. Essentially, the penalty $J(\theta)$ which controls the global properties of the estimates plays no role in the local approximation of the criterion difference within a neighborhood of θ_0. However, to control the local behavior of the linear approximation of the criterion function with penalty, certain assumptions on $J(\theta)$ and λ_n are required.

3. Some examples and illustrative conclusions. In this section, we provide some examples and illustrative conclusions. See Section 5 for a formal discussion of asymptotic efficiency.

EXAMPLE 1 (Nonparametric regression). Suppose

$$Y_i = \theta(X_i) + e_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

where X_i and e_i are independent, and $\{e_i\}_{i=1}^n$ are independently and identically distributed with $Ee_i^2 = \sigma^2$. Assume that $\{X_i\}_{i=1}^n$ are random and the distribution of X_i does not depend on θ. The functional of interest is $f(\theta) = E\theta^k = \int \theta^k \, dP_0$ for some integer $k \geq 1$, where θ is the regression function and P_0 is the distribution of X_i.

In this example, we examine three estimation procedures in terms of the efficiency in the Fisher sense and illustrate the phenomena mentioned in the Introduction. Here the criterion is $l(\theta, y, x) = -(1/(2\sigma^2))(y - \theta(x))^2$. The empirical criterion to be maximized then is $-(1/(2\sigma^2 n)) \sum_{i=1}^n (Y_i - \theta(X_i))^2$.

In the following certain moment conditions on e_i are imposed depending on the case. For simplicity, assume that σ^2 is fixed.

(a) *Estimation without sieve.* Let

$$\Theta = \left\{\theta \in C^m[a,b]\colon \|\theta^{(j)}\|_{\sup} \le L_j,\ \frac{|\theta^{(m)}(x_1) - \theta^{(m)}(x_2)|}{|x_1 - x_2|^\gamma} < L_{m+1},\ j = 0, \ldots, m+1\right\},$$

where $p = m + \gamma > 1/2$, $\{L_j\}_{j=0}^{m+1}$, and a and b are known constants. Assume that $\mathrm{E}|e_1^l| < \infty$ for any real $l > 2$. In the present case, the maximization is done over the whole parameter space Θ, which is compact.

(b) *Sieve estimation.* Let $\Theta = \{\theta \in C^m[a,b]\colon |\theta^{(m)}(x_1) - \theta^{(m)}(x_2)|/|x_1 - x_2|^\gamma < L_{m+1}(\theta)\}$, where $p = m + \gamma > 1/2$. Here the parameter space Θ is not compact since $L_{m+1}(\theta)$ is unknown. For simplicity, assume that $\mathrm{E}\exp(t_0|e_1|) < \infty$, for some $t_0 > 0$. Also see Shen and Wong (1994) for the required condition based on moments.

(1) *Finite-dimensional sieve.* Consider the following truncated series expansion. Let

$$\Theta_n = \left\{\theta \in \Theta\colon \theta(x) = \alpha_0 + \sum_{j=1}^{r_n}(\alpha_j \cos(2\pi jx) + \beta_j \sin(2\pi jx)),\ \alpha_0^2 + \sum_{j=1}^{r_n} j^{2p'}(\alpha_j^2 + \beta_j^2) \le l_n^2\right\},$$

where p' is a constant arbitrarily close to p, $l_n \le n^{(2p-1)/(2p'(2p+1))}$ and $r_n = n^\tau$ with $1/(4p) < \tau < 1/2$. A natural choice of τ is $1/(2p+1)$. The resulting rate is $O_p(n^{-p/(2p+1)})$, which is the optimal convergence rate of the estimate under $\|\cdot\|$.

The above sieve is based on the trigonometric basis functions. Other bases could be used in the same context. Now consider a sieve based on a local basis. Let

$$\Theta_n = \left\{\theta = \sum_{i=1}^{r_n+[p]+1} a_i\phi_i \in \Theta,\ \max_{i=1,\ldots,r_n+[p]+1} |a_i| \le l_n\right\},$$

where $(\phi_1, \ldots, \phi_{r_n+([p]+1)})$ are B-splines of order $[p]+1$ on $[a,b]$ with ϕ_i supported on $[x_i, x_{i+[p]+1}]$, and $(a = x_1, \ldots, x_{r_n+[p]+1} = b)$ is the uniform partition of $[a,b]$ that supports the basis.

(2) *Infinite-dimensional sieve.* Let Θ be the same as in (1) above. Let $J(\theta) = (\int_a^b |\theta^{(p)}(x)|^q\,dx)^{1/q}$ for some real number $p > 1/2$ and $q \ge 1$. The derivative with a fractional power is defined in terms of Fourier series, that is,

$$\theta^\alpha(x) = \sum_{k=1}^{\infty} k^\alpha\left[\left(a_k \cos\frac{\pi}{2}\alpha + b_k \sin\frac{\pi}{2}\alpha\right)\cos kx + \left(b_k \cos\frac{\pi}{2}\alpha - a_k \sin\frac{\pi}{2}\alpha\right)\sin kx\right],$$

for $\theta(x) = \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx)$ and any $0 < \alpha \leq p$. In the penalization estimation, the empirical criterion with penalty often leads to various splines for different integer values of q. The popular choices of q are 2 in the non-parametric regression, and 1 or ∞ in the conditional quantile regression [Koenker and Bassett (1978)].

Let the sieve Θ_n be $\{\theta \in \Theta\colon J(\theta) \leq b_n\}$ with $b_n \to \infty$ as $n \to \infty$ arbitrarily slowly. The sieve estimate is obtained by maximizing the log-likelihood over Θ_n. The relationship between the sieve defined here and the penalized estimation in part (c) can be found in Schoenberg (1964).

(c) *Penalization.* Let

$$\Theta = \left\{\theta \in C^m[a,b]\colon \theta(a) = \theta(b) = 0,\ \frac{|\theta^{(m)}(x_1) - \theta^{(m)}(x_2)|}{|x_1 - x_2|^{\gamma}} < L_{m+1}(\theta)\right\},$$

where $p = m + \gamma > 1/2$, $L_{m+1}(\theta)$ is unknown and the design points $\{X_i\}_{i=1}^n$ are deterministic and equally spaced. In the case of $p \geq 1$, $J(\theta) = \|\theta^{(m)}\|_q + [\int\int(|\theta^{(m)}(x) - \theta^{(m)}(y)|/|x-y|^{\gamma})^q\,dx\,dy]^{1/q}$, where $\|\cdot\|_q$ is the usual L_q-norm and $q \geq 2$. In the case of $p < 1$, $J(\theta) = \sup_{x,y}|\theta(x) - \theta(y)|/|x-y|^{\gamma}$. Assume that $\mathrm{E}\exp(t_0|e_1|) < \infty$, for some $t_0 > 0$.

PROPOSITION 1. *Under the assumptions of Example* 1, *the approximate substitution sieve and penalized estimates are asymptotically normal with variance* $k^2\sigma^2\mathrm{E}_0(\theta_0^{k-1})^2$,

$$n^{1/2}(f(\hat{\theta}_n) - f(\theta_0)) \to_{\mathrm{p}_{\theta_0}} N(0, k^2\sigma^2\mathrm{E}_0(\theta_0^{k-1})^2),$$

where $\hat{\theta}_n$ *is either the sieve or penalized estimates in Example* 1(a)–(c). *In addition, if the error* e_i *is distributed as* $N(0,\sigma^2)$, *then the above estimates are asymptotically efficient in the Fisher sense.*

As already seen, the standard ML method, and the methods of sieves and penalization lead to efficient estimates for the case of $p > 1/2$. Note that different bases may yield different efficient estimates. Here the results for the case of $p \leq 1/2$ are not expected since f is not smooth enough (the degree of smoothness ω is 2 in this case). When f is smooth enough, as illustrated in Examples 2 and 3, the method of sieves leads to efficient estimates, whereas in the same setting the standard and penalized MLE's are inefficient. It is also interesting to note that the requirement on the size of the approximating spaces $(1/(4p) < \log r_n/\log n < \frac{1}{2})$ is not stringent. Therefore, it is possible to choose $r_n = n^{1/(2p+1)}$ so that the sieve estimate $\hat{\theta}_n$ achieves the efficiency for $f(\theta)$ and the optimal convergence rate $(\|\hat{\theta}_n - \theta_0\| = O_p(n^{-p/(2p+1)}))$ [Stone (1982)], which corresponds to the best trade-off phenomenon between the approximation error and the estimation error, as discussed in Shen and Wong (1994). As to be seen from Examples 2 and 3, this cannot be done when $p < 1/2$, that is, the optimal choice of the size r_n which leads to the best convergence rate for the sieve MLE under $\|\cdot\|$ may not yield an efficient estimate for $f(\theta)$. A similar phenomenon also occurs in penalized estimation; that

is, if the penalization coefficient λ_n is chosen to be of order of $n^{-2p/(2p+1)}$ for $p > 1/2$, the penalized estimate $\hat{\theta}_n$ achieves the same optimal rate as above [see Shen (1997a)].

EXAMPLE 2 (Semiparametric model). Instead of modeling the regression function completely nonparametrically in (3.1), we specify θ as $\beta Z + \eta(X)$, where β is the parameter of interest, η is the infinite dimensional nuisance parameter, and Z is the covariate of interest. Let $\Theta = A \times B$, where $A \subset \mathscr{R}^1$ is a bounded open set, and let B be the same as the parameter space in Example 1(b). Assume that $\mathrm{E}\, Z|X$ is smooth enough in X, for example, (X, Z) are normally distributed (not necessarily independent). Next we consider the methods of sieves and penalization.

(a) *Sieve estimation.* First consider the case of $p = m + \gamma > 1/2$. Let $\Theta_n = A \times B_n$, where

$$B_n = \Bigg\{\theta \in B\colon \theta(x) = \alpha_0 + \sum_{j=1}^{r_n}(\alpha_j \cos(2\pi jx) + \beta_j \sin(2\pi jx)), \\ \alpha_0^2 + \sum_{j=1}^{r_n} j^{2p'}(\alpha_j^2 + \beta_j^2) \le l_n^2\Bigg\}.$$

Here $r_n = n^\tau$, with $1/4p < \tau < 1/2$. A natural choice of τ is $1/(2p+1)$. And, p' and l_n are the same as in Example 1. In this case, the empirical criterion to be maximized is $-(1/(2\sigma^2 n))\sum_{i=1}^n (Y_i - (\beta Z_i + \eta(X_i)))^2$.

Next consider the case of $p = m + \gamma \le 1/2$. Let $\{\phi_i\}_{i=1}^\infty$ be an orthonormal basis (Gram–Schmidt orthogonalization based on the trigonometric basis) with respect to $\langle\cdot,\cdot\rangle$. Let

$$B_n = \Bigg\{\theta \in B\colon \theta(x) = \sum_{j=1}^{r_n} \alpha_j \phi_j(x),\ \sum_{j=1}^{r_n} \alpha_j^2 \le l_n^2\Bigg\},$$

where $r_n = n^\tau$. Here $0 < \tau \le 1 - d$ for some $d > 1/2$.

(b) *Penalization.* Let $J(\cdot)$ be the same as defined in Example 1(c). Consider the penalized estimation with penalty $J(\eta)$ for $p = m+\gamma > 1/2$. The penalized empirical criterion to be maximized is

$$-\frac{1}{2\sigma^2 n}\sum_{i=1}^n (Y_i - (\beta Z_i + \eta(X_i)))^2 + \lambda_n J(\eta).$$

PROPOSITION 2. *Under the assumptions of Example* 2, *the approximate sieve and penalized estimates are asymptotically efficient with variance* $\sigma^2/\mathrm{E}(Z - \mathrm{E}\, Z|X)^2$,

$$n^{1/2}(\hat{\beta}_n - \beta_0) \to_{\mathrm{p}_{\theta_0}} N(0, \|v^*\|^2),$$

where $\|v^*\|^2 = \sigma^2/\mathrm{E}(Z - \mathrm{E}\, Z|X)^2$, *and* $\hat{\theta}_n$ *is either the sieve estimate in Example* 2(a) *or the penalized estimate in Example* 2(b).

It is interesting to note that the convergence rate of the sieve estimate in the first case of Example 2(a) is $O_p(n^{-p/(2p+1)})$ (optimal), whereas that in the second case of Example 2(a) is $O_p(n^{-\tau p})$, which is close to $n^{-p/2}$ when τ is close to 1/2. This suboptimal rate is actually the rate of the standard MLE based on the parameter space Θ in Example 1(a) for $p < 1/2$ [see Shen and Wong (1994), part (b) of Example 3, for discussions of this suboptimality].

Compared with the results in Example 2(a) and (b), we found that the sieve estimate is asymptotically efficient when $p = m+\gamma > 0$, whereas as illustrated in Example 3, the same result is only expected for the standard and penalized MLE's when $p = m + \gamma > 1/2$. Note that the case of $p = m + \gamma > 1/2$ corresponds to the local metric entropy index being less than 2.

EXAMPLE 3 (Inefficiency, the degree of smoothness and the size of the parameter space). The purpose of this example is to show that (1) the standard and penalized ML procedures may not yield efficient estimates even for a very smooth functional $f(\theta)$ when the parameter space is very large, which corresponds to the case where the local metric entropy index is larger than 2 ($m + \gamma < 1/2$ in Example 2), and (2) the substitution estimates may not be efficient when the functional $f(\theta)$ is not smooth enough.

To illustrate (1), consider a semiparametric model in which $f(\theta)$ is linear (very smooth). Let (Y_i, X_i, Z_i) and $\lambda = \beta Z + \eta(X)$ be the same as in Example 2. Let $X_i \in [0, 1/2]$ and $Z_i \in [0, 1]$ be uniformly distributed, and let e_i be as in (3.1). In addition, X_i and Z_i are independent. First consider the standard ML estimation with $\Theta = A \times B^\alpha$, where

$$B^\alpha = \left\{ \eta\colon |\eta(x_1) - \eta(x_2)| \le \frac{1}{2^\alpha} |x_1 - x_2|^\alpha \right\}$$

for $0 < \alpha < 1/2$, and $A = (-1, 1)$.

Next we provide a concrete B^α. To define the elements in B^α, consider the following basic functions. For $0 < \alpha < 1/2$ and any $0 \le h \le 1/2$, let

$$B_h(x) = \begin{cases} h^\alpha[(1/2)^\alpha - |x/h - 1/2|^\alpha], & \text{on } (0, h], \\ 0, & \text{otherwise,} \end{cases}$$

and

$$G_h(x) = \begin{cases} ([\mathrm{E}e_1 \operatorname{sgn}(e_1)]/2^{\alpha+2})^{1/2} h^{\alpha/2} x, & \text{on } (0, h/4], \\ ([\mathrm{E}e_1 \operatorname{sgn}(e_1)]/2^{\alpha+2})^{1/2} h^{\alpha/2} (h/2 - x), & \text{on } [h/4, h/2], \\ 0, & \text{otherwise.} \end{cases}$$

Let $\{c_j\}_{j=1}^m \subset [0, 1]$ satisfy the property that $\{(c_j - h, c_j + h)\}_{j=1}^m$ are disjoint and are contained in $[1/8, 3/8]$. Now $B^\alpha = \{\eta_0 + \sum_{j=1}^m s_j B_h(c_j - x) + G_h(c_j - x)\colon 0 \le h \le 1/8\}$ is a class of functions indexed by h with Hölderian exponent α, where $s_j = \pm 1$. Here $B_h(\cdot)$ satisfies (1) $\int (B_h(x) + G_h(x))^i \, dx = d_i h^{i(\alpha/2)+1}(1 + o(h^{\alpha/2}))$, $(d_i > 0)$, $i = 1, 2$, and (2) $B_h(c_j - x)$ are asymmetric and disjoint.

Now consider penalized estimation. Let B^α be the same as above except that $h \in [0, \infty)$ and the design points $\{X_i\}_{i=1}^n$ are deterministic and equally

spaced. Let $J(\eta) = [\int\int(|\eta(x) - \eta(y)|/|x - y|^{\alpha})^p\,dx\,dy]^{1/p}$. The penalty $J(\eta)$ is natural in this case since it penalizes nonsmoothness of η. The penalized likelihood $\tilde{l}(\theta, y)$ is $l(\theta, y) - \lambda_n J(\eta)$. Without loss of generality, assume that $J(\eta_0) > 0$. The conclusion for the standard MLE continues to hold for the penalized MLE in this case.

PROPOSITION 3. *For the standard and penalized MLE's* $\hat{\theta}_n = (\hat{\beta}_n, \hat{\eta}_n)$ (*based on* B^{α}) *in this example*, *we have*, *with a nonzero probability*,

$$|\mathrm{E}_0(\hat{\eta}_n - \eta_0)| \geq c_1 n^{-\alpha/2},$$

for ($c_1 > 0$). *Additionally*, *with a nonzero probability*,

$$|f(\hat{\theta}_n) - f(\theta_0)| = |\hat{\beta}_n - \beta_0| \geq cn^{-\alpha/2}.$$

See Section 9 for the proof.

The conclusion that $|\mathrm{E}_0(\hat{\eta}_n - \eta_0)| \geq c_1 n^{-\alpha/2}$ in Proposition 3 is important because the standard MLE is efficient if $\mathrm{E}_0(\hat{\eta}_n - \eta_0) = o_p(n^{-1/2})$, which means that the nonparametric and the parametric components are not closely related even if the convergence rate $(\mathrm{E}_0(\hat{\eta}_n - \eta_0)^2)^{1/2} = O_p(n^{-\alpha/2})$ is slow.

To illustrate (2), consider $f(\theta) = \mathrm{E}_0\theta^2 = \int \theta^2\,dP_0$ as in Example 1 with $\Theta = B^{\alpha}$. Note that $f(\hat{\theta}_n) - f(\theta_0) = f'_{\theta_0}[\hat{\theta}_n - \theta_0] + \|\hat{\theta}_n - \theta_0\|^2$. Hence, $f(\hat{\theta}_n) - f(\theta_0)$ does not converge at the rate $n^{-1/2}$ when $p = m + \gamma < 1/2$ since $\|\hat{\theta}_n - \theta_0\|^2 \sim O_p(n^{-\alpha})$ is slower than $n^{-1/2}$, which is dominating.

EXAMPLE 4 [The proportional odds model (right censoring)]. The proportional odds model is

$$\log(F(t|x)/(1 - F(t|x))) = -\beta^T x + \log\Gamma(t), \tag{3.2}$$

where $F(t|x) = \mathrm{P}(T \leq t|x)$ is the conditional failure time distribution given covariate values x and $\Gamma(t)$ is a baseline function. Based on the sample $\{(Y_i, \delta_i, X_i)\}_{i=1}^n$, we estimate the regression parameter $\beta = (\beta_1, \ldots, \beta_p)$ while $\Gamma(t)$ is a nuisance parameter. Here $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is an indicator and $X = (X_1, \ldots, X_p)^T$ is $p \times 1$. The following assumptions are made.

(i) Given covariates X, the censoring times $\{C_i\}$ taking values in $[0, U]$, $0 < U < \infty$, are independent of the failure times $\{T_i\}$. In addition, given covariates X, the conditional density of the censoring time $p_C(u|x)$ is differentiable in u with $\int(p'_C(u|x))^2\,du \leq M < \infty$. In order for the Fisher information to exist, we need to assume that the covariates X are not functions of Y, and the covariance matrix of X is nondegenerate. In addition, X are bounded.

(ii) Let $\theta = (\beta, \Gamma(t)) \in \Theta = E \times H$, where $E \subset \mathscr{R}^p$ is a bounded open set and $H = \{\log\Gamma'(t) \in B^s_{p,q}(C)[0, U]\colon \Gamma(0) = 0, \Gamma'(t) > d_1\}$, where $d_1 > 0$ is a

constant. Here the Besov ball is defined as

$$B^s_{p,q}(C)[0,U] = \left\{ \eta \in L_p[0,U]\colon \|\eta\|_{B^s_{p,q}} = \|\eta\|_p + \left\| (t^{-s}\omega_r(\eta,t)_p)\frac{1}{t} \right\|_q < C \right\},$$

where $C > 0$ is unknown, $2 \le p, q \le \infty$, $s > 1/2$, $\|\cdot\|_q$ is the usual L_q-norm and the modulus of smoothness of order r of η at t, $\omega_r(\eta,t)_p$, is defined as $\sup_{|h|\le t} \|\Delta^r_h(\eta,\cdot)\|_p$, with Δ^r_h being the rth-order difference with step h. See Triebel (1983) for more about Besov spaces.

We now construct a sieve based on splines. Let $I = (0 < x_1 < \cdots < x_k < U)$ be the uniformly spaced knots, and let $s(x) = \sum_{i=1}^k \sum_{j=0}^r \eta_{j_i} x^j I(x_i < x \le x_{i+1})$ be the spline with a boundary constraint $G\eta = 0$ for the coefficients $\eta = (\eta_{j_i})$ such that, on each boundary of subinterval $[x_i, x_{i+1}]$, the spline has $r-1$ derivatives. Let the sieve $\Theta_n = B \times H_n$, where $H_n = \{\tilde\Gamma \in H\colon \tilde\Gamma(t) = \int_0^t \exp(s(x))\,dx\}$ be a collection of monotone splines, where the number of knots r_n is chosen to be of order of $n^{1/(2s+1)}$. See also Parzen and Harrington (1993) and Shen (1997b) for related sieve constructions in this model.

From (3.2),

$$F(t|x) = \frac{\exp(-\beta^T x)\Gamma(t)}{1+\exp(-\beta^T x)\Gamma(t)},$$

and the hazard rate

$$\lambda(t) = -\frac{\partial \log(1-F(t|x))}{\partial t} = \frac{\exp(-\beta^T x)\Gamma'(t)}{1+\exp(-\beta^T x)\Gamma(t)}.$$

The log-likelihood $l(\theta, \delta, y, x)$ is

$$\begin{aligned}&\delta \log \lambda(\theta(y), x) - \int_0^y \lambda(\theta(u), x)\,du + \log p_X(x)\\ &\quad = \delta(\log\Gamma'(y) - \beta^T x) - (1+\delta)\log(1+\exp(-\beta^T x)\Gamma(y)) + \log p_X(x),\end{aligned}$$

where $p_X(x)$ is the density of X which is independent of θ. The log-likelihood to be maximized is

$$n^{-1}\sum_{i=1}^n [\delta_i(\log\Gamma'(Y_i) - \beta^T X_i) - (1+\delta_i)\log(1+\exp(-\beta^T X_i)\Gamma(Y_i))].$$

To define the Fisher information for β, let

$$\tilde A_j = -X^{(j)}\left[-\delta + (1+\delta)\frac{\Gamma_0(Y)\exp(-\beta^T X)}{1+\Gamma_0(Y)\exp(-\beta^T X)}\right], \qquad j = 1, \ldots, p,$$

$$\tilde B = \frac{\delta}{\Gamma_0'(Y)}, \qquad \tilde C = -(1+\delta)\frac{\exp(-\beta_0^T X)}{(1+\Gamma_0(Y)\exp(-\beta_0^T X))}.$$

Let $\{\psi_j\}_{j=1}^{\infty}$ be an orthonormal basis in $\mathscr{L}_2$, for example, the trigonometric basis. Let I_n be $I_{11}-I_{12}(s_n)I_{22}^{-1}(s_n)I_{12}^T(s_n)$, where I_{11} is the $p\times p$ matrix whose ijth element is $E_0\tilde{A}_i\tilde{A}_j$, $I_{12}(s_n)$ is the $p\times s_n$ matrix whose ijth element is $E_0\tilde{A}_i(\tilde{B}\psi'_j(Y)+\tilde{C}\psi_j(Y))$ and $I_{22}(s_n)$ is the $s_n\times s_n$ matrix whose ijth element is $E_0(\tilde{B}\psi'_i(Y)+\tilde{C}\psi_i(Y))(\tilde{B}\psi'_j(Y)+\tilde{C}\psi_j(Y))$.

PROPOSITION 4. *Under the assumptions of Example* 4, *the sieve estimate* $\hat{\beta}_n$ *defined here is asymptotically efficient for estimating* β *with covariance matrix* I^{-1}, *where* I *is the Fisher information defined as* $I=\lim_{n\to\infty} I_n$. *The existence of* $\lim_{n\to\infty} I_n$ *is shown in Section* 8.

EXAMPLE 5 (Density estimation). Let $Y_1,\dots,Y_n$ be a random sample independently and identically distributed according to a density $g(y)$ on $[a,b]$. Let $\theta=g^{1/2}$ be the parameter of interest. We estimate the information entropy of the density $f(\theta)=-2\int\theta^2\log\theta$. In this case, the log-likelihood $l(\theta,y)$ is $\log\theta^2(y)$. The standard MLE $\hat{\theta}_n$ is obtained by maximizing $n^{-1}\sum_{i=1}^n 2\log\theta(Y_i)$ over Θ. For simplicity, assume that $\theta_0\ge c_0$ for some constant $c_0>0$.

The classical theory on ML estimation cannot directly handle the constrained case in the infinite-dimensional case because $\mathrm{E}_0 l'_{\theta_0}[\theta-\theta_0, Y]\neq 0$. See Remark 4 after Corollary 1. Here the nonlinear constraints are $\int\theta^2(x)\,dx=1$ and $\theta\ge 0$. We will examine two cases related to the standard ML estimation. Note that the corresponding results for the sieve and penalized MLE's can also be obtained by arguments similar to those presented here and in Example 1.

(a) *Functions with finite amount of smoothness.* Let $\Theta=\{\theta\in C^m[a,b]\colon \theta\ge 0, \theta(a)=\theta(b)=0, \int_a^b\theta^2(x)\,dx=1, \|\theta^{(j)}(x)\|_{\sup}<L_j, \|\theta^{(m)}(x_1)-\theta^{(m)}(x_2)\|_{\sup}<L_{m+1}|x_1-x_2|^{\gamma},\ j=0,1,\dots,m\}$, where $p=m+\gamma>1/2$, and $\{L_j\}_{i=1}^{m+1}$ are known constants.

(b) *Functions with infinite amount of smoothness.* Among infinite-dimensional sets of least massiveness, the class of totally bounded analytic functions is important in the sense that the behavior of the MLE based on such a parameter space is similar to that in the finite-dimensional case.

We now introduce some notation. Let $z=(z_1,\dots,z_s)\in C^s$ (C^s is a complex domain), $k=(k_1,\dots,k_s)$, $c_k=c_{k_1\cdots k_s}$, $K=\{(k_1,\dots,k_s)\}$, $\Pi_i=\{z_i\colon |\operatorname{Im} z_i|<h\}$ $(0<h<\infty)$ and $f(z)=\sum_{k\in K}c_k\phi_k$, where ϕ_k, $k\in K$, are some basis functions such as the trigonometric basis, that is, $\phi_k=e^{i(k,z)}$ $[(k,z)=\sum_{i=1}^s k_i z_i]$. Let $A_h=\{f(z)\colon |f(z)|\le L,\ z\in\Pi_1\times\cdots\times\Pi_s\}$ for some constant $L>0$. Indeed, A_h is a class of analytic functions. Take Θ as $\{\theta(z)=f(z)\in A_h\colon z\in[0,2\pi]^s,\ \theta\ge 0,\ \int\theta^2=1,\ |\theta_0|<L\}$ for some $L>0$.

PROPOSITION 5. *The standard MLE* $f(\hat{\theta}_n)$ *in either Example* 5(a) *or* (b) *is asymptotically efficient for estimating* $f(\theta)$ *with variance* $4\operatorname{Var}_0(\log\theta_0)$, *that is*,

$$n^{1/2}(f(\hat{\theta}_n)-f(\theta_0))\to_{\mathrm{p}_{\theta_0}} N(0,4\operatorname{Var}_0(\log\theta_0)).$$

Actually, the approximate substitution standard MLE in Example 5(b) is asymptotically efficient for any smooth functional with $\omega > 1$ since the convergence rate of $\hat{\theta}_n$ under $\|\cdot\|$ is fast. This aspect is the same as that in the finite-dimensional case.

4. General theory on asymptotic normality.

4.1. *The method of sieves.* To study the asymptotic distribution of the substitution estimates, we discuss linear approximations of the criterion difference by the corresponding derivatives, and the degree of smoothness of f. In the following, all probability calculations are under $p(\theta_0, y)$.

Suppose, for all $\theta \in \Theta$ and all y, there exists $l'_{\theta_0}[\theta - \theta_0, y]$ such that the remainder in the linear approximation can be written as

$$r[\theta - \theta_0, y] = l(\theta, y) - l(\theta_0, y) - l'_{\theta_0}[\theta - \theta_0, y], \tag{4.1}$$

where $l'_{\theta_0}[\theta-\theta_0, y]$ is defined as $\lim_{t\to 0}[l(\theta(\theta_0, t), y)-l(\theta_0, y)/t]$, and $\theta(\theta_0, t) \in \Theta$ is a path in t connecting θ_0 and θ such that $\theta(\theta_0, 0) = \theta_0$ and $\theta(\theta_0, 1) = \theta$. Usually, $\theta(\theta_0, t)$ is chosen as $\theta + t[\theta - \theta_0]$, which is linear in t. In this case, $l'_{\theta_0}[\theta - \theta_0, y]$ becomes the directional derivative of l at θ_0. Sometimes, $\theta(\theta_0, t)$ is nonlinear, which is useful in the constrained problems. Assume that $l'_{\theta_0}[\theta - \theta_0, y] - E_0 l'_{\theta_0}[\theta - \theta_0, y]$ is required to be linear in $\theta - \theta_0$. Note that $l'_{\theta_0}[\theta - \theta_0, y]$ may not be linear in $\theta - \theta_0$ as in the constrained problems. See Example 5 for density estimation.

Suppose the functional f has the following smoothness property: for any $\theta \in \Theta_n$,

$$|f(\theta) - f(\theta_0) - f'_{\theta_0}[\theta - \theta_0]| \le u_n \|\theta - \theta_0\|^{\omega} \quad \text{as } \|\theta - \theta_0\| \to 0, \tag{4.2}$$

where $f'_{\theta_0}[\theta - \theta_0]$ is defined as $\lim_{t\to 0}[f(\theta(\theta_0, t), y) - f(\theta_0, y)]/t$. Here $\omega > 0$ is the degree of smoothness of f at θ_0, $f'_{\theta_0}[\theta - \theta_0]$ is linear in $(\theta - \theta_0)$ and

$$\|f'_{\theta_0}\| = \sup_{\{\theta\in\Theta:\, \|\theta-\theta_0\|>0\}} \frac{|f'_{\theta_0}[\theta - \theta_0]|}{\|\theta - \theta_0\|} < \infty.$$

Let V be the space spanned by $\Theta - \theta_0$. Assume that $\|\cdot\|$ induces an inner product $\langle\cdot, \cdot\rangle$ on the completion of V, denoted as $\bar{V}$. By the Riesz representation theorem, there exists $v^* \in \bar{V}$ such that, for any $\theta \in \Theta$, $f'_{\theta_0}[\theta-\theta_0] = \langle\theta-\theta_0, v^*\rangle$.

Let $K(\theta_0, \theta) = n^{-1}\sum_{i=1}^n \mathrm{E}_0(l(\theta_0, Y_i) - l(\theta, Y_i))$, which is the Kullback-Leibler information number based on n observations when the criterion is a log-likelihood. Let $\nu_n(g) = n^{-1/2}\sum_{i=1}^n (g(Y_i) - \mathrm{E}_0 g(Y_i))$ be the empirical process induced by g. Let the convergence rate of the sieve estimate under $\|\cdot\|$ be $o_p(\delta_n)$ and let $\varepsilon_n = o(n^{-1/2})$.

For $\theta \in \{\theta \in \Theta_n: \|\theta-\theta_0\| \le \delta_n\}$, consider a local alternative value $\theta^*(\theta, \varepsilon_n) = (1-\varepsilon_n)\theta+\varepsilon_n(u^*+\theta_0)$. Denote $P_n(\theta)$ by a projection of θ to Θ_n, where $u^* = \pm v^*$. (The projection does not have to be linear.) Then $P_n(\theta)$ can be chosen as $\pi_n(\theta)$ when $\theta \in \Theta$. Some regularity conditions will be formulated under which the asymptotic distribution of $f(\hat{\theta}_n)$ can be derived.

CONDITION A (Stochastic equicontinuity). For $r[\cdot,\cdot]$ defined in (4.1),

$$\sup_{\{\theta\in\Theta_n:\|\theta-\theta_0\|\le\delta_n\}} n^{-1/2}\nu_n(r[\theta-\theta_0,Y]-r[P_n(\theta^*(\theta,\varepsilon_n))-\theta_0,Y])=O_p(\varepsilon_n^2).$$

CONDITION B (Expectation of criterion difference). We have

$$\sup_{\{\theta\in\Theta_n:0<\|\theta-\theta_0\|\le\delta_n\}} \big[K(\theta_0,P_n\theta^*(\theta,\varepsilon_n))-K(\theta_0,\theta)\big] - \tfrac{1}{2}\big[\|\theta^*(\theta,\varepsilon_n)-\theta_0\|^2-\|\theta-\theta_0\|^2\big]=O(\varepsilon_n^2).$$

CONDITION C (Approximation error). We have

$$\sup_{\{\theta\in\Theta_n:0<\|\theta-\theta_0\|\le\delta_n\}} \|\theta^*(\theta,\varepsilon_n)-P_n(\theta^*(\theta,\varepsilon_n))\|=O(\delta_n^{-1}\varepsilon_n^2).$$

In addition,

$$\sup_{\{\theta\in\Theta_n:\|\theta-\theta_0\|\le\delta_n\}} n^{-1/2}\nu_n(l'_{\theta_0}[\theta^*(\theta,\varepsilon_n)-P_n(\theta^*(\theta,\varepsilon_n)),Y])=O_p(\varepsilon_n^2).$$

CONDITION D (Gradient). We have

$$\sup_{\{\theta\in\Theta_n:\|\theta-\theta_0\|\le\delta_n\}} n^{-1/2}\nu_n(l'_{\theta_0}[\theta-\theta_0,Y])=O_p(\varepsilon_n).$$

Conditions A, B and D can be verified by calculating the corresponding metric entropy [Shen and Wong (1994), Lemma 4], which makes verifications easier. Here the L_2-metric with bracketing $H^B(u,G)$ is defined as the logarithm of the minimal cardinality of the u-covering of the space G in the L_2-metric with bracketing [see, e.g., Pollard (1984) and Example 1 for more details].

Condition A, formulated on stochastic approximations of $n^{-1}\sum_{i=1}^n l'_{\theta_0}[\theta-\theta_0,Y_i]$ to $L_n(\theta)-L_n(\theta_0)$, is a condition related to stochastic equicontinuity. Basically, this condition specifies linear approximations of the empirical criterion by its derivative within a small neighborhood of θ_0. Condition B says that $K(\cdot,\cdot)$ is locally equivalent to $\|\cdot\|^2$, which characterizes the local quadratic behavior of the criterion difference. It is worth noticing that in the infinite-dimensional case, $\hat\theta_n$ is often on the boundary of Θ. For instance, in the density estimation without restriction on the underlying densities, the MLE may wildly oscillate depending on the data points. In the worst case, the MLE will not even be consistent. If certain smoothness such as the derivatives of the density is assumed, the local oscillation then depends on the bounds of the derivatives. See Example 1 for such a phenomenon in the context of nonparametric regression. This implies that the corresponding score function specified by the directional derivative evaluated at $\hat\theta_n$ may not even be close to zero when the parameter space is very large. In addition, interior points of Θ with respect to $\|\cdot\|$ may not exist (see Example 1). Conditions C and D, which can be viewed as a natural generalization of the usual assumption that θ_0 is an interior point of Θ in the finite-dimensional case, are used to deal with

such issues. Condition C, which is always satisfied in the standard ML estimation with $\Theta_n = \Theta$, mainly controls the approximation error of $P_n(\theta^*(\theta, \varepsilon_n))$ to $\theta^*(\theta, \varepsilon_n)$. Condition D controls the gradient of the criterion function in some sense.

THEOREM 1 (Normality). *In addition to Conditions* A–D, *let* f *satisfy* (4.2) *with* $u_n \delta_n^\omega = O(n^{-1/2})$ *and* $\mathrm{Var}_0(l'_{\theta_0}[v^*, Y]) < \infty$. *Then, for the approximate substitution sieve estimate* $f(\hat{\theta}_n)$ *defined in* (2.1),

$$n^{1/2}(f(\hat{\theta}_n) - f(\theta_0)) \to_{\mathrm{p}_{\theta_0}} N(0, \mathrm{Var}_0(l'_{\theta_0}[v^*, Y])).$$

COROLLARY 1. *If Conditions* A–D *hold, then for the approximate sieve estimate defined in* (2.1),

$$n^{1/2}\langle \hat{\theta}_n - \theta_0, s\rangle \to_{\mathrm{p}_{\theta_0}} N(0, \mathrm{Var}_0(l'_{\theta_0}[s, Y])),$$

where $s \in \Theta - \theta_0$.

Typically, $\mathrm{Var}_0(l'_{\theta_0}[s, Y])) = \|f'_{\theta_0}\|^2$.

REMARK 1. The stochastic approximation specified in Condition A follows from the often used condition that $f(\theta, y)$ is differentiable in quadratic mean. The knowledge of δ_n may not be necessary in the verification of Condition A. Typically, Condition A holds even with $\delta_n = O(1)$ in the finite-dimensional case. When the size of Θ is not large, Condition A is implied by a stronger but simpler one:

$$\sup_{\{\|\theta - \theta_0\| \le \delta_n\}} n^{-1/2} \nu_n(r[\theta - \theta_0, Y_i]) = o_p(n^{-1}).$$

Similarly, Condition B is implied by the following condition:

$$\sup_{\{\theta \in \Theta_n : 0 < \|\theta - \theta_0\| \le \delta_n\}} K(\theta_0, \theta) = \tfrac{1}{2}\|\theta_0 - \theta\|^2 + o(n^{-1}).$$

REMARK 2. In the finite-dimensional case, Conditions A and B are implied by the condition on moments of the second derivative of the criterion function since the convergence rate of the supremum of the empirical processes is often $n^{-1/2}$. Condition C is always satisfied if θ_0 is an interior point of Θ. Hence, Theorem 1 recovers the classical results in the finite-dimensional case.

REMARK 3. The symbol $\langle \cdot, \cdot \rangle$ can be any inner product as long as it satisfies the stated regularity conditions, for example, the Fisher inner and the L_2-inner product (see Examples 1, 2 and 5).

REMARK 4. The constraints on the approximating spaces as in Example 1 and the true parameter θ_0 as in Example 5 are allowed in the above formulations. Consequently, Theorem 1 applies directly to the case with constraints

such as density estimation. We emphasize that the classical theory of the ML estimation cannot handle the constraints directly because $\mathrm{E}_0 l'_{\theta_0}[\theta - \theta_0, Y] \neq 0$ (see Example 5).

REMARK 5. Theorems 1 and 2 continue to hold for independent but nonidentically distributed observations if the corresponding quantities are replaced by the average quantities on each observation, and these average quantities are stable in probability law.

The phenomenon of compensation between the degree of smoothness of the functional and the convergence rate of the estimate, as mentioned in the Introduction, can be seen directly from Theorem 1, that is, $\delta_n^\omega = O(n^{-1/2})$. Here the trade-off phenomenon between the approximation errors specified in Condition C and the size of the sieve which determines the stochastic approximations in Conditions A, C and D occurs. The result in Theorem 1 may not hold when the approximation errors related to the sieve approximation are not small due to the size of the space. However, if the related approximation errors are substantially smaller when the special structures such as orthogonality are exploited, then the above result continues to hold even for a larger parameter space. Consequently, the result in Theorem 1 can be sharpened. Some modified regularity conditions are formulated below. A more detailed discussion is deferred to Section 7.

We now focus on the case in which $P_n = \pi_n$ is linear, $P_n(\theta^*(\theta, \varepsilon_n)) = (1 - \varepsilon_n)\theta + \varepsilon_n(\pi_n u^* + \pi_n \theta_0)$ for $\theta \in \{\theta \in \Theta_n \colon \|\theta - \theta_0\| \leq \delta_n\}$, although the corresponding results for the nonlinear case can be established in the same way as the constrained case in Section 6. Suppose the sieve Θ_n has the following orthogonality property:

$$\langle v_1, v_2 \rangle = 0 \quad \text{for any } v_1 \in \Theta_n - \pi_n \theta_0,\ v_2 \in \Theta \setminus \Theta_n - \pi_n \theta_0. \tag{4.3}$$

CONDITION B′ (Expectation of criterion difference). For some positive sequence $\{h_n\} \to 0$ as $n \to \infty$,

$$\sup_{\{\theta \in \Theta_n \colon 0 < \|\theta - \theta_0\| \leq \delta_n\}} K(\theta_0, \theta) = \tfrac{1}{2}\|\theta - \theta_0\|^2 (1 + o(h_n)).$$

CONDITION C′ (Approximation error). We have $\|\pi_n v^* - v^*\| = O(\delta_n^{-1}\varepsilon_n)$ and $|o(h_n)|\|\pi_n \theta_0 - \theta_0\|^2 = O(\varepsilon_n)$.

COROLLARY 2. *Under* (4.3), *Theorem* 1 *continues to hold if Conditions* B *and* C *are replaced by Conditions* B′ *and* C′, *respectively*.

REMARK 6. Condition C′ can be easily satisfied since the approximation error of v^* is often smaller than that of θ_0. See Example 2 for an illustration.

4.2. *The method of penalization.* We now formulate the modified regularity conditions. Let u^* be the same as defined in Section 4.1, and the convergence rate of the penalized estimate under $\|\cdot\|$ and $\|\cdot\|_s$ be $o_p(\delta_n)$ and $o_p(\delta_n^s)$, respectively, where $\|\cdot\| \le d\|\cdot\|_s$ ($\|\cdot\|_s$ is often chosen as the Sobolev norm when the parameter space Θ is related to a Sobolev space). Furthermore, let $\varepsilon_n = o(n^{-1/2})$ and $\theta^{**}(\theta, \varepsilon_n) = (1-\varepsilon_n)\theta + \varepsilon_n(u^* + \theta_0) \in \Theta$ for any $\theta \in \{\theta \in \Theta\colon \|\theta - \theta_0\|_s \le \delta_n^s\}$.

Suppose f has the following smoothness property: for all $\theta \in \{\theta \in \Theta\colon \|\theta - \theta_0\|_s \le \delta_n^s\}$,

$$|f(\theta) - f(\theta_0) - f'_{\theta_0}[\theta - \theta_0]| \le O(\|\theta - \theta_0\|^\omega) \quad \text{as } \|\theta - \theta_0\| \to 0, \tag{4.4}$$

where $\omega > 0$ is the degree of smoothness of f at θ_0, and $f'_{\theta_0}[\theta - \theta_0]$ is linear in $(\theta - \theta_0)$ and $\|f'_{\theta_0}\| < \infty$.

CONDITION A″ (Stochastic equicontinuity). For $r[\cdot,\cdot]$ defined in (4.1), the following hold:

(i) $$\sup_{\{\theta\in\Theta\colon \|\theta-\theta_0\|_s \le \delta_n^s\}} n^{-1/2}\nu_n(r[\theta - \theta_0, Y] - r[\theta^{**}(\theta, \varepsilon_n) - \theta_0, Y]) = O_p(\varepsilon_n^2),$$

(ii) $$\sup_{\{\theta\in\Theta\colon \|\theta-\theta_0\|_s \le \delta_n^s\}} n^{-1/2}\nu_n(r[\theta - \theta_0, Y]) = O_p(\varepsilon_n).$$

CONDITION B″ (Expectation of criterion difference). We have

$$\sup_{\{\theta\in\Theta\colon 0<\|\theta-\theta_0\|_s \le \delta_n^s\}} \big[K(\theta_0, \theta^{**}(\theta, \varepsilon_n)) - K(\theta_0, \theta)\big] - \tfrac{1}{2}\big[\|\theta^{**}(\theta, \varepsilon_n) - \theta_0\|^2 - \|\theta - \theta_0\|^2\big] = O(\varepsilon_n^2).$$

CONDITION C″ (Penalty). For some constant $c > 0$ and any $\theta_i \in \{\theta \in \Theta\colon \|\theta - \theta_0\|_s \le \delta_n^s\}$, $i = 1, 2$,

$$J(\theta_1 + \theta_2) \le c(J(\theta_1) + J(\theta_2)).$$

In addition, $\lambda_n = O(\varepsilon_n)$ and $J(v^*) < \infty$.

CONDITION D″ (Gradient). We have

$$\sup_{\{\theta\in\Theta\colon \|\theta-\theta_0\|_s \le \delta_n^s\}} n^{-1/2}\nu_n(l'_{\theta_0}[\theta - \theta_0, Y]) = O_p(\varepsilon_n).$$

THEOREM 2 (Normality). *In addition to Conditions* A″–D″, *let* f *satisfy* (4.4) *with* $\delta_n^\omega = O(n^{-1/2})$ *and* $\mathrm{Var}_0(l'_{\theta_0}[v^*, Y]) < \infty$. *Then, for the approximate substitution penalized estimate* $f(\hat\theta_n)$ *defined in* (2.2),

$$n^{1/2}(f(\hat\theta_n) - f(\theta_0)) \to_{\mathrm{p}_{\theta_0}} N(0, \mathrm{Var}_0(l'_{\theta_0}[v^*, Y])).$$

COROLLARY 3. *If Conditions* A″–D″ *hold, then for the approximate penalized estimate defined in* (2.2),

$$n^{1/2}\langle \hat{\theta}_n - \theta_0, s\rangle \to_{\mathrm{p}_{\theta_0}} N(0, \mathrm{Var}_0(l'_{\theta_0}[s, Y])),$$

where $s \in \Theta - \theta_0$.

REMARK 7. Conditions A″, B″ and D″ are similar to Conditions A, B and D. Often, when the size of the parameter space, measured by the metric entropy, is not large, the supremum of $r[\theta - \theta_0, Y]$ over a δ_n-neighborhood of θ_0 is of order $O_p(\varepsilon_n^2)$, then the verification of Condition A″(i) is not necessary. Condition C″ and Condition A″(ii) are used to control the behavior of $J(\hat{\theta}_n)$.

5. Efficiency.

5.1. *Lower bound and LAN.* In this section, we restrict our attention to the ML estimation in which the criterion is a log-likelihood. It can be seen from the finite-dimensional case that the notion of locally asymptotic normality (LAN) plays an important role in establishing the lower bounds for evaluating asymptotic performance of estimates.

We begin with a discussion on the LAN. We say a family $(P_\theta,\ \theta \in \Theta)$ is locally asymptotically normal at θ_0 if there exists a normalized factor A_n such that (1) $A_n \to 0$ as $n \to \infty$, (2) for any $h \in V$, $\theta_0 + tA_n h \in \Theta$ if t is small and (3)

$$\frac{dP_{\theta_0 + A_n h}}{dP_{\theta_0}}(Y_1, \ldots, Y_n) = \exp\left(\Sigma_n(h) - \frac{1}{2}\|h\|^2 + R_n(\theta_0, h)\right),$$

where $\Sigma_n(h)$ is linear in h, $\Sigma_n(h) \to_{\mathrm{p}_{\theta_0}} N(0, \|h\|^2)$ and $R_n(\theta_0, h) \to_{\mathrm{p}_{\theta_0}} 0$. The above LAN is a version of "locally asymptotic normality" in general parameter spaces, which is an extension of the LAN in the parametric case [see Le Cam (1960) and Ibragimov and Has'minskii (1991)].

To avoid the "superefficiency" phenomenon, certain conditions on estimates are required. The Fisher information lower bound can then be established for a class of "*regular* estimates" in which the superefficiency phenomenon does not occur. In the finite-dimensional case, Bahadur (1964) established the Fisher information lower bound for the class of "*regular* estimates" consisting of "asymptotic median unbiased" estimates. Hájek (1970) developed the convolution result for the class of estimates with asymptotic distribution representations. In estimating a smooth functional in the infinite-dimensional case, Wong (1992) established the Fisher information lower bound in terms of probability concentration. Ibragimov and Has'minskii (1991) also obtained Hájek's representation theorem. Related works can be found in Levit (1974, 1978), Begun, Hall, Huang and Wellner (1983) and Ibragimov and Has'minskii (1981).

We now define the class of *pathwise regular* estimates in the sense of Bahadur (1964) and Wong (1992). Intuitively, the class of pathwise regular estimates should be as large as possible so that an attainment of the Fisher

lower bound becomes a strong property. We say an estimate $T_n(Y_1, \ldots, Y_n)$ is a *pathwise regular* estimate of f at θ_0 if for any real number $\rho > 0$ and any $h \in V$ we have

$$\limsup_{n\to\infty} \mathrm{P}_{\theta_{n,\rho}}(T_n < f(\theta_{n,\rho})) \le \liminf_{n\to\infty} \mathrm{P}_{\theta_{n,-\rho}}(T_n < f(\theta_{n,-\rho})),$$

where $\theta_{n,\rho} = \theta_0 + A_n \rho h$. This is an extension of the notion of asymptotic median unbiasedness. The following optimality result is a variant of Proposition 4 in Wong (1992) under the above LAN condition which is slightly weaker than Condition L' used in Proposition 4 of that paper.

THEOREM 3 (Lower bound). *In addition to the LAN, suppose f is Frechet-differentiable at θ_0 with $0 < \|f'_{\theta_0}\| < \infty$. Then, for any pathwise regular estimate of f at θ_0 T_n, and any real number $\rho > 0$,*

$$\limsup_{n\to\infty} \mathrm{P}_0(A_n^{-1}|T_n - f(\theta_0)| \le \rho) \le \mathrm{P}_0(|N(0, \|f'_{\theta_0}\|^2)| \le \rho), \tag{5.1}$$

where $N(0, \|f'_{\theta_0}\|^2)$ is a normal distribution with variance $\|f'_{\theta_0}\|^2$.

REMARK 8. The requirement on the degree of smoothness of the functional for this lower bound is weaker than that used for establishing asymptotic normality. In fact, the Frechet-differentiability corresponds to the case of $\omega > 1$ in (4.2) and (4.4). Typically, $A_n = n^{-1/2}$.

5.2. *Asymptotic efficiency of maximum likelihood estimation.* The following result is a trivial consequence of Theorems 1 and 2. It says that (1) the MLE belongs to the class of the pathwise regular estimates and (2) the MLE attains the lower bound in (5.1).

THEOREM 4 (Efficiency). *In addition to the conditions in Theorem* 1 (*Theorem* 2), *if the LAN holds, then for the approximate substitution sieve and penalized estimates of $f(\theta)$, any real number $\rho > 0$ and any $h \in V$,*

$$n^{1/2}(f(\hat{\theta}_n) - f(\theta_n)) \to_{\mathrm{p}_{\theta_n}} N(0, \mathrm{Var}_0(l'_{\theta_0}[v^*, Y])),$$

where $\theta_n = \theta_0 + n^{-1/2}\rho h$.

6. Constrained estimation. In this section, we discuss the estimation in the constrained case in which the true parameter θ_0 is under some restrictions and such restrictions may not be linear. For instance, $\int \theta_0 = 1$ if θ_0 is a density. The situation is slightly complicated since $\mathrm{E}_0 l'_{\theta_0}[\theta - \theta_0, Y]$ may not be zero (see Example 5). Note that $\mathrm{E}_0 l'_{\theta_0}[\theta - \theta_0, Y] = 0$ is a requirement for establishing asymptotic normality for the MLE in the finite-dimensional case. The previous results for the information lower bound are expected to hold if the restrictions are linear only infinitesimally. We now modify the LAN condition in Section 5.1 to take care of the restrictions.

We say a family $(P_\theta,\ \theta \in \Theta)$ is locally asymptotically normal at θ_0 if there exists a normalized factor A_n such that (1) $A_n \to 0$ as $n \to \infty$ and (2)

$$\frac{dP_{m(\theta_0 + A_n h)}}{dP_{\theta_0}}(Y_1, \ldots, Y_n) = \exp\left(\Sigma_n(h) - \frac{1}{2}\|h\|^2 + R_n(\theta_0, h)\right),$$

where $m(\theta_0 + A_n h)$ is the closest point in Θ to $\theta_0 + A_n h$, $\Sigma_n(h)$ is linear in $h \in V$, $\Sigma_n(h) \to_{\mathrm{p}_{\theta_0}} N(0, \|h\|^2)$ and $R_n(\theta_0, h) \to_{\mathrm{p}_{\theta_0}} 0$.

We say an estimate $T_n(Y_1, \ldots, Y_n)$ is a *pathwise regular* estimate of f at θ_0 if, for any real number $\rho > 0$ and any $h \in V$, we have

$$\limsup_{n\to\infty} \mathrm{P}_{m(\theta_{n,\rho})}(T_n < f(m(\theta_{n,\rho}))) \le \liminf_{n\to\infty} \mathrm{P}_{m(\theta_{n,-\rho})}(T_n < f(m(\theta_{n,-\rho}))),$$

where $\theta_{n,\rho} = \theta_0 + A_n \rho h$.

THEOREM 5 (Lower bound). *In addition to the LAN in the constrained case, f is Frechet-differentiable at θ_0 with*

$$0 < \|f'_{\theta_0}\|_m = \limsup_{n\to\infty} \sup_{h\neq 0} \frac{|f'_{\theta_0}[m(\theta_{n,\rho}) - \theta_0]|}{\|\theta_{n,\rho} - \theta_0\|} < \infty.$$

Then, for any pathwise regular estimate of f at θ_0 T_n, and any real number $\rho > 0$,

$$\limsup_{n\to\infty} \mathrm{P}_0(A_n^{-1}|T_n - f(\theta_0)| \le \rho) \le \mathrm{P}_0(|N(0, \|f'_{\theta_0}\|_m^2)| \le \rho).$$

THEOREM 6 (Efficiency). *In addition to the conditions in Theorem 1, if the LAN specified here holds, then for the approximate substitution constrained MLE, any real number $\rho > 0$ and any $h \in V$,*

$$n^{1/2}(f(\hat{\theta}_n) - f(\theta_n)) \to_{\mathrm{p}_{\theta_n}} N(0, \mathrm{Var}_0(l'_{\theta_0}[v^*, Y])),$$

where $\theta_n = \theta_0 + n^{-1/2}\rho h$.

Often, $\|f'_{\theta_0}\|_m^2 = \mathrm{Var}_0(l'_{\theta_0}[v^*, Y])$ in application.

7. Discussion. In estimating a functional, two aspects are essential: (1) the errors in local approximations and (2) the degree of smoothness of the functional. Let $G_1 = \{\theta \in \Theta_n\colon \|\theta - \theta_0\| \le \delta_n\}$ and $G_2 = \{\theta \in \Theta\colon \|\theta - \theta_0\|_s \le \delta_n^s\}$, as specified in Section 4. The stochastic approximations specified in Conditions A, C and D (A′, C′ and D′) are expected to hold when the local size of the underlying space $r_i^* = \log(1/H^B(u, G_i))$, $i = 1, 2$ (for any small $u > 0$) is less than 2. This is because $\sup_{g\in G_i} \nu_n(g) \to 0$ when $\int_0^\delta H^B(u, G_i)\,du < \infty$ [Ossiander (1987)]. Consequently, Theorems 1 and 2 hold in such cases. On the other hand, as known in empirical process theory, the empirical processes specified in the corresponding conditions may converge to zero very slowly when the size of the underlying space is very large. As a result, as illustrated in Example 3, the asymptotic normality result does not hold.

As discussed in Section 4.1, when the parameter space is very large, certain orthogonality properties are useful in reducing the approximation errors. This suggests the following construction scheme using the method of sieves. Let $\Theta_n = \{\theta = \sum_{i=1}^{r_n} a_i \phi_i \in \Theta\}$ (possibly with some restrictions on the sieve), where $\{\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis with respect to $\langle \cdot, \cdot \rangle$. The existence of such an orthonormal basis is guaranteed by the Gram–Schmidt orthogonalization procedure. However, a certain initial estimate of $\langle \cdot, \cdot \rangle$ is required when $\langle \cdot, \cdot \rangle$ depends on unknown θ. Let $\langle\langle \cdot, \cdot \rangle\rangle$ be an estimate of $\langle \cdot, \cdot \rangle$ such that $|\langle v_1, v_2 \rangle - \langle\langle v_1, v_2 \rangle\rangle| = O_p(n^{-1/2})$ for any $v_1 \in \Theta_n - \pi_n \theta_0$ and any $v_2 \in \Theta \setminus \Theta_n - \pi_n \theta_0$. As illustrated in Example 2, such an estimate for $\langle \cdot, \cdot \rangle$ is not necessary when $\langle \cdot, \cdot \rangle$ does not depend on θ. In this setting, the size of this constructed sieve r_n needs to be chosen so that the required conditions for the stochastic approximations are satisfied. Note that when the approximation error of the sieve is reduced by orthogonality, the trade-off phenomenon as discussed in the paragraph before (4.3) no longer exists. The constructed substitution sieve estimate is therefore efficient in the Fisher sense. In contrast, as illustrated in Examples 2 and 3, the standard ML and penalized procedures do not have this flexibility. This phenomenon is particularly manifested when ω is large, such as in linear functional and semiparametric estimations.

We need to point out that the phenomenon that the sieve MLE is asymptotically optimal and the standard and penalized MLE's may not be, as shown in Examples 2 and 3, also occurs in the context of convergence rate. See Birgé and Massart (1993) and Shen and Wong (1994) for the suboptimality of the standard MLE's and the construction of the optimal sieve which achieves the best convergence rate. As shown in Shen and Wong (1994), the size of the sieve should be balanced in order to achieve the best convergence rate. In the efficiency context, the optimal size of the sieve for the convergence rate $\|\hat{\theta}_n - \theta_0\|$ also leads to an efficient estimate $f(\hat{\theta}_n)$ when the index r_1^* is less than 2, whereas there may not exist such a choice when r_1^* is at least 2; that is, the best possible choice of the size of sieve for $f(\hat{\theta}_n)$ may only yield a rate which is close to the suboptimal rate of the corresponding MLE. A similar phenomenon also occurs in penalized estimation, as shown in Examples 2 and 3.

8. Applications. We now apply the general theory to obtain the results (Propositions 1–5) presented in Section 2.

EXAMPLE 1 (Nonparametric regression, continued).

(a) *Estimation without sieve.* The convergence rate of this regression estimate $\|\hat{\theta}_n - \theta_0\|$ is $O_p(n^{-p/(2p+1)})$ when $E|e_1^l| < \infty$ for $l > 2$, which can be found in part (b) of Example 3 of Shen and Wong (1994). It can be verified by a Taylor expansion that

$$\left| f(\theta) - f(\theta_0) - k \int \theta_0^{k-1}(x)(\theta - \theta_0)(x)\, dP_0(x) \right| = O(\|\theta - \theta_0\|^2).$$

Then (4.2) follows with $\omega = 2$ and the representer $v^* = k\sigma^2 \theta_0^{k-1} \in V$.

We now verify Conditions A–D. Note that $P(\theta^*(\theta, \varepsilon_n)) = \theta^*(\theta, \varepsilon_n)$ for $u^* \in V$, where P is the projection specified in Condition A. In addition,

$$r[\theta - \theta_0, y] = l(\theta, y) - l(\theta_0, y) - l'_{\theta_0}[\theta - \theta_0, y] = -(\theta - \theta_0)^2$$

and

$$\begin{aligned} &r[\theta - \theta_0, y] - r[\theta^*(\theta, \varepsilon_n) - \theta_0, y] \\ &\quad = -\varepsilon_n(2 - \varepsilon_n)[(\theta - \theta_0)^2 + 2(\theta - \theta_0)u^*] + \varepsilon_n^2(u^*)^2. \end{aligned}$$

For Condition A, it is sufficient to calculate the convergence rates of supremums of the empirical processes

$$\sup_{\{\theta \in \Theta: \|\theta - \theta_0\| \le \delta_n\}} n^{-1/2} \nu_n((\theta - \theta_0)u^*) \tag{8.1}$$

and

$$\sup_{\{\theta \in \Theta: \|\theta - \theta_0\| \le \delta_n\}} n^{-1/2} \nu_n((\theta - \theta_0)^2). \tag{8.2}$$

Let $S_n = \{(\theta - \theta_0)u^*: \|\theta - \theta_0\| \le \delta_n, \theta \in \Theta\}$. It follows from Kolmogorov and Tihomirov (1959) that $H_2^B(u, S_n) \le H(u, \Theta, \|\cdot\|_{\sup}) \le cu^{-1/p}$ for some constant $c > 0$, where $H(\cdot, \cdot, \|\cdot\|_{\sup})$ is the metric entropy under the supremum norm, defined as the logarithm of the minimal cardinality of the u-covering of S_n in the L_∞-metric [see, e.g., Ossiander (1987)]. Therefore, by Lemma 4 of Shen and Wong (1994), the convergence rate of the empirical process in (8.1) is of order $O_p(n^{-2p/(2p+1)})$. Similarly, the empirical process in (8.2) is also bounded by $O_p(n^{-2p/(2p+1)})$. Hence Condition A holds for $2p/(2p+1) \ge d > 1/2$. We need to point out that the rate $n^{-2p/(2p+1)}$ for the empirical processes in (8.1) and (8.2) cannot be improved in general, even if the knowledge of δ_n is available. Moreover, Condition B is fulfilled because $K(\theta_0, \theta) = \frac{1}{2}\|\theta - \theta_0\|^2$. Condition C is satisfied with $P(\theta^*(\theta, \varepsilon_n)) = \theta^*(\theta, \varepsilon_n)$. Condition D follows from the fact that $\sup_{\{\theta \in \Theta: \|\theta - \theta_0\| \le \delta_n\}} n^{-1/2} \nu_n(l'_{\theta_0}[\theta - \theta_0, Y]) = O_p(n^{-2p/(2p+1)}) = O_p(\varepsilon_n)$. By Theorem 1, we conclude that the approximate substitution regression estimate $f(\hat{\theta}_n)$ is asymptotically normal with variance $\mathrm{Var}_0(l'_{\theta_0}[v^*, Y]) = \|v^*\|^2 = k^2\sigma^2 \mathrm{E}_0(\theta_0^{k-1})^2$ for $p = m + \gamma > 1/2$.

To discuss the efficiency issue, we need to specify the distribution of the error term e_i. Under the assumption that e_i is distributed as $N(0, \sigma^2)$, we have

$$\begin{aligned} &\frac{dP_{\theta_0 + A_n h}}{dP_{\theta_0}}((Y_1, X_1), \ldots, (Y_n, X_n)) \\ &\quad = \exp\left(\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \theta_0(X_i)) A_n h(X_i) - \frac{1}{2\sigma^2} \sum_{i=1}^n A_n h^2(X_i) \right). \end{aligned}$$

The LAN holds with $A_n = n^{-1/2}$, $\Sigma(h) = n^{-1/2} \sum_{i=1}^n (Y_i - \theta_0(X_i))h(X_i)$, and $R_n(\theta_0, h) = n^{-1} \sum_{i=1}^n (h^2(X_i) - \mathrm{E}_0 h^2(X_i))$. By Theorem 4, the standard MLE $f(\hat{\theta}_n)$ is asymptotically efficient for $f(\theta)$.

(b) *Sieve estimation.*

(1) *Finite-dimensional sieve.* In the following, only the case when $p > (2+\sqrt{5})/4$ will be discussed. The result for the case when $1/2 < p < (2+\sqrt{5})/4$ is exactly the same except for the rate calculation [see Shen and Wong (1994), Example 3]. The detailed calculations will not be repeated here. The approximation error of this sieve is well known [see, e.g., Lorentz (1966)]. For any $\theta \in \Theta$, there exists $\pi_n\theta \in \Theta_n$ such that

$$\|\pi_n\theta - \theta\| \le \sup_x |\pi_n\theta(x) - \theta(x)| \le c(\theta) r_n^{-p},$$

for $c(\theta) > 0$ depending on θ. The convergence rate of the sieve estimate is

$$O_p\big(\max\big(n^{-(1-\tau)/2}, n^{-\tau p}\big)\big), \tag{8.3}$$

as given in part (a) of Example 3 of Shen and Wong (1994). The best possible convergence rate for the sieve estimate under $\|\cdot\|$ can be obtained from (8.3). Note that $P_n(\theta^*(\theta, \varepsilon_n)) = (1-\varepsilon_n)\theta + \varepsilon_n(\pi_n u^* + \pi_n\theta_0)$ for $u^* \in \Theta$, and

$$\begin{aligned} &r[\theta - \theta_0, y] - r[P_n(\theta^*(\theta, \varepsilon_n)) - \theta_0, y] \\ &\quad = -(2-\varepsilon_n)\varepsilon_n(\theta-\theta_0)^2 + 2(1-\varepsilon_n)\varepsilon_n(\theta-\theta_0)(\pi_n u^* + \pi_n\theta_0 - \theta_0) \\ &\qquad + \varepsilon_n^2[\pi_n u^* + \pi_n\theta_0 - \theta_0]^2; \end{aligned}$$

it is then sufficient to calculate

$$\sup_{\{\theta\in\Theta:\, \|\theta-\theta_0\|\le\delta_n\}} n^{-1/2}\nu_n((\theta - \pi_n\theta_0)(\pi_n u^* + \pi_n\theta_0 - \theta_0)) \tag{8.4}$$

and

$$\sup_{\{\theta\in\Theta:\, \|\theta-\theta_0\|\le\delta_n\}} n^{-1/2}\nu_n((\theta-\theta_0)^2) \tag{8.5}$$

in the verification of Condition A. Let $S_n = \{(\theta-\theta_0)(\pi_n u^* + (\pi_n\theta_0 - \theta_0))\colon \theta \in \Theta_n, \|\theta-\theta_0\| \le \delta_n\}$. Applying an argument similar to that in part (a) of Example 3 of Shen and Wong (1994), we obtain after some calculations that $H_2^B(u, S_n) \le H(u, \Theta_n, \|\cdot\|_{\sup}) \le c r_n \log(\delta_n/u)$ for some constant $c > 0$. By Lemma 4 of Shen and Wong (1994), the empirical process in (8.4) is bounded by $O_p(n^{-(1-\tau)})$. Similarly, the empirical process in (8.5) is also of order $O_p(n^{-(1-\tau)})$. Thus Condition A is fulfilled if $(1-\tau) \ge d > 1/2$. Conditions B and D can be verified easily. It remains to verify Condition C. The second condition in Condition C can be checked easily. Note that $\|P_n(\theta^*(\theta, \varepsilon_n)) - \theta^*(\theta, \varepsilon_n)\| = \varepsilon_n\|\pi_n(u^* + \theta_0) - (u^* + \theta_0)\|$. Thus Condition C holds if

$$n^{-p\tau} = \varepsilon_n\delta_n^{-1}. \tag{8.6}$$

Then τ can be determined by (8.3) and (8.6). The solution is $1/(4p) < \tau < 1/2$. A natural choice of τ is $1/(2p+1)$ when $p > 1/2$.

By Theorems 3 and 4, the approximate substitution sieve estimate is asymptotically efficient for $p > 1/2$ ($\tau = 1/(2p+1)$ and $l_n \le n^{(2p-1)/(2p'(2p+1))}$).

(2) *Infinite-dimensional sieve.* We now calculate the convergence rate of this sieve estimate. By the Sobolev embedding theorem [Zeidler (1990)] and a

result from Kolmogorov and Tihomirov (1959), $H(u, \Theta_n, \|\cdot\|_{\sup}) \le c(b_n/u)^{1/p}$ for any small $\varepsilon > 0$ and some constant $c > 0$. It now follows from Theorem 2 of Shen and Wong (1994) with $r_0 = (\log b_n^{1/p})/2$ and $\mathrm{E}|e_1|^l < \infty$, $l > 2$, that the convergence rate of the sieve estimate under $\|\cdot\|$ is $O_p(n^{-p/(2p+1)} b_n^{1/(2p+1)})$ for $p > 1/2$.

Note that $P\theta^*(\theta, \varepsilon_n) = \theta^*(\theta, \varepsilon_n)$ for any large n. Conditions A–D can be verified easily by arguments similar to those in parts (a) and (b). Therefore, by Theorems 2 and 3, the approximate substitution sieve MLE is asymptotically efficient.

(c) *Penalization.* Under the assumption that $\mathrm{E}\exp(t_0|e_1|) < \infty$ for some $t_0 > 0$, the convergence rate of the penalized estimate under $\|\cdot\|$ and $\|\cdot\|_s$ are $O_p(n^{-p/(2p+1)})$ and $O_p(n^{-(p-s)/(2p+1)})$, respectively, with penalization coefficient $\lambda_n = n^{-2p/(2p+1)}$ [see, e.g., Shen (1997a), Example 3]. Clearly, Conditions A″, B″ and D″ can be verified as in part (b). Furthermore, Condition C″ follows from Minkowski's inequality and (4.4) is implied by a Taylor expansion and the Sobolev embedding theorem [Zeidler (1990)]. Therefore, the approximate substitution penalized MLE is asymptotically efficient for $p > 1/2$.

EXAMPLE 2 (Semiparametric model, continued).

(a) *Sieve estimation.* Let $\theta = (\beta, \eta)$ and $\theta_0 = (\beta_0, \eta_0)$. Let $\|\theta - \theta_0\|^2 = \mathrm{E}_0(\lambda - \lambda_0)^2/\sigma^2 = \mathrm{E}_0[(\beta - \beta_0)Z + (\eta(X) - \eta_0(X))]^2/\sigma^2$. To determine the representer v^*, note that $f(\theta) = \beta$ is linear. It is easy to see that (4.2) is satisfied with $\omega = \infty$ and $f'_{\theta_0}[\theta - \theta_0] = \beta - \beta_0$ since $f(\theta)$ is linear. By definition,

$$\|v^*\|^2 = \sup_{\{\theta-\theta_0:\, \|\theta-\theta_0\|>0\}} \frac{(\beta-\beta_0)^2}{\|\theta-\theta_0\|^2} = \sup_{\{h:\, \|Z+h\|>0\}} \frac{1}{\|Z+h\|^2}.$$

It suffices to find the minimizer h^* of $\|Z + h\|^2 = \mathrm{E}(Z + h(X))^2/\sigma^2$ (h^* is often called the least favorable direction). By a conditional argument, we obtain $h^* = -\mathrm{E}\, Z|X$. Thus, $\|v^*\|^{-2} = \mathrm{E}\,(Z - \mathrm{E}\, Z|X)^2/\sigma^2$ and $v^* = \|v^*\|^2(1, -\mathrm{E}\, Z|X)$, which agrees with the usual definition of the minimal Fisher information for β [see Lindsay (1980)]. Here $v^* \in V$ because the conditional density of $Z|X$ is smooth enough.

Consider the first case when $p = m + \gamma > 1/2$. Note that

$$H_2^B(u, \Theta_n) \le H_2^B(u, A) + H_2^B(u, B_n) \le \log\left(\frac{1}{u}\right) + cr_n \log\left(\frac{\delta_n}{u}\right),$$

then Conditions A–D and the LAN can be verified as in Example 1 with replacement of θ in Example 1 by λ here. By Theorems 3 and 4, the approximate sieve MLE $\hat{\beta}_n$ is asymptotically efficient.

Now consider the case when $p = m + \gamma \le 1/2$. Conditions A, B′, C′ and D and (4.3) are satisfied with $d \ge (1 - \tau) > 1/2$. In the above verification, $\|\pi_n v^* - v^*\|$ is required to be $O(n^{-d})$. This is so because $\mathrm{E}\, Z|X$ is smooth enough. By Corollary 2, the sieve MLE $\hat{\beta}_n$ is asymptotically efficient.

It is interesting to note that the convergence rate of the sieve estimate under $\|\cdot\|$ is $O_p(n^{-\tau p})$, which is close to $n^{-p/2}$ when τ is close to $1/2$. This suboptimal

rate is actually the rate of the standard MLE based on the parameter space Θ in Example 1(a) for $p < 1/2$ [see part (b) of Example 3 of Shen and Wong (1994) for discussions of this suboptimality].

(b) *Penalization.* Conditions A″–D″, (4.4) and the LAN can be verified. In addition, $J(v^*) < \infty$ because $v^* \in \Theta$. By Theorems 3 and 4, we conclude that the approximate penalized MLE $\hat{\beta}_n$ is asymptotically efficient.

EXAMPLE 4 [The proportional odds model (right censoring), continued]. Let θ_0 be $(\beta_0, \Gamma_0) = ((\beta_1^0, \ldots, \beta_p^0), \Gamma_0)$, $X = (X^{(1)}, \ldots, X^{(p)})$. After some calculations, we obtain that $l'_{\theta_0}[\theta - \theta_0, \delta, Y, X] = \sum_{j=1}^p \tilde{A}_j(\beta_j - \beta_j^0) + \tilde{B}(\Gamma'(Y) - \Gamma_0'(Y)) + \tilde{C}(\Gamma(Y) - \Gamma_0(Y))$, where $\tilde{A}_j$, $\tilde{B}$ and $\tilde{C}$ are given in Section 3. Define

$$\|\theta - \theta_0\|^2 = \mathrm{E}_0\Bigg[\sum_{j=1}^p \tilde{A}_j(\beta_i - \beta_i^0) + \tilde{B}(\Gamma'(Y) - \Gamma_0'(Y)) + \tilde{C}(\Gamma(Y) - \Gamma_0(Y))\Bigg]^2.$$

By Theorem 8.2 of Devore and Lorentz (1991),

$$\begin{aligned}
\|\pi_n\theta_0 - \theta_0\|_2 &\le O(\|\tilde{\Gamma} - \Gamma_0\|_2 + \|\tilde{\Gamma}' - \Gamma_0'\|_2) \\
&\le O\left(\left\|\int_0^t [\exp(\log\Gamma_0'(x)) - \exp(s(x))]\,dx\right\|_2\right) \\
&\le O(\|\log\Gamma_0'(x) - s(x)\|_2) \\
&\le O\big(r_n^{-s}\|\log\Gamma_0'\|_{B^s_{(s+1/q)^{-1},(s+1/p)^{-1}}}\big) \\
&\le O(r_n^{-s}C).
\end{aligned}$$

Similarly, by Theorem 8.2 of Devore and Lorentz (1991) (the case of $p = \infty$ follows by passing the limit $p \to \infty$ and by the fact that $\log\Gamma_0'$ is continuous), we have

$$\begin{aligned}
\|\pi_n\theta_0 - \theta_0\|_{\sup} &\le O\big(r_n^{-s}\|\log\Gamma_0'\|_{B^s_{s^{-1},s^{-1}}}\big) \\
&\le O(r_n^{-s}C).
\end{aligned}$$

Furthermore, for $\theta_i = (\beta^{(i)}, \Gamma_i)$, $i = 1, 2$,

$$\begin{aligned}
|l(\theta_1, \delta, y, z) - l(\theta_2, \delta, y, z)| \le c|\log\Gamma_1'(y) - \log\Gamma_2'(y)| \\
+ \sum_{j=1}^p |\beta^{(1)} - \beta^{(2)}| + |\Gamma_1(y) - \Gamma_2(y)|.
\end{aligned}$$

By Theorem 2 of Shen and Wong (1994), the convergence rate of this sieve estimate under $\|\cdot\|$ is $O_p(\delta_n)$ with $\delta_n = \max(n^{-1/2}r_n^{1/2}, r_n^{-s}) = n^{-s/(2s+1)}$ with $r_n = n^{1/(2s+1)}$. For more details about this type of rate calculation, see part (c) of Example 3 of Shen and Wong (1994).

By the assumptions, we know that $\|\Gamma_1 - \Gamma_2\|_2 + \|\Gamma_1' - \Gamma_2'\|_2 \le c_1\|\theta_1 - \theta_2\|$ and $\|\beta^{(1)} - \beta^{(2)}\|_E \le c_2\|\theta_1 - \theta_2\|$, where $\|\cdot\|_2$ is the usual L_2-norm and $\|\beta^{(1)} - \beta^{(2)}\|_E =$

$(\sum_{i=1}^{p}(\beta_i^{(1)} - \beta_i^{(2)})^2)^{1/2}$ is the usual Euclidean norm. These facts will be used in rate calculations and verifications of Conditions A–D.

Let $f(\theta) = \beta^T t$, where $t^T = (t_1, \ldots, t_k)$ is an arbitrary vector. Then (4.2) is satisfied with $\omega = \infty$. We now calculate v^* and the corresponding Fisher information for $f(\theta)$. By definition,

$$\begin{aligned}\|v^*\|^2 &= \sup_{\{\theta-\theta_0:\|\theta-\theta_0\|>0\}} \frac{((\beta-\beta_0)^T t)^2}{\|\theta-\theta_0\|^2} \\ &= \sup_{\{(\tau_1,\ldots,\tau_p,h)\}} \frac{(s^T t)^2}{\|\sum_{i=1}^{p}\tau_i(e_i,0)+(0,\ldots,0,h)\|^2},\end{aligned}$$

where $s = (\tau_1, \ldots, \tau_p)$ and e_i is the p-dimensional vector with a one in the ith location and zero elsewhere. Here the minimizer v^* exists with $\|v^*\| < \infty$ because there does not exist a $h(Y)$ which is independent of X such that $\|\sum_{i=1}^{p}\tau_i(e_i,0)+(0,\ldots,0,h)\| = 0$ for a nonzero $s = (\tau_1, \ldots, \tau_p) \in \mathscr{R}^p$. It follows from the relationship between the L_2-norm and the $\|\cdot\|$ that $\|h^*\|_2 < \infty$ and $\|(h^*)'\|_2 < \infty$. Unfortunately, the above infinite-dimensional optimization problem does not appear to have an explicit solution even for $p = 1$. When $p = 1$, the above optimization problem is equivalent to $\inf_h \|(1,0)+(0,h)\|^2 = \inf_h \mathrm{E}_0(\tilde{A}_1 + \tilde{B}h'(Y) + \tilde{C}h(Y))^2$.

To characterize v^*, we use finite-dimensional approximations. Let $\mu = \int(p(y)/\Gamma_0'(y))\,dy$, where $p(y)$ is the density of Y. Let $h = \sum_{j=1}^{\infty} w_j\psi_j(x)$, where $\{\psi_j\}$ is an orthonormal basis (Gram–Schmidt orthogonalization based on the trigonometric basis in $\mathscr{L}_2[0,U]$) with respect to the L_2-norm induced by μ, denoted by $\|\cdot\|_{2,\mu}$. Let $h_n = \sum_{j=1}^{s_n} w_j\psi_j(x)$. Consider the following finite-dimensional approximations:

$$\begin{aligned}\|v_n^*\|^2 &= \sup_{\{(s,w)\in\mathscr{R}^p\times\mathscr{R}^{s_n}\}} \frac{(s^T t)^2}{\|\sum_{i=1}^{p}\tau_i e_i^{(1)} + \sum_{i=1}^{s_n} w_i e_i^{(2)}\|^2} \\ &= \sup_{\{(s,w)\in\mathscr{R}^p\times\mathscr{R}^{s_n}\}} \frac{(s^T t)^2}{s^T I_{11}s + 2s^T I_{12}(s_n)h + w^T I_{22}(s_n)w} \\ &= \sup_{\{(s,w)\}} \Big((s^T t)^2\Big[s^T(I_{11} - I_{12}(s_n)I_{22}^{-1}(s_n)I_{12}^T(s_n))s \\ &\qquad + (w + I_{22}^{-1}(s_n)I_{12}^T(s_n)s)^T I_{22}(s_n)(w + I_{22}^{-1}(s_n)I_{12}^T(s_n)s)\Big]^{-1}\Big) \\ &= t^T(I_{11} - I_{12}(s_n)I_{22}^{-1}(s_n)I_{12}^T(s_n))^{-1}t,\end{aligned}$$

where $e_i^{(1)}$ and $e_j^{(2)}$ are the $(p+s_n)$-dimensional vectors with a one in the ith location and zero elsewhere and with $\psi_j(Y)$ in the $(p+j)$th location and zero elsewhere, respectively. Here $w = (w_1, \ldots, w_{s_n})$, I_{11}, $I_{12}(s_n) = I_{21}(s_n)$, $I_{22}(s_n)$ and $I_n = I_{11} - I_{12}(s_n)I_{22}^{-1}(s_n)I_{12}^T(s_n)$ are defined in Section 3. The optimizer $\tilde{v}_n^* = (I_n^{-1}t, h_n^*)$, where $h_n^* = \sum_{j=1}^{s_n} w_j^*\psi_j$ with $w^*(s_n) = (w_1^*, \ldots, w_{s_n}^*) = -I_{22}^{-1}(s_n)I_{12}^T(s_n)(I_n^{-1}t)$.

Under the assumption on the censoring distribution, we know that $g_{i1}(y) = \mathrm{E}_0(\tilde{A}_i\delta|Y = y)$ and $g_{i2}(y) = \mathrm{E}_0(\tilde{A}_i\tilde{C}|Y = y)$ have the same amount of smoothness as $\Gamma_0(y)$ with $\int g_{ij}^2(y)\,dy < \infty$ for $i = 1, \ldots, p$ and $j = 1, 2$. This implies that, for any k, $\|I_{12}(k)\|_M^2 \le C(\sum_{i=1}^p \sum_{j=1}^k (\int g_{i1}(y)\psi'_j(y)\,d\mu(y))^2 + (\int g_{i2}(y)\psi_j(y)\,d\mu(y))^2)$ is bounded by $\sum_{i=1}^p (\|g'_{i1}\|_{2,\mu}^2 + \|g_{i2}\|_{2,\mu}^2) < \infty$, where $\|\cdot\|$ is the corresponding matrix (vector) norm. Note here that

$$\inf_{\{\|s\|=1\}} \frac{u^T I_{22}(k) u}{u^T u} = \lambda,$$

where $u^T = (u_1, \ldots, u_k)$ is an arbitrary vector and λ is the smallest eigenvalue of $I_{22}(k)$. Hence, for any integer k, $\lambda \ge \inf_{\{\|u\|=1\}} u^T I_{22}(k) u \ge \inf_{\{\|h\|_{2,\mu}=1\}} E(\tilde{B}h'(Y) + \tilde{C}h(Y))^2$, which is bounded from below by an argument similar to that for the existence of $\|v^*\|$. This implies that, for any integer k, $\|I_{22}^{-1}(k)\| \le 1/\lambda$ is bounded. To show that h_n^* converges in L_2, note that

$$\|w_k^*\|_M^2 \le \|I_{22}^{-1}(k)\|_M^2 \|I_{12}^T(k)\|_M^2 \|I_n^{-1} t\|_M^2.$$

The convergence of h_n^* in L_2 then follows from the fact that $\|I_n\|_M$ is bounded.

Note that $\|v_n^*\|^2 \to \|v^*\|^2$. This implies that $\lim_{n\to\infty} I_n$ exists. Let $\tilde{h}^* = \sum_{j=1}^\infty w_j^* \psi_j$ and $\tilde{v}^* = (I^{-1}t, \tilde{h}^*)$, where $I = \lim_{n\to\infty} I_n$. Applying a similar argument to that above, we know that $\tilde{h}^*$ has the same amount of smoothness as Γ_0. This implies that $\tilde{h}^*$ belongs to the completion of the space spanned by $\Gamma - \Gamma_0$. Write v^* as (s^*, h^*), where $h^* = \sum_{j=1}^\infty a_j \psi_j$. Define $v_n^* = (s^*, h_n^*)$, where $h_n^* = \sum_{j=1}^{s_n} a_j \psi_j$. By definition, for any s_n, $\|v_n^*\|^2 \ge \|\tilde{v}_n^*\|^2$, which implies that $\|v^*\|^2 = \|\tilde{v}^*\|^2$. By the uniqueness of the Riesz representer, $v^* = \tilde{v}^*$. The above construction of v^* is based on $\{\psi_j\}$. Because $\|\cdot\|_{2,\mu}$ is equivalent to $\|\cdot\|_2$, the h_n^* based on the trigonometric basis also leads to an approximated information $\|v_n^*\|^2$ that converges to $\|v^*\|^2$.

For any $\theta \in \Theta_n$, let $P_n\theta = (\beta, P_n\Gamma)$, where

$$P_n\theta^*(\theta, \varepsilon_n) = (\beta + \varepsilon_n \tau^*, P_n\Gamma(\theta, \varepsilon_n)),$$

$$P_n\Gamma(\theta, \varepsilon_n)(t) = \int_0^t \exp(s(x) + \varepsilon_n \pi_n[\exp(-s(x))(h^*(x))']) \, dx$$

and $\Gamma(t) = \int_0^t \exp(s(x))\,dx$. Applying Theorem 8.2 of Devore and Lorentz (1991), after some calculations, we have

$$\sup_{\{\theta\in\Theta_n: \|\theta-\theta_0\|\le\delta_n\}} \|P_n\theta^*(\theta, \varepsilon_n) - \theta^*(\theta, \varepsilon_n)\| = O(\varepsilon_n r_n^{-s}) + O(\varepsilon_n^2) = O_p(\delta_n^{-1}\varepsilon_n^2).$$

Thus (i) of Condition C is satisfied for any $s > 1/2$.

Applying a Taylor expansion up to order 5 and Lemma 4 of Shen and Wong (1994), the rate of convergence of the empirical process in Condition A is $O_p((\varepsilon_n^2 + \varepsilon_n r_n^{-s} + \varepsilon_n) n^{-1} r_n) + O_p(\delta_n^4) = O_p(\varepsilon_n^2)$. Thus Condition A is satisfied

for any $s > 1/2$. Applying the same argument, we have

$$\left|\left[K(\theta_0, \theta) - K(\theta_0, P_n\theta^*(\theta, \varepsilon_n))\right] - \tfrac{1}{2}\left[\|\theta - \theta_0\|^2 - \|P_n\theta^*(\theta, \varepsilon_n) - \theta_0\|^2\right]\right| = O\Big((\varepsilon_n^2 + \varepsilon_n r_n^{-s} + \varepsilon_n)((\delta_n^2 + \delta_n^{8/3})\Big) + O(\delta_n^4) = O(\varepsilon_n^2).$$

Condition B is satisfied for any $s > 1/2$. For (ii) of Condition C, applying Lemma 4 of Shen and Wong (1994) as in the verification of Condition A and using the relationship between $\|\cdot\|$ and the L_2-norm, we know that the rate of convergence of the empirical process there is $O_p((\varepsilon_n^2 + \varepsilon_n r_n^{-s})n^{-1}r_n) = O(\varepsilon_n^2)$. Therefore, (ii) of Condition C is fulfilled for any $s > 1/2$. Condition D can be verified using the same argument. Finally, the LAN is satisfied with $A_n = n^{-1/2}$ and $\Sigma(h) = n^{-1/2}\sum_{i=1}^n l'_{\theta_0}[h, \delta_i, Y_i, X_i]$.

By Theorems 3 and 4, $\hat{\beta}_n^T t$ is asymptotically efficient for estimating $\beta^T t$ with variance $t^T I^{-1} t$. Therefore, the sieve MLE $\hat{\beta}_n$ is asymptotically efficient for β with covariance matrix I^{-1}.

EXAMPLE 5 (Density estimation, continued). As mentioned in Section 4, the Fisher norm cannot be directly used. We therefore consider the Hellinger distance instead. Let

$$\|\theta - \theta_0\| = 2\left(\int_0^1 (\theta(x) - \theta_0(x))^2\, dx\right)^{1/2} \quad \text{and} \quad \varepsilon_n = n^{-d},$$

where d will be specified later.

(a) *Functions with finite amount of smoothness.* The convergence rate of the standard MLE under $\|\cdot\|$ is $O_p(n^{-p/(2p+1)})$ for $p > 1/2$ [see Wong and Shen (1995), Example 1]. It is easy to see that (4.2) is satisfied with $\omega = 2$ and $v^* = -\theta_0(\log\theta_0 - \int \theta_0^2 \log\theta_0)$ following from a Taylor expansion and the facts that $\int \theta_0(\theta - \theta_0) = 0$, and

$$\left|f(\theta) - f(\theta_0) - \int 4(\theta - \theta_0)\theta_0(\log\theta_0 - \int \theta_0^2 \log\theta_0)\right| \le O(\|\theta - \theta_0\|^2).$$

We now proceed to verify Conditions A–D. Consider the path $\theta(\theta_0, t) = (\theta_0 + t(\theta - \theta_0))/(1 + (t^2 - t)\|\theta - \theta_0\|^2)^{1/2} \in \Theta$ for $0 \le t \le 1$ and $\theta \in \{\theta \in \Theta\colon \|\theta - \theta_0\| \le \delta_n\}$. Then $P(\theta^*(\theta, \varepsilon_n)) = \theta^*(\theta, \varepsilon_n)/[\int \theta^*(\theta, \varepsilon_n)^2]^{1/2} = \theta^*(\theta, \varepsilon_n)/[1 + 2\varepsilon_n \int u^*(\theta - \theta_0) + \varepsilon_n^2\|u^*\|^2]^{1/2}$ for $u^* \in V$, where P is the projection specified in Condition A. Note that $l'_{\theta_0}[\theta - \theta_0] = 2(\theta - \theta_0)/\theta_0 + \|\theta - \theta_0\|^2$ and $\nu_n(r[\theta - \theta_0, y]) = 2\nu_n([\log(1 + (\theta - \theta_0)/\theta_0) - (\theta - \theta_0)/\theta_0])$. Applying a Taylor expansion, we obtain, for $\theta \in \{\theta \in \Theta\colon \|\theta - \theta_0\| \le \delta_n\}$,

$$\begin{aligned} &r[\theta - \theta_0, y] - r[P\theta^*(\theta, \varepsilon_n) - \theta_0, y] \\ &\quad = [(P\theta^*(\theta, \varepsilon_n) - \theta)/\theta_0][(\theta - \theta_0)/\theta] + r^{(1)}(y) \\ &\quad \le \varepsilon_n(\theta - \theta_0)/\theta_0 + u^*(\theta - \theta_0)/\theta_0\theta + r^{(1)}(y), \end{aligned}$$

where $r^{(1)}$ is the remainder in the expansion. To bound $r^{(1)}$, note that $\int \theta_0 u^* = 0$. It follows that

$$\sup_{\{\theta\in\Theta:\,\|\theta-\theta_0\|\le\delta_n\}} |r^{(1)}| = O\Big(\sup_{\{\theta\in\Theta:\,\|\theta-\theta_0\|\le\delta_n\}} \|P\theta^*(\theta,\varepsilon_n)-\theta\|^3\Big) = O(\varepsilon_n^2).$$

By Proposition 6, $\|\theta-\theta_0\|_{\sup} \le O(\delta_n^{2p/(2p+1)})$ implies that the likelihood ratios θ/θ_0 are uniformly bounded above and below for θ within a δ_n-neighborhood of θ_0. Condition A then follows from the fact that

$$\sup_{\{\theta\in\Theta:\,\|\theta-\theta_0\|\le\delta_n\}} n^{-1/2}\max\big(\nu_n((\theta-\theta_0)/\theta_0),\,\nu_n(u^*(\theta-\theta_0)/(\theta_0\theta))\big) = O_p(n^{-2p/(2p+1)})$$

with $2p/(2p+1) \ge d > 1/2$. To verify Condition B, note that

$$\begin{aligned} K(\theta_0,\theta) &= -2\,\mathrm{E}_0\log(1+(\theta-\theta_0)/\theta_0) \\ &= -2\int(\theta-\theta_0)(x)\theta_0(x)\,dx + \|\theta-\theta_0\|^2 + r^{(2)}[\theta-\theta_0] \\ &= \tfrac{1}{2}\|\theta-\theta_0\|^2 + r^{(2)}[\theta-\theta_0], \end{aligned}$$

where $r^{(2)}[\theta-\theta_0] = \mathrm{E}_0(\sum_{j=1}^6(-1)^j((\theta-\theta_0)/\theta_0)^j/j + r^{(3)}[\theta-\theta_0])$ is the remainder in the above expansion. Condition B then follows from the fact that, for any $\theta_i \in \{\theta\in\Theta\colon \|\theta-\theta_0\| \le \delta_n\}$ and some constant $c>0$,

$$\begin{aligned} &|r^{(2)}[\theta^*(\theta,\varepsilon_n)-\theta_0] - r^{(2)}[\theta-\theta_0]| \\ &\quad\le c\Big[\delta_n \sup_{\{\theta\in\Theta:\,\|\theta-\theta_0\|\le\delta_n\}} |P\theta^*(\theta,\varepsilon_n)-\theta| + \sup_{\{\theta\in\Theta:\,\|\theta-\theta_0\|\le\delta_n\}} 2|r^{(3)}[\theta-\theta_0]|\Big] \\ &\quad\le c\varepsilon_n\delta_n^2 + c\delta_n^{2+10p/(2p+1)} \\ &\quad= O(\varepsilon_n^2). \end{aligned}$$

In the above calculations, Proposition 6 has been used for bounding $r^{(3)}$. For Condition C, note that, for any $\theta \in \{\theta\in\Theta_n\colon \|\theta-\theta_0\| \le \delta_n\}$, $\|P(\theta^*(\theta,\varepsilon_n)) - \theta^*(\theta,\varepsilon_n)\| = O(\varepsilon_n\delta_n) = O(\delta_n^{-1}\varepsilon_n^2)$. The first statement in Condition C is satisfied. Using the expression for $P(\theta^*(\theta,\varepsilon_n))$ and Chebyshev's inequality, we have

$$\begin{aligned} &\sup_{\{\theta\in\Theta_n:\,\|\theta-\theta_0\|\le\delta_n\}} n^{-1/2}\nu_n(l'_{\theta_0}[\theta^*(\theta,\varepsilon_n) - P(\theta^*(\theta,\varepsilon_n)), Y]) \\ &\quad\le C\varepsilon_n\delta_n \sup_{\{\theta\in\Theta_n:\,\|\theta-\theta_0\|\le\delta_n\}} n^{-1/2}\nu_n(\theta^*(\theta,\varepsilon_n)) \\ &\quad= O_p(\varepsilon_n^2). \end{aligned}$$

Condition C is therefore fulfilled. Condition D can be verified similarly.

As for the LAN specified in Section 5, we have

$$\frac{dP_{m(\theta_0+A_nh)}}{dP_{\theta_0}}(Y_1,\ldots,Y_n)=\exp\biggl(\sum_{i=1}^n 2\log\biggl[1+\frac{h(X_i)}{2n^{1/2}\theta_0(X_i)}+O\biggl(\frac{\|h\|^2}{n}\biggr)\biggr]\biggr)$$
$$=\exp\biggl(\Sigma(h)-\frac{1}{2}\|h\|^2+R_n(\theta_0,h)\biggr).$$

Then the LAN holds with $\Sigma(h)=n^{-1/2}\sum_{i=1}^n h(X_i)/\theta_0(X_i)$, $A_n=(2n^{1/2})^{-1}$ and $R_n(\theta_0,h)\to_{\mathrm{p}_{\theta_0}}0$.

We conclude by Theorems 5 and 6 that $f(\hat\theta_n)$ is asymptotically efficient for estimating $f(\theta)$ with variance $\mathrm{Var}_0(l'_{\theta_0}[\theta-\theta_0,Y])=4\,\mathrm{Var}_0(\log\theta_0)$ for $p>1/2$.

(b) *Functions with infinite amount of smoothness.* We now calculate the convergence rate of the MLE. It follows from Section 7 of Kolmogorov and Tihomirov (1959) that

$$H(u,\Theta,\|\cdot\|_{\sup})\le\frac{2^{s+1}}{(s+1)!(\log e)^s h^s}\biggl(\log\frac{1}{u}\biggr)^{s+1}+O\biggl(\biggl(\log\frac{1}{u}\biggr)^{s+1}\log\log\frac{1}{u}\biggr)$$

for all $u>0$. Applying Theorem 2 of Wong and Shen (1995), we obtain that the convergence rate of the standard MLE is $O_p(n^{-1/2}(\log n)^{(s+1)/2})$. The extra $\log n$ factor may be eliminated or the power of the $\log n$ term may be reduced if the local metric entropy is calculated [Wong and Shen (1995), Theorem 2], but the above rate is enough for the application in this case. Conditions A-D and the LAN can be verified by applying arguments similar to those above and by the Sobolev interpolation inequality [Zeidler (1990), Example 21.66]. By Theorems 5 and 6, we conclude that $f(\hat\theta_n)$ is asymptotically efficient for estimating $f(\theta)$ with variance $4\,\mathrm{Var}_0(\log\theta_0)$.

9. Technical proofs.

PROOF OF THEOREM 1. The key idea is to use a linear approximation of $n^{-1}\sum_{i=1}^n l'_{\theta_0}[\theta-\theta_0]$ to approximate $L_n(\theta)-L_n(\theta_0)$ characterized by stochastic equicontinuity. Such an approximation is crucial since a poor approximation in a large parameter space cannot yield an asymptotic distribution with the rate of $n^{-1/2}$. Here $f(\hat\theta_n)-f(\theta_0)$ is approximated by $\langle\hat\theta_n-\theta_0,v^*\rangle$, which builds a bridge between $f(\hat\theta_n)$ and $L'_n(\cdot)$. After careful comparisons between $L_n(\hat\theta_n)$ and $L_n(\theta^*(\hat\theta_n,\varepsilon_n))$, we obtain that $\langle\hat\theta_n-\theta_0,v^*\rangle$ is approximated by $n^{-1}\sum_{i=1}^n l'_{\theta_0}[v^*,Y_i]$ with the desired precision.

The following local linear approximation of the empirical criterion can be established from (4.1). For any $P_n\theta_n\in\{P_n\theta_n\in\Theta_n\colon\|P_n\theta_n-\theta_0\|\le\delta_n\}$,

$$L_n(P_n\theta_n)=L_n(\theta_0)-K(\theta_0,P_n\theta_n)+n^{-1/2}\nu_n(l'_{\theta_0}[P_n\theta_n-\theta_0,Y])$$
$$+n^{-1/2}\nu_n(r[P_n\theta_n-\theta_0,Y]).\tag{9.1}$$

Substituting $P_n\theta_n$ by $\hat{\theta}_n$ in (9.1), we obtain that

$$L_n(\hat{\theta}_n) = L_n(\theta_0) - K(\theta_0, \hat{\theta}_n) + n^{-1/2}\nu_n(l'_{\theta_0}[\hat{\theta}_n - \theta_0, Y]) + n^{-1/2}\nu_n(r[\hat{\theta}_n - \theta_0, Y]). \tag{9.2}$$

Note that $\|\theta^*(\hat{\theta}_n, \varepsilon_n) - \theta_0\| = \|(1-\varepsilon_n)(\hat{\theta}_n - \theta_0) + \varepsilon_n u^*\| \le \delta_n$ with probability tending to 1. Subtracting (9.2) from (9.1) and substituting θ_n by $\theta^*(\hat{\theta}_n, \varepsilon_n)$ in (9.2), we have, by Conditions A and B,

$$\begin{aligned} L_n(\hat{\theta}_n) &= L_n(P_n\theta^*(\hat{\theta}_n, \varepsilon_n)) - [K(\theta_0, \hat{\theta}_n) - K(\theta_0, P_n\theta^*(\hat{\theta}_n, \varepsilon_n))] \\ &\quad + n^{-1/2}\nu_n(l'_{\theta_0}[\hat{\theta}_n - P_n\theta^*(\hat{\theta}_n, \varepsilon_n), Y]) \\ &\quad + n^{-1/2}\nu_n(r[\hat{\theta}_n - P_n\theta^*(\hat{\theta}_n, \varepsilon_n), Y]) \\ &= L_n(P_n\theta^*(\hat{\theta}_n, \varepsilon_n)) - \tfrac{1}{2}[\|\hat{\theta}_n - \theta_0\|^2 - \|P_n\theta^*(\hat{\theta}_n, \varepsilon_n) - \theta_0\|^2] \\ &\quad + n^{-1/2}\nu_n(l'_{\theta_0}[\hat{\theta}_n - P_n\theta^*(\hat{\theta}_n, \varepsilon_n), Y]) \\ &\quad + O_p(\varepsilon_n^2). \end{aligned}$$

By Condition C and (2.1), we get

$$\begin{aligned} -O(\varepsilon_n^2) &\le -\tfrac{1}{2}[\|\hat{\theta}_n - \theta_0\|^2 - \|P_n\theta^*(\hat{\theta}_n, \varepsilon_n) - \theta_0\|^2] \\ &\quad + n^{-1/2}\nu_n(l'_{\theta_0}[\hat{\theta}_n - \theta^*(\hat{\theta}_n, \varepsilon_n), Y]) + O_p(\varepsilon_n^2). \end{aligned} \tag{9.3}$$

By (9.3) and Conditions C and D, we have

$$\begin{aligned} -O(\varepsilon_n^2) &\le -\tfrac{1}{2}[1-(1-\varepsilon_n)^2]\|\hat{\theta}_n - \theta_0\|^2 \\ &\quad + (1-\varepsilon_n)\|\hat{\theta}_n - \theta_0\|\|P_n\theta^*(\hat{\theta}_n, \varepsilon_n) - \theta^*(\hat{\theta}_n, \varepsilon_n)\| \\ &\quad + (1-\varepsilon_n)\langle\hat{\theta}_n - \theta_0, \varepsilon_n u^*\rangle - n^{-1/2}\nu_n(l'_{\theta_0}[\varepsilon_n(u^* - (\hat{\theta}_n - \theta_0)), Y]) \\ &\quad + O_p(\varepsilon_n^2) \\ &\le -\varepsilon_n\|\hat{\theta}_n - \theta_0\|^2 + \|\hat{\theta}_n - \theta_0\|\|P_n\theta^*(\hat{\theta}_n, \varepsilon_n) - \theta^*(\hat{\theta}_n, \varepsilon_n)\| \\ &\quad + (1-\varepsilon_n)\langle\hat{\theta}_n - \theta_0, \varepsilon_n u^*\rangle - n^{-1/2}\nu_n(l'_{\theta_0}[\varepsilon_n u^*, Y]) + O_p(\varepsilon_n^2) \\ &\le (1-\varepsilon_n)\langle\hat{\theta}_n - \theta_0, \varepsilon_n u^*\rangle - n^{-1/2}\nu_n(l'_{\theta_0}[\varepsilon_n u^*, Y]) + O_p(\varepsilon_n^2). \end{aligned}$$

Hence,

$$-(1-\varepsilon_n)\langle\hat{\theta}_n - \theta_0, u^*\rangle + n^{-1/2}\nu_n(l'_{\theta_0}[u^*, Y]) = O(\varepsilon_n) + O_p(\varepsilon_n) = o_p(n^{-1/2}). \tag{9.4}$$

This gives, together with the inequality in (9.4) with u^* being replaced by $-u^*$,

$$|\langle\hat{\theta}_n - \theta_0, u^*\rangle - n^{-1/2}\nu_n(l'_{\theta_0}[u^*, Y])| = o_p(n^{-1/2}),$$

whence $\langle \hat{\theta}_n - \theta_0, v^* \rangle = n^{-1/2} \nu_n(l'_{\theta_0}[v^*, Y]) + o_p(n^{-1/2})$. Hence, by (4.2),

$$\begin{aligned} f(\hat{\theta}_n) - f(\theta_0) &= f'_{\theta_0}[\hat{\theta}_n - \theta_0] + o_p(u_n \|\hat{\theta}_n - \theta_0\|^{\omega}) \\ &= \langle \hat{\theta}_n - \theta_0, v^* \rangle + o_p(n^{-1/2}) \\ &= n^{-1} \sum_{i=1}^{n} l'_{\theta_0}[v^*, Y_i] + o_p(n^{-1/2}). \end{aligned}$$

The result then follows from the classical central limit theorem. □

PROOF OF COROLLARY 1. The result follows from the same argument as in the proof of Theorem 1 with v^* being replaced by s. □

PROOF OF COROLLARY 2. The main body of the proof is as shown in Theorem 1. Due to the possible poor approximations by the sieve, we need to bound the related error terms precisely. Note that

$$P_n \theta^*(\hat{\theta}_n, \varepsilon_n) - \theta^*(\hat{\theta}_n, \varepsilon_n) = \varepsilon_n(\pi_n(u^* + \theta_0) - (u^* + \theta_0)).$$

By (4.3), Conditions B′ and C′ and the Cauchy–Schwarz inequality, we have, after some calculations, that

$$\begin{aligned} &-\tfrac{1}{2}\|\hat{\theta}_n - \theta_0\|^2(1 - |o(h_n)|) + \tfrac{1}{2}\Big\|(P_n\theta^*(\hat{\theta}_n, \varepsilon_n) - \theta^*(\hat{\theta}_n, \varepsilon_n)) \\ &\qquad + (\theta^*(\hat{\theta}_n, \varepsilon_n) - \theta_0)\Big\|^2(1 + |o(h_n)|) \\ &\quad \le -\varepsilon_n(1 - |o(h_n)|)\|\hat{\theta}_n - \theta_0\|^2 \\ &\qquad + \varepsilon_n(1 + |o(h_n)|)\langle \hat{\theta}_n - \theta_0, \pi_n(u^* + \theta_0) - (u^* + \theta_0)\rangle \\ &\qquad + (1 - \varepsilon_n)(1 + |o(h_n)|)\langle \varepsilon_n u^*, \hat{\theta}_n - \theta_0\rangle + O_p(\varepsilon_n^2) + O(\varepsilon_n^2) \\ &\quad \le -\varepsilon_n(1 - |o(h_n)|)[\|\hat{\theta}_n - \pi_n\theta_0\|^2 + \|\pi_n\theta_0 - \theta_0\|^2] \\ &\qquad + \varepsilon_n[\|\pi_n\theta_0 - \theta_0\|^2 + \langle \hat{\theta}_n - \theta_0, \pi_n u^* - u^*\rangle] \\ &\qquad + (1 - \varepsilon_n)(1 + |o(h_n)|)\langle \varepsilon_n u^*, \hat{\theta}_n - \theta_0\rangle + O_p(\varepsilon_n^2) \\ &\quad \le 2\varepsilon_n|o_p(h_n)|\|\pi_n\theta_0 - \theta_0\|^2 + \varepsilon_n(1 + |o_p(h_n)|)\|\hat{\theta}_n - \theta_0\|\|\pi_n u^* - u^*\| \\ &\qquad + (1 - \varepsilon_n)(1 + |o(h_n)|)\langle \varepsilon_n u^*, \hat{\theta}_n - \theta_0\rangle + O_p(\varepsilon_n^2) \\ &\quad \le (1 - \varepsilon_n)(1 + |o(h_n)|)\langle \varepsilon_n u^*, \hat{\theta}_n - \theta_0\rangle + O_p(\varepsilon_n^2). \end{aligned}$$

The result then follows from the same arguments as in (9.3) and (9.4). □

PROOF OF THEOREM 2. The basic idea is the same as that presented in Theorem 1. However, we need to control the penalty $J(\hat{\theta}_n)$. Similarly, by (2.2),

Condition D″ and (ii) of Condition A″, we obtain

$$\begin{aligned}-O(\varepsilon_n^2) &\le \tilde{L}_n(\hat{\theta}_n) - \tilde{L}_n(\theta_0)\\ &\le -\tfrac{1}{2}\|\hat{\theta}_n - \theta_0\|^2 + n^{-1/2}\nu_n(l'_{\theta_0}[\hat{\theta}_n - \theta_0, Y]) + n^{-1/2}\nu_n(r[\hat{\theta}_n - \theta_0, Y])\\ &\quad - \lambda_n(J(\hat{\theta}_n) - J(\theta_0)) + O_p(\varepsilon_n)\\ &\le -\lambda_n(J(\hat{\theta}_n) - J(\theta_0)) + O_p(\varepsilon_n).\end{aligned}$$

Thus, $\lambda_n[J(\hat{\theta}_n) - J(\theta_0)] \le O_p(\varepsilon_n)$. By Condition C″ and the fact that $J(u^*) < \infty$, we have

$$\begin{aligned}\lambda_n(J(\theta^{**}(\hat{\theta}_n, \varepsilon_n)) - J(\hat{\theta}_n)) &\le c\lambda_n J(\varepsilon_n[-\hat{\theta}_n + \theta_0 + u^*])\\ &\le c\lambda_n\varepsilon_n(J(\hat{\theta}_n - \theta_0) + J(u^*))\\ &= O_p(\varepsilon_n^2),\end{aligned}$$

for some constant $c > 0$. Comparing $L_n(\hat{\theta}_n)$ with $L_n(\theta^{**}(\hat{\theta}_n, \varepsilon_n))$ as in (9.3)–(9.4), we obtain the desired result. □

PROOF OF COROLLARY 3. Use the same arguments as in Corollary 1. □

PROOF OF THEOREM 3. Applying testing arguments similar to those in Bahadur (1964) and Wong (1992) on H_0: $P_{\theta_n, \rho}$ versus H_A: P_{θ_0}, we obtain the desired result. See the proof of Proposition 4 in Wong (1992). □

PROOF OF THEOREM 4. Note that

$$\begin{aligned}(f(\hat{\theta}_n) - f(\theta_n)) &= (f(\hat{\theta}_n) - f(\theta_0)) - (f(\theta_n) - f(\theta_0))\\ &= \langle\hat{\theta}_n - \theta_0, v^*\rangle - \langle n^{-1/2}\rho h, v^*\rangle + o_p(n^{-1/2})\\ &= n^{-1}\sum_{i=1}^{n} l'_{\theta_0}[v^*, Y] - \langle n^{-1/2}\rho h, v^*\rangle + o_p(1).\end{aligned}$$

The desired result follows from the LAN condition and Le Cam's third lemma. □

PROOF OF THEOREM 5. The proof is straightforward and thus omitted. □

PROOF OF PROPOSITION 3. First consider the standard MLE. By Lemma 4 of Shen and Wong (1994), we have $\sup_{\eta\in B^\alpha} n^{-1/2}\nu_n(\eta - \eta_0) = O_p(n^{-\alpha})$. We then obtain after some calculations that $\max(|\hat{\beta}_n - \beta_0|, (\mathrm{E}_0(\hat{\eta}_n - \eta_0)^2)^{1/2}) \le O((2/\sigma)\|\hat{\theta}_n - \theta_0\|) = O_p(n^{-\alpha/2})$. Hence, we restrict our attention to an $n^{-\alpha/2}$-neighborhood of θ_0 that includes the standard MLE.

Let $S = \{X_{j_1}, \ldots, X_{j_{N_n}}\}$ be the maximal subset of $\{X_1, \ldots, X_n\}$ such that S satisfies (1) $\{(X_{j_i} - 0.2/(n+1), X_{j_i} + 0.2/(n+1))\}_{i=1}^{N_n}$ are disjoint and

contained in $[1/8, 3/8]$ and (2) for any $X_i \in S$ and $X_j \notin S$, $|X_i - X_j| \geq 0.2/(n+1)$. Let $\tilde{\eta}(x) = \eta_0 + \sum_{i=1}^{N_n}[\operatorname{sgn}(e_{j_i})B_h(X_{j_i} + h/2 - x) + G_h(X_{j_i} + h/2 - x)]$ and $\tilde{\theta} = (\tilde{\beta}, \tilde{\eta})$, where $\operatorname{sgn}(\cdot)$ is the sign function, $\tilde{\beta} = \beta_0 - (\mathrm{E}_0 Z_1/\mathrm{E}_0 Z_1^2)^{1/2}\,\mathrm{E}_0(\tilde{\eta} - \eta_0)]^2$. By Lemma 2 of Birgé and Massart (1993), $N_n \geq n/16$ with probability 1. Using the property of S, we obtain that $B_h(X_{j_i} + h/2 - X_j) = 0$ for $j \neq j_i$ and $B_h(h/2) = (1/2^\alpha)h^\alpha$. Note that $0 \leq G(x) \leq ([\mathrm{E}\, e_1 \operatorname{sgn}(e_1)]/2^{\alpha+2})^{1/2}h^{\alpha/2}$. For any $0.2/(n+1) \leq h \leq \tilde{c}/(n+1)$, $0.2 \leq \tilde{c} \leq 1$,

$$\begin{aligned}
L_n(\tilde{\theta}) - L_n(\theta_0) &= \frac{1}{2\sigma^2 n}\sum_{j=1}^{n}[2e_j(\tilde{\eta} - \eta_0)(X_j) - (\tilde{\eta} - \eta_0)^2(X_j)] \\
&\quad + (\mathrm{E}_0 Z_1)^2[\mathrm{E}_0(\tilde{\eta} - \eta_0)]^2 + o_p(n^{-\alpha}) \\
&\geq \frac{1}{\sigma^2 n}\sum_{j=1}^{n} e_j\Bigg[\sum_{i=1}^{N_n}\operatorname{sgn}(e_{j_i})B_h\Big(X_{j_i} + \frac{h}{2} - X_j\Big) \\
&\qquad\qquad - 2\Big(\sum_{i=1}^{N_n} G_h\Big(X_{j_i} + \frac{h}{2} - X_j\Big)\Big)^2\Bigg] + o_p(n^{-\alpha}) \\
&\geq \frac{1}{\sigma^2 n}\sum_{j\in S} e_j\Big[\operatorname{sgn}(e_j)\frac{1}{2^\alpha}h^\alpha - \frac{[\mathrm{E}e_1 \operatorname{sgn}(e_1)]}{2^{\alpha+1}}h^\alpha\Big] + o_p(n^{-\alpha}) \\
&\geq \frac{[\mathrm{E}e_1 \operatorname{sgn}(e_1)]}{2^{\alpha+1}\sigma^2}h^\alpha = c_2(n+1)^{-\alpha},
\end{aligned}$$

with a nonzero probability, where $c_2 > 0$ is a constant. On the other hand, using an empirical process inequality [see, e.g., Shen and Wong (1994), Lemma 4], we have, for any $\theta \in \Theta$,

$$L_n(\theta) - L_n(\theta_0) \leq c_3 h^\alpha + o_p(n^{-\alpha}),$$

which implies that an approximate maximizer $\hat{h}_n = c_4 n^{-1}$ for some constant $c_4 > 0$, where $c_3 > c_2$ is a constant. Thus, $\mathrm{E}_0(\hat{\eta}_n - \eta_0) = \hat{h}_n^{\alpha/2}(1 + o(1))$ with a nonzero probability.

For any θ in the neighborhood,

$$\begin{aligned}
L_n(\theta) - L_n(\theta_0) &= \frac{1}{2\sigma^2 n}\sum_{j=1}^{n}\Big[2e_j(\lambda(Z_j, X_j) - \lambda_0(Z_j, X_j)) \\
&\qquad\qquad - (\lambda(Z_j, X_j) - \lambda_0(Z_j, X_j))^2\Big] \\
&= \frac{1}{2\sigma^2 n}\sum_{j=1}^{n}[2e_j(\eta - \eta_0)(X_j) - (\eta - \eta_0)^2(X_j)] \\
&\quad + (\mathrm{E}_0 Z_1)^2[\mathrm{E}_0(\eta - \eta_0)]^2 \\
&\quad - \mathrm{E}_0 Z_1^2\Big[(\beta - \beta_0) + \Big(\frac{\mathrm{E}_0 Z_1}{\mathrm{E}_0 Z_1^2}\Big)^{1/2}\mathrm{E}_0(\eta - \eta_0)\Big]^2 + o_p(n^{-\alpha})
\end{aligned}$$

$$= -\frac{1}{2\sigma^2 n}\sum_{j=1}^{n}[e_j - (\eta-\eta_0)(X_j)]^2 + (\mathrm{E}_0 Z_1)^2[\mathrm{E}_0(\eta-\eta_0)]^2$$

$$- (\mathrm{E}_0 Z_1^2)\left[(\beta-\beta_0) + \left(\frac{\mathrm{E}_0 Z_1}{\mathrm{E}_0 Z_1^2}\right)^{1/2}\mathrm{E}_0(\eta-\eta_0)\right]^2 + o_p(n^{-\alpha}).$$

In order to maximize $L_n(\theta) - L_n(\theta_0)$ in the above expression, it is necessary to minimize the second negative term. Hence,

$$|(\hat{\beta}_n - \beta_0) + (\mathrm{E}_0 Z_1/\mathrm{E}_0 Z_1^2)^{1/2}\,\mathrm{E}_0(\hat{\eta}_n - \eta_0)| = o_p(n^{-\alpha/2}).$$

The result in Proposition 3 then follows.

Now consider the penalized MLE. The convergence rate for the penalized MLE is $O_p(n^{-\alpha/2})$ when λ_n is chosen to be of order $n^{-\alpha}$ [Shen (1997a), Theorem 3]. Furthermore, by Theorem 4 of Shen (1997a), we know that, for any small $\delta > 0$, $J(\hat{\eta}_n) \le (1+\delta)J(\eta_0)$ with probability tending to 1. We therefore restrict our attention to the set $\{\eta \in B^\alpha\colon J(\eta) \le (1+\delta)J(\eta_0)\}$. Note that

$$\tilde{L}_n(\theta) - \tilde{L}_n(\theta_0) = L_n(\theta) - L_n(\theta_0) - \lambda_n(J(\eta) - J(\eta_0)).$$

The conclusion for the standard MLE continues to hold for the penalized MLE since the contribution of $\lambda_n(J(\eta) - J(\eta_0))$ $[-J(\eta_0)n^{-\alpha} \le \lambda_n(J(\eta) - J(\eta_0)) \le \delta J(\eta_0)\lambda_n = \delta J(\eta_0)n^{-\alpha}]$ is negligible. □

PROPOSITION 6. *Let* $f \in C^{m+\gamma}[a,b] = \{f \in C^m[a,b]\colon f(a) = f(b) = 0, \|f^{(j)}(x)\|_{\sup} \le L_j, \|f^{(m)}\|_H \le L_{m+1}\}$ *for* $j = 0,\ldots,m$, *where the Hölder norm is defined as*

$$\|f^{(m)}\|_H = \sup_{x\in[a,b]} \frac{|f^{(m)}(x) - f^{(m)}(y)|}{|x-y|^\gamma}.$$

Then

$$\|f\|_{\sup} \le \|f\|_2^a L^{1-a},$$

where $a = 2(m+\gamma)/(2(m+\gamma)+1)$ *and* L *is a positive constant independent of* f.

PROOF. The above result can be obtained by applying an argument similar to that in Theorem 1 of Gabushin (1967). The detailed proof for the case of $m = 0$ is given in Lemma 7 of Shen and Wong (1994). □

Acknowledgments. The author would like to thank the Associate Editors and an anonymous referee for helpful comments and suggestions. The author would also like to thank Professors John Rice and Andrew Barron for expediting the review process.

REFERENCES

BAHADUR, R. R. (1964). On Fisher's bound for asymptotic variances. *Ann. Math. Statist.* **35** 1545–1552.

BAHADUR, R. R. (1967). Rate of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303–324.

BEGUN, J., HALL, W., HUANG, W. and WELLNER, J. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452.

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

BICKEL, P. J., KLASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1994). *Efficient and Adaptive Inference in Semi-parametric Models.* Johns Hopkins Univ.

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

BIRGÉ, L. and MASSART, P. (1994). Minimum contrast estimators on sieves. Unpublished manuscript.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

DEVORE, R. and LORENTZ, G. (1991). *Constructive Approximation*. Springer, New York.

GABUSHIN, V. N. (1967). Inequalities for norms of functions and their derivatives in the L_p metric. *Mat. Zametki* **1** 291–298.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

HÁJEK, J. (1970). A characterisation of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation*. Springer, New York.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1991). Asymptotically normal families of distributions and efficient estimation. *Ann. Statist.* **19** 1681–1724.

KOENKER, R. and BASSETT, F. (1978). Regression quantiles. *Econometrica* **46** 33–50.

KOLMOGOROV, A. N. and TIHOMIROV, V. M. (1961). ε-entropy and ε-capacity of sets in function spaces. *Amer. Math. Soc. Transl.* **2** 227–304.

LE CAM, L. (1960). Local asymptotically normal families of distributions. *Univ. California Publ. Statist.* **3** 37–98.

LEVIT, B. (1974). On optimality of some statistical estimates. In *Proceedings of the Prague Symposium on Asymptotic Statistics* (J. Hajek, ed.) **2** 215–238. Univ. Karlova, Prague.

LEVIT, B. (1978). Infinite dimensional informational inequalities. *Theory Probab. Appl.* **23** 371–377.

LINDSAY, B. G. (1980). Nuisance parameters, mixture models and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London Ser. A* **296** 639–665.

LORENTZ, G. (1966). *Approximation of Functions*. Holt, Reinehart and Winston, New York.

OSSIANDER, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15** 897–919.

PARZEN, M. and HARRINGTON, D. (1993). Proportional odds regression with right-censored data using adaptive integrated splines. Technical report, Dept. Biostatistics, Harvard Univ.

PFANZAGL, J. (1982). *Contribution to a General Asymptotic Statistical Theory*. Springer, New York.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

RITOV, Y. and BICKEL, P. J. (1990). Achieving information bounds in non and semiparametric models. *Ann. Statist.* **18** 925–938.

SCHOENBERG, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. U.S.A.* **52** 947–950.

SEVERINI, T. A. and WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20** 1768–1862.

SHEN, X. (1997a). On the method of penalization. *Statist. Sinica*. To appear.

SHEN, X. (1997b). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika*. To appear.

SHEN, X. and WONG, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615.

STONE, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

TIKHONOV, A. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet. Math. Dokl.* **5** 1035–1038.

TRIEBEL, H. (1983). *Theory of Function Spaces*. Birkhäuser, Boston.

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

WAHBA, G. (1990). *Spline Models for Observational Data*. IMS, Hayward, CA.

WHITTAKER, E. (1923). On new method of graduation. *Proc. Edinburgh Math. Soc.* **2** 41.

WONG, W. H. (1992). On asymptotic efficiency in estimation theory. *Statist. Sinica* **2** 47–68.

WONG, W. H. and SEVERINI, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter space. *Ann. Statist.* **19** 603–632.

WONG, W. H. and SHEN, X. (1995). A probability inequality for the likelihood surface and convergence rate of the maximum likelihood estimate. *Ann. Statist.* **23** 339–362.

ZEIDLER, E. (1990). *Nonlinear Functional Analysis and Its Applications II/A*. Springer, New York.

DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
1958 NEIL AVE.
COLUMBUS, OHIO 43210-1247
E-MAIL: xshen@mle.mps.ohio-state.edu