The Annals of Statistics
1990, Vol. 18, No. 4, 1676–1695

ASYMPTOTIC ANALYSIS OF PENALIZED LIKELIHOOD AND RELATED ESTIMATORS

BY DENNIS D. COX[1] AND FINBARR O'SULLIVAN[2]

University of Illinois and University of Washington

A general approach to the first order asymptotic analysis of penalized likelihood and related estimators is described. The method gives expansions for the systematic and random error. Asymptotic convergence rates in a family of spectral norms are obtained. The theory applies to a broad range of function estimation problems including nonparametric density, hazard and generalized regression curve estimation. Some examples are provided.

1. Introduction. Regularization is an old technique for obtaining well behaved solutions to overparameterized estimation problems. Several historical instances of the method can be identified including one by Whittaker (1923) who used the method to smooth time series data. Tikhonov (1963) was the first to systematically study regularization for solving a broad range of *inverse problems* in applied mathematics and the introduction of the term *regularization* is generally credited to him.

Let θ be the parameter of interest. The method of regularization has two components: a data fit functional l_n which measures how well θ predicts the observed set of n-dimensional data and a penalty functional J which assesses the physical plausibility of θ. Smaller values of $J(\theta)$ generally correspond to more desirable values for θ. The method of regularization chooses a θ which minimizes

$$l_{n\lambda}(\theta) = l_n(\theta|\text{data}) + \lambda J(\theta), \qquad \lambda > 0. \tag{1.1}$$

λ is the regularization parameter. Larger values of λ produce more regular estimators. The above formulation was initially proposed in the statistics literature by Good and Gaskins (1971), who used the term penalized likelihood. In their context, l_n corresponded to the negative log-likelihood for the data given θ and J was referred to as a penalty, roughness or flamboyancy functional. Penalized likelihood is familiar to Bayesians as a maximum a posteriori (MAP) procedure [Leonard (1978)]. It is clear that there are close connections with the various methods of sieves as well, see Grenander (1981).

In this paper we use the term penalized likelihood to refer to the functional $l_{n\lambda}$ even though the data fit functional l_n need not be a negative log-likelihood. We shall assume that the data fit functional $l_n(\theta)$ approaches a limiting functional $l(\theta)$ as $n \to \infty$. It is easy to identify $l(\theta)$ in most concrete examples;

Received December 1988; revised July 1989.
[1]Research partially supported by NSF Grant MCS-820-2560.
[2]Research supported by NSF Grant MCS-840-3239 and by the Department of Energy Grant DE-FG06-85ER25006.
AMS 1980 *subject classifications*. Primary, 62G05; secondary, 62J05, 41A35, 41A25, 47A53, 45L10, 45M05.
Key words and phrases. Linearization, spectral analysis, rates of convergence.

see Section 1.1 below. The limiting functional is used to identify a target parameter θ_0; again see Section 1.1. Penalized likelihood procedures have been used in practice for density and hazard estimation, generalized nonparametric regression and more general nonlinear inverse problems. The purpose of this paper is to provide a framework for studying the error characteristics of such estimation schemes. Although a fair bit of functional analysis is required, the main ideas are rather simple. An informal description of the error analysis is given in Section 1.2.

1.1. *Examples.* We begin by giving some examples of the results obtained from the analysis to follow. Applications to a general family of nonparametric regression estimators are detailed in Cox and O'Sullivan (1989). For each example discussed here, the parameter of interest is a univariate real-valued function defined on a bounded interval which we take to be $[0, 1]$. The familiar smoothing spline type penalty

$$J(\theta) = \int_0^1 \left[\theta^{(m)}(t)\right]^2 dt$$

is used. We obtain, under regularity conditions, the usual upper bounds on the integrated squared error convergence rate.

EXAMPLE 1. Log-density estimation. $X_1, X_2, \ldots, X_n$ is a random sample from a density f_0: $[0, 1] \to R$. Suppose f_0 is bounded away from zero and infinity and let $\theta_0 = \log f_0$. Silverman (1982) introduced a penalized likelihood procedure for estimating θ_0. In this scheme $l_n(\theta)$ is

$$l_n(\theta) = \int_0^1 e^{\theta(t)}\, dt - \frac{1}{n}\sum_{i=1}^{n} \theta(X_i). \tag{1.2}$$

The second term is a negative log-likelihood for the data. The integral term is included to guarantee the unitary constraint for a probability density. The limiting form of $l(\theta)$ is

$$l(\theta) = \int_0^1 e^{\theta(t)}\, dt - \int_0^1 \theta(t) e^{\theta_0(t)}\, dt. \tag{1.3}$$

It is readily verified that θ_0 is the unique minimizer of l.

EXAMPLE 2. Log-hazard estimation. Let $(X_1, \delta_1), (X_2, \delta_2), \ldots, (X_n, \delta_n)$ be a random sample in which $X_i = \min(Y_i, C_i)$ and $\delta_i = I_{C_i}(Y_i)$ (I_t is the characteristic function of $[0, t]$). This corresponds to a sample of censored failure time data common in survival analysis. We assume that the censoring time C_i and failure time Y_i are independent. Let $\lambda_0(t)$ be the hazard function for the failure time distribution and suppose λ_0: $[0, 1] \to R$ is bounded away from zero and infinity. Put $\theta_0 = \log \lambda_0$. Following Anderson and Senthilselvan (1980), a penalized likelihood estimator of θ_0 may be developed; see O'Sullivan

(1988). Here l_n is

$$l_n(\theta) = \int e^{\theta(t)} S_n(t)\, dt - \frac{1}{n}\sum_{i=1}^{n} \delta_i \theta(X_i), \tag{1.4}$$

where S_n is the empirical survival function of the sample $X_1, X_2, \ldots, X_n$. If the limiting survival function is denoted S, then

$$l(\theta) = \int_0^1 e^{\theta(t)} S(t)\, dt - \int_0^1 \theta(t) e^{\theta_0(t)} S(t)\, dt. \tag{1.5}$$

From this it can be verified that θ_0 is the unique minimizer of l.

EXAMPLE 3. Nonparametric logistic regression. (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ is a random sample in which $Y|X = t$ is Bernoulli with success probability $p(t)$. X has density f: $[0,1] \to R$, which we assume is bounded away from the zero and infinity. Suppose $p(t)$ is bounded away from zero and one and let $\theta_0(t) = \text{logit}(p(t)) = \log(p(t)/(1 - p(t)))$. For penalized likelihood estimation of θ_0 [see O'Sullivan, Yandell and Raynor (1986)], we let

$$l_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \{\log(1 + e^{\theta(X_i)}) - Y_i \theta(X_i)\}. \tag{1.6}$$

It follows that

$$l(\theta) = \int_0^1 \{\log(1 + e^{\theta(t)}) - p(\theta_0(t))\theta(t)\} f(t)\, dt. \tag{1.7}$$

Again θ_0 may be shown to be the unique minimizer of l.

The parameter space for each of these examples will be a Sobolev space given by

$$W_2^m[0,1] = \{\theta: [0,1] \to R | \theta, \theta^{(1)}, \ldots, \theta^{(m-1)} \text{ are absolutely continuous and } \theta^{(m)} \in L_2[0,1]\}.$$

This is a natural choice given the penalty. There are many possible inner products on W_2^m under which it is a Hilbert space, for instance,

$$\langle \theta, \zeta \rangle_{W_2^m} = \langle \theta, \zeta \rangle_{L_2} + \langle \theta^{(m)}, \zeta^{(m)} \rangle_{L_2}. \tag{1.8}$$

The Sobolev spaces will be useful not only as a parameter spaces, but we will use their norms to measure the loss. Sobolev spaces with noninteger order m are also used, see Adams (1975) for definitions. As an illustration of the theory we obtain Theorem 1.1, which applies to the previous examples.

THEOREM 1.1. *Suppose* $m \geq 2$ *and* $\theta_0 \in W_2^{mp}[0,1]$, *where* $3/(2m) < p \leq 1$. *If* λ_n *is a sequence such that* $\lambda_n \to 0$ *and for some* $\alpha \in (1/(2m), (p - 1/(2m))/2]$, $n^{-1}\lambda_n^{-2(\alpha + 1/(2m))} \to 0$, *then for* $0 \leq b \leq \alpha$,

$$\lim_{M \to \infty} \liminf_{n \to \infty} P\Big[\text{a unique minimizer } \theta_{n\lambda_n} \text{ of } l_{n\lambda_n}(\theta) \text{ exists and satisfies } \|\theta_{n\lambda_n} - \theta_0\|^2_{W_2^{mb}} \leq M\big(\lambda_n^{p-b} + n^{-1}\lambda_n^{-(b+1/(2m))}\big)\Big] = 1,$$

in any of the examples (1, 2 *or* 3).

The proof is given in Section 4.

1.2. *Informal description of the error analysis.* Let the parameter space be denoted Θ and for the purposes of this description let Θ be a subset of R^p. The symbol D denotes differentiation with respect to the parameter θ. The theory developed in this paper relates to the large sample behavior of roots of the penalized likelihood equations. That is, we look at the *score vector* $Z_{n\lambda}$ given by

$$Z_{n\lambda}(\theta) = Dl_{n\lambda}(\theta) \tag{1.9}$$

and discuss the properties of roots of the equations $Z_{n\lambda} = 0$ as the sample size $n \to \infty$. The limiting score vector is defined as

$$Z_\lambda(\theta) = Dl(\theta) + \lambda DJ(\theta), \tag{1.10}$$

where $l(\theta)$ is the limiting version of l_n.

One step linearizations. The roots of Z_λ and $Z_{n\lambda}$, which will be shown to exist in Section 3, may be approximated by linearized forms using a first order Taylor series expansion of Z_λ and $Z_{n\lambda}$. Thus the method of analysis is similar in spirit to Cramer's (1946) analysis of maximum likelihood estimators. Let $G_\lambda(\theta)$ be the Hessian of the limiting penalized likelihood. We assume that the penalty functional J has the form $J(\theta) = \frac{1}{2}\theta' W\theta$, where W is symmetric positive semidefinite. Thus

$$G_\lambda(\theta) = D^2 l(\theta) + \lambda W \equiv U(\theta) + \lambda W. \tag{1.11}$$

In general $U(\theta)$ and W are linear operators but for the purposes of this discussion, Θ is p-dimensional, so they are $p \times p$ matrices. The true parameter θ_0 is assumed to be a locally unique root of $Dl(\theta_0) = 0$. Linearizing the *continuous* score Z_λ about the true parameter value θ_0 and setting the result equal to zero gives the approximate root

$$\bar{\theta}_\lambda = \theta_0 - G_\lambda(\theta_0)^{-1} Z_\lambda(\theta_0). \tag{1.12}$$

We call this the *continuous* linearization and we use it to show that for all λ sufficiently small there is a locally unique root θ_λ of Z_λ in a neighborhood of θ_0 and $\theta_\lambda \approx \bar{\theta}_\lambda$. There is a corresponding *discrete* linearization which arises by first order linearization of the discrete score vector $Z_{n\lambda}$ about θ_λ, but replacing $DZ_{n\lambda}$ by DZ_λ. The discrete linearization is the root $\bar{\theta}_{n\lambda}$ of the linearized equation and is given by

$$\bar{\theta}_{n\lambda} = \theta_\lambda - G_\lambda(\theta_\lambda)^{-1} Z_{n\lambda}(\theta_\lambda). \tag{1.13}$$

It will be shown that if λ_n is a sequence which does not tend to zero too fast (see Theorems 3.2 and 4.2), then (with $\lambda = \lambda_n$) for all n sufficiently large with probability approaching unity there is a unique root $\theta_{n\lambda}$ of $Z_{n\lambda}$ in a neighborhood of θ_λ satisfying $\theta_{n\lambda} \approx \bar{\theta}_{n\lambda}$.

Taken together, these linearizations yield an expansion for the estimation error as

$$\theta_{n\lambda} - \theta_0 = (\theta_\lambda - \theta_0) + (\theta_{n\lambda} - \theta_\lambda) \approx (\bar{\theta}_\lambda - \theta_0) + (\bar{\theta}_{n\lambda} - \theta_\lambda). \tag{1.14}$$

The continuous linearization provides information on the asymptotic bias (first term) of the estimator while the discrete linearization gives information on its asymptotic sampling variability (second term). Explicit error estimates are derived using spectral analysis which is described in Section 2. The entire analysis draws heavily on the techniques developed in Cox (1988).

2. Theoretical framework. This section provides some of the formal framework needed for the analysis to follow. For later reference we use the notation A.1–A.4 to label the assumptions. We could have stayed in a framework of function spaces but the arguments seem easier when given in a more general setting.

2.1. *Parameter space and penalty functional.*

ASSUMPTION A.1. Θ is a real Hilbert space with norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$. The penalty functional J is a quadratic form $J(\theta) = \frac{1}{2}\langle\theta, W\theta\rangle$, where W is a nonnegative definite linear operator on Θ.

The latter requirement means that W is self adjoint on Θ ($W = W^*$, where W^* is defined through $\langle\theta, W\zeta\rangle = \langle W^*\theta, \zeta\rangle$ for all $\theta, \zeta \in \Theta$), and that $\langle\theta, W\zeta\rangle \geq 0$ (compare with matrices). For each of the examples in Section 1, Θ is the Sobolev space $W_2^m[0,1]$. In general, we will not require that the target parameter θ_0 be an element of Θ. For instance, it may happen that θ_0 is in $W_2^{pm}[0,1]$ for some $p < 1$.

2.2. *Spectral decomposition.* To study convergence it is necessary to define some appropriate norms. Let $\mathbf{B}(\Theta,\Theta)$ be the collection of bounded linear operators on Θ. $\mathbf{B}(\Theta,\Theta)$ is equipped with the usual operator norm. We introduce an auxiliary operator $U \in \mathbf{B}(\Theta,\Theta)$, which is equivalent in some sense to $D^2 l(\theta)$ (see Assumption A.4). Unfortunately, we need U for technical purposes (defining norms and spaces) before being able to rigorously indicate where it comes from [the derivative operator $D^2 l(\theta)$ will be defined on spaces derived from U].

ASSUMPTION A.2. U is a compact positive definite operator and there exist strictly positive constants m and M such that for all $\zeta \in \Theta$,

$$m\|\zeta\|^2 \leq \langle\zeta, U\zeta\rangle + \langle\zeta, W\zeta\rangle \leq M\|\zeta\|^2. \tag{2.1}$$

There is a convenient parameterized family of norms which are defined in terms of U and W. From Section 3.3 of Weinberger (1974), (A.2) and the construction in Section 2 of Cox (1988) there is a sequence $\{\phi_\nu\colon \nu = 1, 2, \ldots\}$ of eigenfunctions and $\{\gamma_\nu\colon \nu = 1, 2, \ldots\}$ of eigenvalues which satisfy

$$\langle\phi_\mu, U\phi_\nu\rangle = \delta_{\nu\mu} \quad \text{and} \quad \langle\phi_\mu, W\phi_\nu\rangle = \gamma_\nu\delta_{\nu\mu}, \tag{2.2}$$

for all pairs ν, μ of positive integers, where $\delta_{\nu\mu}$ is Kronecker's delta.

For $b \geq 0$, let

$$\|\theta\|_b = \left\{ \sum_{\nu=1}^{\infty} (1 + \gamma_\nu^b) \langle \theta, U\phi_\nu \rangle^2 \right\}^{1/2} \tag{2.3}$$

and let Θ_b denote the normed linear space obtained by completing $\{\theta \in \Theta: \|\theta\|_b < \infty\}$ in $\|\cdot\|_b$ norm. Θ_b is a Hilbert space with inner product

$$\langle \theta, \zeta \rangle_b = \sum_{\nu=1}^{\infty} (1 + \gamma_\nu^b) \langle \theta, U\phi_\nu \rangle \langle \zeta, U\phi_\nu \rangle. \tag{2.4}$$

For $0 \leq b \leq 1$, Θ_b can be obtained by applying the K-method of interpolation [Triebel (1978)] to Θ_0 and $\Theta = \Theta_1$. Here, the notation $=$ means the Banach spaces are equal as sets and have equivalent norms. The equivalence $\Theta_1 = \Theta$ follows from Assumption A.2. Further, if $b \leq a$, then $\Theta_a \subset \Theta_b$ and has a stronger norm, i.e., there is a constant C_{ab} such that

$$\|\theta\|_b \leq C_{ab} \|\theta\|_a, \tag{2.5}$$

for all $\theta \in \Theta_a$. Without loss of generality (and in the interests of some simplification) we will assume that $C_{ab} \leq 1$. Formally, $C_{ab} \leq 1$ whenever $\gamma_\nu \geq 1$ which can be accomplished by trivial rescaling of W. The following result is needed to show that our linearizations are well-defined.

LEMMA 2.1. *For $0 \leq b \leq 1$, U extends to an element of $\mathbf{B}(\Theta_b, \Theta_2)$ and W extends to an element of $\mathbf{B}(\Theta_b, \Theta_b)$.*

PROOF. We first note that if $\theta \in \Theta = \Theta_1$, then $\theta \in \Theta_{1+c}$ $(c \in R)$ if and only if

$$\sup_{\zeta \in \Theta, \|\zeta\|_{1-c} = 1} \langle \zeta, \theta \rangle < \infty.$$

Further, the previous supremum defines an equivalent norm on Θ_{1+c}. These follow from Lemma 3.1(a) of Cox (1988). Now, $\langle \zeta, U\theta \rangle = \langle \zeta, \theta \rangle_0$ for $\zeta, \theta \in \Theta$, so $\langle \zeta, U\theta \rangle \leq \|\zeta\|_0 \|\theta\|_0$ and it follows that $U\theta \in \Theta_2$. Further,

$$\|U\theta\|_2 \leq \text{constant} \sup_{\|\zeta\|_0 = 1} \langle \zeta, U\theta \rangle \leq \text{constant} \|\theta\|_0,$$

so $U \in \mathbf{B}(\Theta_0, \Theta_2)$ and hence $U \in \mathbf{B}(\Theta_b, \Theta_2)$, for any $b \geq 0$.

For W, first note that $W\phi_\nu = \gamma_\nu U\phi_\nu$, so if $\zeta \in \Theta_{2-b}$,

$$\begin{aligned} \langle W\theta, \zeta \rangle &= \left\langle W\theta, \sum_\nu \langle \zeta, U\phi_\nu \rangle \phi_\nu \right\rangle \\ &= \sum_\nu \gamma_\nu \langle \zeta, U\phi_\nu \rangle \langle \theta, U\phi_\nu \rangle \\ &\leq \left\{ \sum_\nu \gamma_\nu^{2-b} \langle \zeta, U\phi_\nu \rangle^2 \right\}^{1/2} \left\{ \sum_\nu \gamma_\nu^b \langle \theta, U\phi_\nu \rangle^2 \right\}^{1/2} \\ &\leq \|\zeta\|_{2-b} \|\theta\|_b, \end{aligned} \tag{2.6}$$

so $W\theta \in \Theta_b$ and $\|W\theta\|_b \leq \text{constant} \|\theta\|_b$, thus proving the result. □

2.3. *Derivatives.* Derivatives must be introduced in order to define and analyze the linearizations. Here we use the Frechet derivative which is the strongest notion of a derivative in a normed linear space; see, for example, Chapter 3 of Rall (1969). We use the standard notation D for the Frechet derivative operator. Let $l(\theta)$ denote the limiting data fit functional.

ASSUMPTION A.3. For some $\alpha \in (0, 1]$, there is a $\Theta_0 \in \Theta_\alpha$ and a Θ_α-neighborhood N_{θ_0} of θ_0 such that

(i) l_n and l are three times continuously Frechet differentiable in N_{θ_0}.
(ii) θ_0 is the unique root of $Dl(\theta_0) = 0$ in N_{θ_0}.

Part (ii) of A.3 may be viewed as a definition of the *true* parameter θ_0. Note that $D^k l(\theta)$ (and $D^k l_n(\theta)$) is a bounded multilinear operator of order k. $Dl(\theta)$ and $Dl_n(\theta)$ are bounded linear functionals on Θ_α. It will be necessary to carefully represent the dual Θ_α^*. As in the proof of Lemma 2.1, a linear functional $f \in \Theta_\alpha^*$ may be represented via the Θ duality pairing as $f(\theta) = \langle \zeta, \theta \rangle$, for all $\theta \in \Theta_\alpha$, where $\zeta \in \Theta_{2-\alpha}$ and $\|f\|_{\Theta_\alpha^*}$ may be bounded above and below by strictly positive multiples of $\|\zeta\|_{2-\alpha}$. With this representation of Θ_α^*, we have $Dl(\theta) \in \Theta_{2-\alpha}$ for $\theta \in N_{\theta_0}$ and in particular $Dl(\theta) \in \Theta$. Now Dl is a mapping from Θ_α to $\Theta_{2-\alpha}$, so for $\theta \in N_{\theta_0}$,

$$U(\theta) \overset{\text{def}}{=} D^2 l(\theta) \in \mathbf{B}(\Theta_\alpha, \Theta_{2-\alpha}). \tag{2.7}$$

Thus if $\zeta, \eta \in \Theta_\alpha$, then

$$D^2 l(\theta)\zeta\eta = \langle \zeta, U(\theta)\eta \rangle. \tag{2.8}$$

[On the l.h.s. one sees our notation for multilinear operators which we have borrowed from Rall (1969).] The next assumption guarantees in fact that $U(\theta) \in B(\Theta_b, \Theta_2)$, similar to U.

ASSUMPTION A.4. There are strictly positive constants m and M such that for all $\Theta \in N_{\theta_0}$ and for all $\zeta \in \Theta$,

$$m\langle \zeta, U\zeta \rangle \le \langle \zeta, U(\theta)\zeta \rangle \le M\langle \zeta, U\zeta \rangle. \tag{2.9}$$

Further, for $\theta \in N_{\theta_0}$, $D^3 l(\theta) \in B(\Theta_\alpha, B(\Theta_\alpha, \Theta_{2-\alpha}))$ and so $D^3 l(\theta)\zeta\eta \in \Theta_{2-\alpha}$, for $\zeta, \eta \in \Theta_\alpha$. All remarks above also hold for $D^k l_n(\theta)$. For each $\theta_* \in N_{\theta_0}$, there is a sequence $\{\phi_{*\nu}: \nu = 1, 2, \dots\}$ of eigenfunctions and $\{\gamma_{*\nu}: \nu = 1, 2, \dots\}$ of eigenvalues which satisfy

$$\langle \phi_{*\mu}, U(\theta_*)\phi_{*\nu} \rangle = \delta_{\nu\mu} \quad \text{and} \quad \langle \phi_{*\mu}, W\phi_{*\nu} \rangle = \gamma_{*\nu}\delta_{\nu\mu}, \tag{2.10}$$

and for $b \ge 0$, we have norms

$$\|\theta\|_{*b} = \left\{ \sum_{\nu=1}^{\infty} (1 + \gamma_{*\nu}^b)\langle \theta, U(\theta_*)\phi_{*\nu} \rangle^2 \right\}^{1/2}. \tag{2.11}$$

Let Θ_{*b} denote the normed linear space obtained by completing $\{\theta \in \Theta:$

$\|\theta\|_{*b} < \infty\}$ in $\|\cdot\|_{*b}$ norm. From A.4, there are positive constants c_1 and c_2 such that for all ν large enough,

$$c_1\gamma_\nu \le \gamma_{*\nu} \le c_2\gamma_\nu. \tag{2.12}$$

The following proposition for $b = 0$ follows from A.4 and for $0 < b < 1$ from the K-method of interpolation.

PROPOSITION 2.1. *For* $0 \le b \le 1$, $\Theta_{*b} = \Theta_b$.

Various elementary facts relating to the convergence norms should be noted. The proof of Lemma 2.2 is straightforward; see Cox (1988). We will make repeated use of this in Section 4.

LEMMA 2.2. *For* $\theta_* \in N_{\theta_0}$, $b > 0$ *and* $\nu = 1, 2, \ldots,$

(i) $\|\phi_\nu\|_b^2 = 1 + \gamma_\nu^b$ *and* $\|\phi_{*\nu}\|_{*b}^2 = 1 + \gamma_{*\nu}^b$.

(ii) $[U + \lambda W]^{-1}U\phi_\nu = (1 + \lambda\gamma_\nu)^{-1}\phi_\nu$ *and* $[U(\theta_*) + \lambda W]^{-1}U(\theta_*)\phi_{*\nu} = (1 + \lambda\gamma_{*\nu})^{-1}\phi_{*\nu}$.

(iii) *Suppose* $\gamma_\nu \approx \nu^r$ *for some* $r > 0$, *meaning* γ_ν/ν^r *is bounded away from* 0 *and* ∞ *as* $\nu \to \infty$. *Then for* $b \ge 0$ *and* $c \ge 0$ *with* $b + c < 2 - 1/r$, *uniformly in* $\theta_* \in N_{\theta_0}$,

$$\sum_\nu \left(1 + \gamma_{*\nu}^b\right)\left(1 + \gamma_{*\nu}^c\right)\left(1 + \lambda\gamma_{*\nu}\right)^{-2} \approx \lambda^{-(b+c+1/r)} \quad \text{as } \lambda \to 0, \tag{2.13}$$

meaning that the supremum (infimum), over $\theta_* \in N_{\theta_0}$, *of the ratio of the quantity on the left to that on the right remains bounded away from* $\infty(0)$ *as* $\lambda \to 0$.

Now we may show that the linearizations are well-defined. By the Lax–Milgram theorem [Section 3.6 of Aubin (1979)], A.2 and A.4, the operator

$$G_\lambda(\theta) = U(\theta) + \lambda W \tag{2.14}$$

has a bounded inverse on Θ. Hence, $G_\lambda(\theta)^{-1}Dl(\zeta)$ and $G_\lambda(\theta)^{-1}Dl_n(\zeta)$ are well-defined elements of Θ for $\theta, \zeta \in N_{\theta_0}$ [recall $Dl(\zeta), Dl_n(\zeta) \in \Theta_{2-\alpha} \subset \Theta$, since $\alpha \le 1$]. Thus the continuous linearization in (1.12) is well-defined. To show $G_\lambda(\theta)^{-1}Z_{n\lambda}(\zeta)$ is well-defined for $\theta, \zeta \in N_{\theta_0}$ [see (1.14)], we need to show $G_\lambda(\theta)^{-1}\{\lambda W\zeta\} = [I - G_\lambda(\theta)^{-1}U(\theta)]\zeta$ is well-defined. But taking $\theta_* = \theta$,

$$\begin{aligned} G_\lambda(\theta_*)^{-1}U(\theta_*)\zeta &= \sum_\nu \langle G_\lambda(\theta_*)^{-1}U(\theta_*)\zeta, U(\theta_*)\phi_{*\nu}\rangle\phi_{*\nu} \\ &= \sum_\nu (1 + \lambda\gamma_{*\nu})^{-1}\langle\zeta, U(\theta_*)\phi_{*\nu}\rangle\phi_{*\nu} \end{aligned} \tag{2.15}$$

and the series converges for $\zeta \in \Theta_{*0}$ to an element of Θ_{*2}. Finally note that if $\zeta \in \Theta_\alpha$, then for $\theta_* \in N_{\theta_0}$,

$$\left\|G_\lambda(\theta_*)^{-1}U(\theta_*)\zeta\right\|_{*b}^2 = \sum_\nu \left(1 + \gamma_{*\nu}^b\right)(1 + \lambda\gamma_{*\nu})^{-2}\langle\zeta, U(\theta_*)\phi_{*\nu}\rangle^2. \tag{2.16}$$

2.4. *Linear expansions with bounds on remainders.* Taylor series expansion for the generalized score vectors is a key device used in the analysis to follow. If $f:H \to H$, where H is a Hilbert space, then assuming the requisite Frechet derivatives exist, we have $f(\theta + \phi) = f(\theta) + \int_0^1 Df(\theta + s\phi)\phi\, ds$. A second expansion inside the integral sign yields

$$f(\theta + \phi) = f(\theta) + Df(\theta)\phi + \int_0^1\int_0^1 sD^2f(\theta + s's\phi)\phi\phi\, ds'\, ds. \tag{2.17}$$

Note the H-valued integrals are of functions mapping $[0,1] \to H$. Under continuity assumptions (which will hold for us by A.3), such integrals are readily defined as limits of Riemann sums [see Rall (1969), for example]. It is easy to verify from the triangle inequality applied to the Riemann sums that the norm of the integral is bounded by the integral of the norm.

Before analyzing expansions for the generalized score vectors, we need to define some norms for derivative operators. For $0 \le b \le \alpha$, $\lambda > 0$, $\theta_1, \theta_2 \in N_{\theta_0}$ and u, ν unit elements in Θ_α (so $\|u\|_\alpha = \|\nu\|_\alpha = 1$), let

$$\begin{aligned}
K_{2n}(\lambda, b) &= \sup_{\theta_1,\theta_2}\sup_{u} \left\| G_\lambda(\theta_1)^{-1}\left[D^2l_n(\theta_2)u - D^2l(\theta_2)u\right]\right\|_b, \\
K_3(\lambda, b) &= \sup_{\theta_1,\theta_2}\sup_{u,\nu} \left\| G_\lambda(\theta_1)^{-1}\left[D^3l(\theta_2)u\nu\right]\right\|_b, \\
K_{3n}(\lambda, b) &= \sup_{\theta_1,\theta_2}\sup_{u,\nu} \left\| G_\lambda(\theta_1)^{-1}\left[D^3l_n(\theta_2)u\nu\right]\right\|_b.
\end{aligned} \tag{2.18}$$

[The discussions following Assumption A.3 and Lemma 2.2 may be adapted to show that quantities such as $G_\lambda(\theta_1)^{-1}[D^2l_n(\theta_2)u - D^2l(\theta_2)u]$ appearing in these expressions are well-defined elements of Θ.] By Taylor series expansion if $\theta_0 + \phi \in N_{\theta_0}$, then

$$\begin{aligned}
&Z_\lambda(\theta_0 + \phi) \\
&\quad = Z_\lambda(\theta_0) + D^2l_\lambda(\theta_0)\phi + \int_0^1\int_0^1 sD^3l(\theta_0 + s's\phi)\phi\phi\, ds'\, ds \\
&\quad = Z_\lambda(\theta_0) + G_\lambda(\theta_0)\phi + \int_0^1\int_0^1 sD^3l(\theta_0 + s's\phi)\phi\phi\, ds'\, ds.
\end{aligned} \tag{2.19}$$

From the definition of $K_3(\lambda, b)$, the integral remainder is bounded as

$$\left\| G_\lambda(\theta_0)^{-1}\int_0^1\int_0^1 sD^3l(\theta_0 + s's\phi)\phi\phi\, ds'\, ds \right\|_b \le \left\{\tfrac{1}{2}K_3(\lambda, b)\|\phi\|_\alpha\right\}\|\phi\|_\alpha. \tag{2.20}$$

(Note that a bounded linear operator can be interchanged with the integral.) Similarly, for $\theta_\lambda, \theta_\lambda + \phi \in N_{\theta_0}$,

$$\begin{aligned}
Z_{n\lambda}(\theta_\lambda + \phi) &= Z_{n\lambda}(\theta_\lambda) + D^2l_{n\lambda}(\theta_\lambda)\phi \\
&\quad + \int_0^1\int_0^1 sD^3l_n(\theta_\lambda + s's\phi)\phi\phi\, ds'\, ds.
\end{aligned} \tag{2.21}$$

So

$$
\begin{aligned}
&Z_{n\lambda}(\theta_\lambda + \phi) - Z_{n\lambda}(\theta_\lambda) - G_\lambda(\theta_\lambda)\phi \\
&\quad = \{D^2 l_n(\theta_\lambda)\phi - D^2 l(\theta_\lambda)\phi\} + \int_0^1 \int_0^1 sD^3 l_n(\theta_\lambda + s's\phi)\phi\phi \, ds' \, ds.
\end{aligned} \tag{2.22}
$$

Using the definitions of K_{2n} and K_{3n}, the remainder is bounded as

$$
\begin{aligned}
&\|G_\lambda(\theta_\lambda)^{-1}\{Z_{n\lambda}(\theta_\lambda + \phi) - Z_{n\lambda}(\theta_\lambda) - G_\lambda(\theta_\lambda)\phi\}\|_b \\
&\quad \le \|G_\lambda(\theta_\lambda)^{-1}\{D^2 l_n(\theta_\lambda)\phi - D^2 l(\theta_\lambda)\phi\}\|_b \\
&\qquad + \left\| G_\lambda(\theta_\lambda)^{-1}\left\{ \int_0^1 \int_0^1 sD^3 l_n(\theta_\lambda + s's\phi)\phi\phi \, ds' \, ds \right\} \right\|_b \\
&\quad \le \{K_{2n}(\lambda, b) + \tfrac{1}{2}K_{3n}(\lambda, b)\|\phi\|_\alpha\}\|\phi\|_\alpha.
\end{aligned} \tag{2.23}
$$

3. General linearization results. Now we show that the linearizations introduced in Section 1 accurately approximate the asymptotic bias and variability of roots of penalized likelihood variational equations. The existence of locally unique roots of Z_λ and $Z_{n\lambda}$ is first established. The method of analysis uses a fixed point argument similar to the one employed by Huber (1973). Assumptions A.1–A.4 are in force throughout the section.

3.1. *Continuous linearization and bias approximation.* It will be convenient to have the following notation for balls in Θ_b:

$$
S_\theta(r, b) = \{\eta \in \Theta_b : \|\eta - \theta\|_b \le r\}, \qquad S(r, b) = S_0(r, b). \tag{3.1}
$$

For $0 \le b \le \alpha$ and $0 < \lambda < \infty$, put $d(\lambda, b) = \|\bar{\theta}_\lambda - \theta_0\|_b$ and $r(\lambda, b) = K_3(\lambda, b)d(\lambda, \alpha)$.

ASSUMPTION A.5. As $\lambda \to 0$, both $d(\lambda, \alpha) \to 0$ and $r(\lambda, \alpha) \to 0$.

THEOREM 3.1 (Existence of θ_λ and the bias approximation). *Under* A.5 *there exists* $\lambda_0 > 0$ *such that for* $\lambda \in [0, \lambda_0]$,

(i) *there is a unique* $\theta_\lambda \in S_{\theta_0}(2d(\lambda, \alpha), \alpha)$ *satisfying* $Z_\lambda(\theta_\lambda) = 0$ *and* $\theta_\lambda \in N_{\theta_0}$,

(ii) $\|\bar{\theta}_\lambda - \theta_\lambda\|_b \le 4r(\lambda, b)d(\lambda, \alpha)$ *as* $\lambda \to 0$, *for* $b \in (0, \alpha]$.

PROOF. Let

$$
F_\lambda(\phi) = \theta - G_\lambda(\theta_0)^{-1} Z_\lambda(\theta_0 + \phi). \tag{3.2}
$$

For convenience, put $t_\lambda = 2d(\lambda, \alpha)$. The proof has three steps: (a) $F_\lambda(S(t_\lambda, \alpha)) \subset S(t_\lambda, \alpha)$; (b) F_λ is a contraction map on $S(t_\lambda, \alpha)$; and (c) obtaining the estimate in part (ii) of the theorem.

For (a), by A.5, let λ_0 be chosen so that $S_{\theta_0}(t_\lambda, \alpha) \subset N_{\theta_0}$ and $r(\lambda, \alpha) < \frac{1}{2}$ for all $\lambda \in (0, \lambda_0]$. For $\phi \in S(t_\lambda, \alpha)$,

$$\|F_\lambda(\phi)\|_b \leq \left\|\phi - G_\lambda(\theta_0)^{-1} Z_\lambda(\theta_0 + \phi) - (\bar{\theta}_\lambda - \theta_0)\right\|_b + \|\bar{\theta}_\lambda - \theta_0\|_b. \tag{3.3}$$

Using the definition of $\bar{\theta}_\lambda$ and a Taylor series expansion of $Z_\lambda(\theta_0 + \phi)$ as described in Section 2,

$$\begin{aligned}
&\left\|\phi - G_\lambda(\theta_0)^{-1} Z_\lambda(\theta_0 + \phi) - (\bar{\theta}_\lambda - \theta_0)\right\|_b \\
&\quad = \left\|G_\lambda^{-1}(\theta_0)\left[Z_\lambda(\theta_0 + \phi) - Z_\lambda(\theta_0) - G_\lambda(\theta_0)\phi\right]\right\|_b \\
&\quad \leq \tfrac{1}{2} K_3(\lambda, b)\|\phi\|_\alpha^2.
\end{aligned}$$

Thus (with $b = \alpha$) for $\phi \in S(t_\lambda, \alpha)$,

$$\|F_\lambda(\phi)\|_\alpha \leq \left[\tfrac{1}{2} K_3(\lambda, \alpha) t_\lambda + \tfrac{1}{2}\right] t_\lambda = \left[r(\lambda, \alpha) + \tfrac{1}{2}\right] t_\lambda < t_\lambda, \tag{3.4}$$

which completes step (a) of the proof.

For step (b), let $\phi_1, \phi_2 \in S(t_\lambda, \alpha)$, since

$$\begin{aligned}
Z_\lambda(\theta_0 + \phi_2) &= Z_\lambda(\theta_0 + \phi_1) \\
&\quad + \int_0^1 DZ_\lambda(\theta_0 + \phi_1 + s(\phi_2 - \phi_1))(\phi_2 - \phi_1)\, ds,
\end{aligned} \tag{3.5}$$

by Taylor series expansion about θ_0, inside the integral we have

$$\begin{aligned}
Z_\lambda(\theta_0 + \phi_2) &= Z_\lambda(\theta_0 + \phi_1) + G_\lambda(\theta_0)(\phi_2 - \phi_1) \\
&\quad + \int_0^1 \int_0^1 D^2 Z_\lambda(\theta_0 + s'(\phi_1 + s(\phi_2 - \phi_1))) \\
&\qquad \times (\phi_2 - \phi_1)(\phi_1 + s(\phi_2 - \phi_1))\, ds'\, ds.
\end{aligned}$$

Thus

$$\begin{aligned}
F_\lambda(\phi_1) - F_\lambda(\phi_2) = G_\lambda(\theta_0)^{-1} \Bigg\{ &\int_0^1 \int_0^1 D^3 l(\theta_0 + s'(\phi_1 + s(\phi_2 - \phi_1))) \\
&\times (\phi_2 - \phi_1)(\phi_1 + s(\phi_2 - \phi_1))\, ds'\, ds \Bigg\}.
\end{aligned}$$

Since $\phi_1 + s(\phi_2 - \phi_1) \in S(t_\lambda, \alpha)$ by convexity, for $0 \leq b \leq \alpha$,

$$\|F_\lambda(\phi_1) - F_\lambda(\phi_2)\|_b \leq K_3(\lambda, b) t_\lambda \|\phi_2 - \phi_1\|_\alpha = 2r(\lambda, b)\|\phi_2 - \phi_1\|_\alpha. \tag{3.6}$$

From our choice of λ_0, it follows that F_λ is a contraction on $S(t_\lambda, \alpha)$, for $\lambda \in (0, \lambda_0]$ (put $b = \alpha$ in the above line). The contraction mapping theorem [Rudin (1976)] gives a unique $\phi_\lambda \in S(t_\lambda, \alpha)$ for which $F_\lambda(\phi_\lambda) = \phi_\lambda$. Letting $\theta_\lambda = \theta_0 + \phi_\lambda$, θ_λ is the unique root of Z_λ in $S_{\theta_0}(t_\lambda, \alpha)$. This completes step (b) of the proof.

For the final step of the proof, note, using the definition of F_λ, that

$$\theta_\lambda - \bar{\theta}_\lambda = (\theta_\lambda - \theta_0) - (\bar{\theta}_\lambda - \theta_0) = F_\lambda(\phi_\lambda) - F_\lambda(0). \tag{3.7}$$

Thus, since $\phi_\lambda \in S(t_\lambda, \alpha)$,

$$\|\theta_\lambda - \bar{\theta}_\lambda\|_b = \|F_\lambda(\phi_\lambda) - F_\lambda(0)\|_b \le 2r(\lambda, b)\|\phi_\lambda\|_\alpha \le 4r(\lambda, b)d(\lambda, \alpha). \tag{3.8}$$

This completes the final step of the proof. □

3.2. *Discrete linearization and variability approximation.* The next result justifies the discrete linearization for a fixed sequence of λ's. For $0 \le b \le \alpha$ and $\theta_\lambda \in N_{\theta_0}$, put

$$d_n(\lambda, b) = \|\bar{\theta}_{n\lambda} - \theta_\lambda\|_b \text{ and } r_n(\lambda, b) = K_{2n}(\lambda, b) + K_{3n}(\lambda, b)d(\lambda, \alpha). \tag{3.9}$$

ASSUMPTION A.6. λ_n is a sequence such that for all n sufficiently large, $\theta_{\lambda_n} \in N_{\theta_0}$ and

$$d_n(\lambda_n, \alpha) \to_p 0 \quad \text{and} \quad r_n(\lambda_n, \alpha) \to_p 0.$$

THEOREM 3.2 (Existence of $\theta_{n\lambda}$ and the variability approximation). *Let λ_n be a sequence satisfying* A.6. *Then with probability tending to unity as $n \to \infty$,*

(i) *there is a unique root $\theta_{n\lambda_n}$ of $Z_{n\lambda_n}(\theta) = 0$ in $S_{\theta_0}(2d_n(\lambda_n, \alpha), \alpha)$,*
(ii) *for $b \in [0, \alpha]$, $\|\theta_{n\lambda_n} - \bar{\theta}_{n\lambda_n}\|_b \le 4r_n(\lambda_n, b)d_n(\lambda_n, \alpha)$.*

PROOF. For convenience, drop the subscript on λ_n and let $t_{n\lambda} = 2d_n(\lambda, \alpha)$. Let

$$F_{n\lambda}(\phi) = \phi - G_\lambda(\theta_\lambda)^{-1}Z_{n\lambda}(\theta_\lambda + \phi). \tag{3.10}$$

The proof proceeds in three steps analogous to the proof of Theorem 3.1, except extra terms are introduced in approximating $DZ_{n\lambda}$ by DZ_λ. Take n large enough so that the event $S_{\theta_\lambda}(t_{n\lambda}, \alpha) \subset N_{\theta_0}$ and $r_n(\lambda, \alpha) < \frac{1}{2}$ occurs with probability arbitrarily close to unity. This is possible by A.6. For the remainder of the theorem we restrict attention to this event. Manipulating the expression for $F_{n\lambda}$ gives

$$\begin{aligned} F_{n\lambda}(\phi) = {} & -G_\lambda(\theta_\lambda)^{-1}\left[D^2 l_n(\theta_\lambda)\phi - D^2 l(\theta_\lambda)\phi\right] \\ & - G_\lambda(\theta_\lambda)^{-1}\left[Dl_n(\theta_\lambda + \phi) - Dl_n(\theta_\lambda) - D^2 l_n(\theta_\lambda)\phi\right] \\ & + (\bar{\theta}_{n\lambda} - \theta_\lambda). \end{aligned} \tag{3.11}$$

Now using the definitions of $K_{2n}(\lambda, \alpha)$ and $K_{3n}(\lambda, \alpha)$ and Taylor series expansions as before,

$$\begin{aligned} \|F_{n\lambda}(\phi)\|_\alpha &\le \left\|G_\lambda(\theta_\lambda)^{-1}\left[D^2 l_n(\theta_\lambda)\phi - D^2 l(\theta_\lambda)\phi\right]\right\|_\alpha \\ &\quad + \sup_{s\in[0,1]} \left\|G_\lambda(\theta_\lambda)^{-1}D^3 l_n(\theta_\lambda + s\phi)\phi\phi\right\|_\alpha + \tfrac{1}{2}t_{n\lambda} \\ &\le \left[K_{2n}(\lambda, \alpha) + \tfrac{1}{2}K_{3n}(\lambda, \alpha)t_{n\lambda} + \tfrac{1}{2}\right]t_{n\lambda} \\ &\le \left[r_n(\lambda, \alpha) + \tfrac{1}{2}\right]t_{n\lambda} < t_{n\lambda}. \end{aligned} \tag{3.12}$$

Thus, $F_{n\lambda}(S(t_{n\lambda}, \alpha)) \subset S(t_{n\lambda}, \alpha)$.

For the second step, by considering the expression for $F_{n\lambda}$ above and expanding as in the proof of step (b) in Theorem 3.1, we have for $\phi_1, \phi_2 \in S(t_{n\lambda}, \alpha)$,

$$\begin{aligned} \|F_{n\lambda}(\phi_1) - F_{n\lambda}(\phi_2)\|_b &\le [K_{2n}(\lambda, b) + K_{3n}(\lambda, b)t_{n\lambda}]\|\phi_1 - \phi_2\|_\alpha \\ &\le 2r_n(\lambda, b)\|\phi_1 - \phi_2\|_\alpha, \end{aligned} \tag{3.13}$$

for $0 \le b \le \alpha$. Letting $b = \alpha$, this shows $F_{n\lambda}$ is a contraction map on $S(t_{n\lambda}, \alpha)$. This suffices to establish part (i) of the theorem.

Letting $\theta_{n\lambda} = \theta_\lambda + \phi_{n\lambda}$, where $\phi_{n\lambda}$ is the fixed point of $F_{n\lambda}$, $Z_{n\lambda}(\theta_{n\lambda}) \equiv 0$ and

$$\begin{aligned} \|\theta_{n\lambda} - \bar{\theta}_{n\lambda}\|_b &= \left\|G_\lambda(\theta_\lambda)^{-1}G_\lambda(\theta_\lambda)(\theta_{n\lambda} - \bar{\theta}_{n\lambda})\right\|_b \\ &= \left\|G_\lambda(\theta_\lambda)^{-1}[G_\lambda(\theta_\lambda)\theta_{n\lambda} - G_\lambda(\theta_\lambda)\theta_\lambda + Z_{n\lambda}(\theta_\lambda)]\right\|_b \\ &= \|F_{n\lambda}(\phi_{n\lambda}) - F_{n\lambda}(0)\|_b \le 4r_n(\lambda, b)d_n(\lambda, \alpha), \end{aligned} \tag{3.14}$$

proving part (ii) of the theorem. □

4. Some illustrative applications. We now consider the examples introduced in Section 1. Our goal is to prove Theorem 1.1. We proceed by verifying Assumptions A.1–A.6.

Assumption A.1 is immediate. For Assumption A.2, take $U = J^*J$, where J is the injection of Θ into $L_2[0, 1]$, i.e.,

$$\langle \theta, U\zeta \rangle = \int_0^1 \theta(t)\zeta(t)\, dt. \tag{4.1}$$

(2.1) follows immediately. U is compact; see Cox (1988). From Cox (1988), the convergence norms are immediately associated with Sobolev norms:

$$\Theta_b = W_2^{mb}[0, 1] \qquad 0 \le b \le 1, \tag{4.2}$$

and $\gamma_\nu \approx \nu^{2m}$ as $\nu \to \infty$.

4.1. *Derivative formulas and assumptions* A.3 *and* A.4. The following expressions are easily established.

(4.3)

$$Dl(\theta)\phi = \begin{cases} \int_0^1 [e^{\theta(t)} - e^{\theta_0(t)}]\phi(t)\, dt & \text{(i) log density,} \\ \int_0^1 [e^{\theta(t)} - e^{\theta_0(t)}]\phi(t)S(t)\, dt & \text{(ii) log hazard,} \\ \int_0^1 [p(\theta(t)) - p(\theta_0(t))]\phi(t)f(t)\, dt & \text{(iii) logistic regression.} \end{cases}$$

(Future instances of such multiline formulae will refer to Examples 1, 2 and 3 in that order.)

$$Dl_n(\theta)\phi = \begin{cases} \int_0^1 e^{\theta(t)}\phi(t)\,dt - \dfrac{1}{n}\sum_{i=1}^n \phi(X_i), \\ \int_0^1 e^{\theta(t)}\phi(t)S_n(t)\,dt - \dfrac{1}{n}\sum_{i=1}^n \delta_i\phi(X_i), \\ \dfrac{1}{n}\sum_{i=1}^n \left[p(\theta(X_i))\phi(X_i) - Y_i\phi(X_i)\right]. \end{cases} \tag{4.4}$$

For the second and third order derivatives there are the representations:

$$D^2l(\theta)\phi\zeta = \int_0^1 \phi(t)\zeta(t)g(t,\theta(t))\,dA(t)$$

and

$$D^2l_n(\theta)\phi\zeta = \int_0^1 \phi(t)\zeta(t)g(t,\theta(t))\,dA_n(t).$$

Furthermore,

$$D^3l(\theta)\phi\zeta\psi = \int_0^1 \phi(t)\zeta(t)\psi(t)h(t,\theta(t))\,dA(t)$$

and

$$D^3l_n(\theta)\phi\zeta\psi = \int_0^1 \phi(t)\zeta(t)\psi(t)h(t,\theta(t))\,dA_n(t).$$

$A(t)$ and $A_n(t)$ are given by

$$A(t) = \begin{cases} t, \\ \int_0^t S(s)\,ds, \\ F(t), \end{cases} \qquad A_n(t) = \begin{cases} t, \\ \int_0^1 S_n(s)\,ds, \\ F_n(t). \end{cases} \tag{4.5}$$

Here F and F_n are the c.d.f. and empirical c.d.f. of the X_i's in the density and logistic regression settings; $g(t,\mu)$ and $h(t,\mu)$ are given by

$$g(t,\mu) = \begin{cases} e^\mu, \\ e^\mu, \\ p(\mu)(1-p(\mu)), \end{cases}$$

$$h(t,\mu) = \begin{cases} e^\mu, \\ e^\mu, \\ p(\mu)(1-p(\mu))(1-2p(\mu)). \end{cases} \tag{4.6}$$

The true parameter lies in Θ_p, where $3/(2m) < p \le 1$. We choose $1/(2m) < \alpha \le p$. It is not difficult to verify that the above derivatives are well-defined and continuous in Θ_α. From (4.3), θ_0 is a root of $Dl(\theta)$ and since $l(\theta)$ is strictly convex, θ_0 is the unique root of $Dl(\theta)$ in Θ_α. This verifies Assumption

A.3.

For any $R > 0$, let $N_{\theta_0} = S_{\theta_0}(R, \alpha)$. If θ_0 lies in the interior of some constraint set $C \subset \Theta_\alpha$, then we should choose R so that $N_{\theta_0} \subset C$. For $\theta \in N_{\theta_0}$, there are constants m_R and M_R such that for $t \in [0, 1]$,

$$0 < m_R \le g(t, \theta(t)) \le M_R, \tag{4.7}$$

$$|h(t, \theta(t))| \le M_R. \tag{4.8}$$

Assumption A.4 follows from the first of these relations and the fact that $dA(t) = \nu(t)\,dt$, where ν is bounded away from zero and infinity.

4.2. *Error analysis.* The error analysis is carried out via the linearization technique described in Section 3. We only consider results for a fixed sequence of λ's. [Some uniformity in λ should be possible, see O'Sullivan (1989) for example, but we do not pursue that here.] Convergence characteristics in $\|\cdot\|_b$-norm for $0 \le b \le \alpha$ are studied. The $b = 0$ case corresponds to the usual integrated squared error.

ASSUMPTION A.5 AND THE SYSTEMATIC ERROR APPROXIMATION. From Theorem 2.3 of Cox (1988), uniformly in b as $\lambda \to 0$,

$$d(\lambda, b) = \|\bar{\theta}_\lambda - \theta_0\|_b = O(\lambda^{(p-b)/2})\|\theta_0\|_p \quad \text{for } 0 \le b \le \alpha. \tag{4.9}$$

Using the expansion indicated after Lemma 2.2,

$$\begin{aligned} &\left\|G_\lambda^{-1}(\theta_1) D^3 l(\theta_0 + u)\nu w\right\|_b^2 \\ &\quad \approx \sum_{\nu=1}^{\infty} \left[1 + \gamma_{*\nu}^b\right]\left[1 + \lambda\gamma_{*\nu}\right]^{-2} \langle D^3 l(\theta_0 + u)\nu w, \phi_{*\nu}\rangle^2, \end{aligned} \tag{4.10}$$

where $\theta_* = \theta_1 \in N_{\theta_0}$. Using the derivative representations, the second part of (4.8), Holder's inequality and Sobolev's imbedding theorem [$\sup|\nu| \le M\|\nu\|_\alpha$ for $\alpha > 1/(2m)$, see Adams (1975)],

$$\begin{aligned} \langle D^3 l(\theta_0 + u)\nu w, \phi\rangle^2 &= \left[\int_0^1 h(t, \theta_0(t) + n(t)) w(t)\nu(t)\phi(t)\,dt\right]^2 \\ &\le M\|w\|_\alpha^2 \|\nu\|_\alpha^2 \|\phi\|_0^2, \end{aligned} \tag{4.11}$$

where M is a positive constant. Invoking Lemma 2.2 we obtain

$$\left\|G_\lambda^{-1}(\theta_1) D^3 l(\theta_0 + u)\nu w\right\|_b \le M\|w\|_\alpha \|\nu\|_\alpha \lambda^{-(b+1/(2m))/2}, \tag{4.12}$$

so

$$K_3(\lambda, b) \le M\lambda^{-(b+1/(2m))/2}.$$

From this and (4.9),

$$r(\lambda, b) \le M\lambda^{(p-\alpha-b-1/(2m))/2}. \tag{4.13}$$

If $p > 3/(2m)$, then there is an $\alpha \in (1/(2m), (p - 1/(2m))/2]$ such that $r(\lambda, \alpha) \to 0$ as $\lambda \to 0$. This verifies Assumption A.5 so from Theorem 3.1 and

the strict convexity of l_λ, we have the result:

THEOREM 4.1 (Systematic error bound). *Suppose* $\theta_0 \in \Theta_p$, *for* $3/(2m) < p \le 1$ *and* $\alpha \in (1/(2m), (p - 1/(2m))/2]$. *Then there is some* $\lambda_0 > 0$ *such that for any* $0 < \lambda < \lambda_0$ *and* $0 \le b \le \alpha$, *there is a unique root* θ_λ *of* Z_λ *which satisfies*

$$\|\theta_\lambda - \bar{\theta}_\lambda\|_b \le M\lambda^{(p-b)/2}\lambda^{(p-2\alpha-1/(2m))/2}. \tag{4.14}$$

Furthermore, for $0 \le b \le \alpha$,

$$\|\theta_\lambda - \theta_0\|_b \le M\lambda^{(p-b)/2}. \tag{4.15}$$

ASSUMPTION A.6 AND THE STOCHASTIC ERROR APPROXIMATION. Since $\lambda = \lambda_n \to 0$, we can (by Theorem 4.1) choose λ such that $\theta_\lambda \in N_{\theta_0}$. An expansion of the norm gives

$$\begin{aligned} d_n(\lambda, b)^2 &= \|\bar{\theta}_{n\lambda} - \theta_\lambda\|_b^2 \\ &\approx \sum_{\nu=1}^{\infty} \left[1 + \gamma_{*\nu}^b\right]\left[1 + \lambda\gamma_{*\nu}\right]^{-2}\left[\{Dl_n(\theta_*) - Dl(\theta_*)\}\phi_{*\nu}\right]^2, \end{aligned} \tag{4.16}$$

where $\theta_* = \theta_\lambda$. We will need to study the behavior of $[\{Dl_n(\theta_\lambda) - Dl(\theta_\lambda)\}\phi]^2$. From (4.3) and (4.4),

$$Dl_n(\theta_\lambda)\phi - Dl(\theta_\lambda)\phi = \int_0^1 \phi(t)\, dU_n(t; \theta_\lambda), \tag{4.17}$$

with

$$U_n(t; \theta_\lambda) = \begin{cases} F_n(t) - F(t), \\ \int_0^1 e^{\theta_\lambda(s)}\left[S_n(s) - S(s)\right] ds - V_n(t), \\ \int_0^1 p(\theta_\lambda(s))\, d\left[F_n(s) - F(s)\right] - W_n(t). \end{cases} \tag{4.18}$$

$V_n(t)$ and $W_n(t)$ are

$$V_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i I_t(Y_i) - E\delta_i I_t(Y_i)\right], \tag{4.19}$$

$$W_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i I_t(X_i) - EY_i I_t(X_i)\right]. \tag{4.20}$$

Here I_t is the characteristic function of $[0, t]$. A direct computation of expectations gives

$$E\{Dl_n(\theta_\lambda)\phi - Dl(\theta_\lambda)\phi\}^2 = E\left\{\int_0^1 \phi(t)\, dU_n(t; \theta_\lambda)\right\}^2 \le Mn^{-1}\|\phi\|_0^2. \tag{4.21}$$

Thus substituting into (4.16) and using Lemma 2.2 we have for $0 \le b \le \alpha$,

$$d_n(\lambda, b)^2 = O_p(n^{-1}\lambda^{-(b+1/(2m))}). \tag{4.22}$$

Second and third order derivative analysis. Now we need to analyze the second and third order derivatives. Note since

$$\begin{aligned} D^3 l_n(\theta_*) &= D^3 l(\theta_*) + \{D^3 l_n(\theta_*) - D^3 l(\theta_*)\}, \\ K_{3n}(\lambda, b) &\le K_3(\lambda, b) \\ &\quad + \sup_{\substack{u,\nu \in S(1,\alpha) \\ \theta_1, \theta \in N_{\theta_0}}} \left\| G_\lambda(\theta_\lambda)^{-1}\{D^3 l_n(\theta) u\nu - D^3 l(\theta) u\nu\} \right\|_b. \end{aligned} \tag{4.23}$$

Thus it is only necessary to study the quantities

$$\begin{aligned} D^2 l_n(\theta) u\phi - D^2 l(\theta) u\phi &= \int_0^1 u(t)\phi(t) g(t, \theta(t))\, dH_n(t), \\ D^3 l_n(\theta) u\nu\phi - D^3 l(\theta) u\nu\phi &= \int_0^1 u(t)\nu(t)\phi(t) h(t, \theta(t))\, dH_n(t), \end{aligned} \tag{4.24}$$

where $H_n(t) = A_n(t) - A(t)$ [see equation (4.5)] and $\theta \in N_{\theta_0}$. For the density case these quantities are both zero ($U_n(t) \equiv 0$). In the hazard case, using Kolmogorov's inequality, Sobolev's imbedding theorem and the assumption that $\alpha > 1/(2m)$,

$$\begin{aligned} \{D^2 l_n(\theta) u\phi - D^2 l(\theta) u\phi\}^2 &\le \|u\|_\alpha^2 \|\phi\|_0^2 O_p(n^{-1}), \\ \{D^3 l_n(\theta) u\nu\phi - D^3 l(\theta) u\nu\phi\}^2 &\le \|u\|_\alpha^2 \|\nu\|_\alpha^2 \|\phi\|_0^2 O_p(n^{-1}) \end{aligned} \tag{4.25}$$

and from this using the expansion similar to (4.16) and applying Lemma 2.2 we have

$$\begin{aligned} K_{2n}(\lambda, b) &\le O_p(n^{-1/2})\lambda^{-(b+1/(2m))/2}, \\ K_{3n}(\lambda, b) &\le \lambda^{-(b+1/(2m))/2} + O_p(n^{-1/2})\lambda^{-(b+1/(2m))/2}. \end{aligned}$$

A more elaborate argument is employed in the logistic regression case. For $u, \nu \in S(1, \alpha)$,

$$\begin{aligned} &\left\| G_\lambda(\theta_\lambda)^{-1}\{D^3 l_n(\theta) u\nu - D^3 l(\theta) u\nu\} \right\|_b \\ &\quad = \sum_\nu \left[1 + \lambda_{*\nu}^b\right]\left[1 + \lambda\gamma_{*\nu}\right]^{-2}\{D^3 l_n(\theta) u\nu\phi_{*\nu} - D^3 l(\theta) u\nu\phi_{*\nu}\}^2 \end{aligned}$$

with $\theta_* = \theta_\lambda$. Let $\zeta(t) = u(t)\nu(t)\phi_{*\nu}(t)h(t, \theta(t))$ and expand ζ in terms of the eigenvalues ϕ_ν defined in Section 2.

$$\zeta = \sum_{\nu'} \zeta_{\nu'}\phi_{\nu'}, \tag{4.26}$$

where $\zeta_{\nu'} = \langle \zeta, U\phi_{\nu'} \rangle$. Choose a such that $\alpha \ge a = 1/(2m) + \varepsilon$ where $\varepsilon \in$

$(0, 1/(2m)]$ is arbitrarily small. By the Cauchy–Schwartz inequality,

$$
\begin{aligned}
&\{D^3 l_n(\theta) u v \phi_{*\nu} - D^3 l(\theta) u v \phi_{*\nu}\}^2 \\
&\quad = \left\{ \sum_{\nu'} \zeta_{\nu'} \int_0^1 \phi_{\nu'} \, d(F_n(t) - F(t)) \right\}^2 \\
&\quad \le \sum_{\nu} [1 + \gamma_\nu^a] \zeta_\nu^2 \sum_{\nu'} [1 + \gamma_{\nu'}^a]^{-1} \left\{ \int_0^1 \phi_{\nu'} \, dH_n(t) \right\}^2 \\
&\quad = \|\zeta\|_a^2 \left[\sum_{\nu'} [1 + \gamma_{\nu'}^a]^{-1} \left\{ \int_0^1 \phi_{\nu'} \, d(F_n(t) - F(t)) \right\}^2 \right].
\end{aligned} \tag{4.27}
$$

But since $a > 1/(2m)$ and $\gamma_\nu \approx \nu^{2m}$, by Markov's inequality the stochastic order of the term in square brackets is $O_p(n^{-1})$, because

$$
E\left[\sum_{\nu'} [1 + \gamma_{\nu'}^a]^{-1} \left\{ \int_0^1 \phi_{\nu'} \, d[F_n(t) - F(t)] \right\}^2 \right] \le M n^{-1}, \tag{4.28}
$$

where M is an appropriate positive constant. Note that

$$
\|\zeta\|_a = \|u v \phi_{*\nu} h(\cdot, \theta(\cdot))\|_a \le \|u\|_a \cdot \|\nu\|_a \cdot \|\phi_{*\nu}\|_a \cdot \|h(\cdot, \theta(\cdot))\|_a. \tag{4.29}
$$

This follows by iterating the well-known inequality

$$
\|fg\|_{W_2^s} \le C_s \|f\|_{W_2^s} \cdot \|g\|_{W_2^s}, \tag{4.30}
$$

which is true for $f, g \in W_2^s$ and $s > \frac{1}{2}$, see Lemma X4 of Kato and Ponce (1988) or the proof of Theorem 2.1 of Strichartz [(1967), page 1047]. Note $h(\cdot, \theta(\cdot))$ is clearly in Θ_α by the chain rule since h is infinitely differentiable and $\theta \in \Theta_\alpha$. Also, $\|\phi_{*\nu}\|_a = 1 + \gamma_{*\nu}^a$. Therefore, we have using Lemma 2.2 again that

$$
\begin{aligned}
K_{3n}(\lambda, b) &\le K_3(\lambda, b) + \lambda^{-(b+a+1/(2m))/2} O_p(n^{-1/2}) \\
&= O(\lambda^{-(b+1/(2m))/2}) + \lambda^{-(b+a+1/(2m))/2} O_p(n^{-1/2}).
\end{aligned} \tag{4.31}
$$

An identical analysis for $K_{2n}(\lambda, b)$ yields

$$
K_{2n}(\lambda, b) \le \lambda^{-(b+a+1/(2m))/2} O_p(n^{-1/2}). \tag{4.32}
$$

Combining these results we have $r_n(\lambda, b) = O_p(a_n(\lambda, b))$, where

$$
a_n(\lambda, b)^2 = \begin{cases}
0 + n^{-1}\lambda^{-(\alpha+1/(2m))}\lambda^{-(b+1/(2m))} + 0, \\
\quad n^{-1}\lambda^{-(b+1/(2m))} \\
\quad + n^{-1}\lambda^{-(\alpha+1/(2m))}\{\lambda^{-(b+1/(2m))} + n^{-1}\lambda^{-(b+1/(2m))}\}, \\
\quad n^{-1}\lambda^{-(b+a+1/(2m))} \\
\quad + n^{-1}\lambda^{-(\alpha+1/(2m))}\{\lambda^{-(b+1/(2m))} + n^{-1}\lambda^{-(b+a+1/(2m))}\}.
\end{cases}
$$

On each row in this display the first term comes from the $K_{2n}(\lambda, \alpha)$ term, the second term comes from $K_3(\lambda, b)d(\lambda, \alpha)$ and the third term comes from the

obvious quantity in (4.23). Simplifying

$$(4.33)\quad a_n(\lambda, b)^2 = n^{-1}\lambda^{-2(\alpha+1/(2m))}\lambda^{\alpha-b}\begin{cases}1,\\ \lambda^{\alpha+1/(2m)} + 1 + n^{-1},\\ \lambda^{\alpha+1/(2m)-a} + 1 + n^{-1}\lambda^{-a}.\end{cases}$$

It is clear that if λ_n is a sequence for which $n^{-1}\lambda_n^{-2(\alpha+1/(2m))} \to 0$ for some α in the range $\min(1/(2m), b) < \alpha \le p$, then we can choose α and a such that $1/(2m) < a < \min(\alpha, 1/m)$ and then $r_n(\lambda_n, \alpha) \to_p 0$. Note $\lambda_n \to 0$, so by choosing n sufficiently large we have $\lambda_n \le \lambda_0$, where λ_0 satisfies Theorem 4.1. Thus Assumption A.6 is satisfied and using the convexity of $l_{n\lambda}$, we have the following result.

THEOREM 4.2 (Stochastic error bound). *Let $\lambda_n \le \lambda_0$ be a sequence such that for some $\alpha \in (1/(2m), (p - 1/(2m))/2]$ $n^{-1}\lambda_n^{-2(\alpha+1/(2m))} \to 0$. Then there exists $M > 0$ such that with probability approaching unity as $n \to 0$, $\theta_{n\lambda}$ (with $\lambda = \lambda_n$) is uniquely defined and for $0 \le b \le \alpha$,*

$$(4.34)\quad \|\theta_{n\lambda} - \bar{\theta}_{n\lambda}\|_b^2 \le Mn^{-1}\lambda^{-(b+1/(2m))}\{n^{-1}\lambda^{-2(\alpha+1/(2m))}\},$$

and so for $0 \le b \le \alpha$,

$$(4.35)\quad \|\theta_{n\lambda} - \theta_\lambda\|_b \le Mn^{-1/2}\lambda^{-(b+1/(2m))/2}.$$

Theorem 1 follows immediately from Theorems 4.1 and Theorem 4.2. An optimal upper bound on the rate of convergence is obtained by equating the asymptotic orders of the systematic and stochastic errors. We find that the optimal upper bound on the rate of convergence applies if

$$(4.36)\quad \lambda_n^* = n^{-2m/(2m+1)}$$

and the resulting rate of convergence of the penalized likelihood estimator is

$$(4.37)\quad \|\theta_{n\lambda_n^*} - \theta_0\|_b^2 = O_p(n^{-2m(p-b)/(2m+1)}).$$

The upper bound can be achieved in W_2^m (i.e., λ_n^* satisfies the constraint) provided $m > \frac{3}{2}$.

Acknowledgments. We thank Professors G.. Wahba and T. Leonard for stimulating our interest in this topic. Professors R. J. Beran and C. J. Stone were encouraging on some welcome occasions. We are grateful to Professor M. E. Taylor for directing us to the papers by Kato and Ponce (1988) and Strichartz (1967).

REFERENCES

ADAMS, R. (1975). *Sobolev Spaces*. Academic, New York.

ANDERSON, J. A. and SENTHILSELVAN, A. (1980). Smooth estimates for the hazard function. *J. Roy. Statist. Soc. Ser. B* **42** 322–327.

AUBIN, J. P. (1979). *Applied Functional Analysis*. Wiley, New York.

COX, D. D. (1988). Approximation of method of regularization estimators. *Ann. Statist.* **16** 694–712.

COX, D. D. and O'SULLIVAN, F. (1989). Generalized nonparametric regression via penalized likelihood. Unpublished manuscript.

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

GOOD, I. J. and GASKINS, R. A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

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

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1** 799–821.

KATO, T. and PONCE, G. (1988). Commutator estimates and the Euler and Navier–Stokes equations. *Comm. Pure Appl. Math.* **41** 891–907.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

O'SULLIVAN, F., YANDELL, B. and RAYNOR, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–104.

O'SULLIVAN, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.* **9** 363–379.

O'SULLIVAN, F. (1989). Nonparametric estimation in the Cox proportional hazards model. Unpublished manuscript.

RALL, L. B. (1969). *Computational Solution of Nonlinear Operator Equations*. Wiley, New York.

RUDIN, W. (1976). *Principles of Mathematical Analysis*. McGraw–Hill, New York.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalised likelihood method. *Ann. Statist.* **10** 795–810.

STRICHARTZ, R. (1967). Multipliers in fractional Sobolev spaces. *J. Math. Mech.* **16** 1031–1060.

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

TRIEBEL, H. (1978). *Interpolation Theory, Function Spaces, Differential Operators*. North-Holland, Amsterdam.

WEINBERGER, H. F. (1974). *Variational Methods for Eigenvalue Approximation*. SIAM, Philadelphia.

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

DEPARTMENT OF STATISTICS
101 ILLINI HALL
725 S. WRIGHT STREET
UNIVERSITY OF ILLINOIS
61820 CHAMPAIGN, ILLINOIS

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195