Analysis of Penalized Likelihood-type Estimators with application to Generalized Smoothing in Sobolev Spaces

Dennis Cox

Department of Statistics
University of Illinois

Finbarr O'Sullivan

Department of Statistics
University of California
Berkeley

Technical Report No. 51
October 1985

Research partially supported by the National Science Foundation under Grant No. MCS-820-2560 and MCS-840-3239.

Department of Statistics
University of California
Berkeley, California

Analysis of Penalized Likelihood-type Estimators with application to Generalized Smoothing in Sobolev Spaces

Dennis D. Cox[1]

Department of Statistics
University of Illinois
Urbana. Il 61801.

Finbarr O'Sullivan[2]

Department of Statistics
University of California
Berkeley. CA 94720.

ABSTRACT

A general approach to the analysis of penalized likelihood estimators is described. Focusing on the asymptotics, our main result uses a well known fixed-point theorem to develop an asymptotically valid linear approximation to the roots of penalized likelihood equations. The behavior of the linearized estimators can be conveniently studied in a Hilbert space setting, where there is a rich spectral theory at one's disposal. The theory is well illustrated in the context of generalized smoothing in Sobolev spaces, and here the rates of convergence of some interesting classes of estimators are worked out in detail. The results apply to a broad range of important practical problems including hazard function estimation, density estimation, and the smoothing of regression functions in generalized linear models.

AMS 1980 subject classifications. Primary, 62-G05, Secondary, 62J05, 41-A35, 41-A25, 47-A53, 45-L10, 45-M05.

Key words and phrases. Linearization, Penalty Infornation Scale, Fixed-Point, Banach Space, Spectral Theory, Rates of Convergence.

Running Head: Penalized Liklelihood Estimation.

October 30, 1985

[1] Research partially supported by the National Science Foundation under Grant No. MCS-820-2560.
[2] Research supported by the National Science Foundation under Grant No. MCS-840-3239.

Analysis of Penalized Likelihood-type Estimators with application to Generalized Smoothing in Sobolev Spaces

Dennis D. Cox[1]

Department of Statistics

University of Illinois

Urbana. Il 61801.

Finbarr O'Sullivan[2]

Department of Statistics

University of California

Berkeley. CA 94720.

1. Introduction

1.1. Motivation

Suppose T is an operator defined on a space of probability measures $\mathbf{P}$. To estimate $T(P)$ using data from an unknown element P of $\mathbf{P}$, one can sometimes replace P by an empirical measure $P^{(n)}$ and compute $T(P^{(n)})$. However, this straightforward approach is frequently inapplicable as T may not be defined on the possible empirical probability measures. Furthermore, T may not be continuous in any topology in which the empirical $P^{(n)}$ converges to P, so that the small departure of $P^{(n)}$ from P translates into a large departure of $T(P^{(n)})$ from $T(P)$. A classical example of the former difficulty is density estimation: the empirical distribution does not

[1] Research partially supported by the National Science Foundation under Grant No. MCS-820-2560.
[2] Research supported by the National Science Foundation under Grant No. MCS-840-3239.

possess a (Lebesgue) density. Nonparametric regression exemplifies both problems. Suppose (X, Y) have a jointly continuous distribution, and we wish to estimate $\mathrm{E}[Y \mid X = x_0]$. The corresponding empirical quantity is not even defined unless x_0 is an observed value of X (which occurs with probability zero), and even if x_0 is observed, the corresponding estimate is not consistent. In many scientific problems, the quantity of interest is only indirectly related to the data, and this can compound matters further. For instance, in retrieval of temperature profiles from satellite data (see O'Sullivan and Wahba[20]) one is interested in solving an integral equation

$$R(x) = \int k(x, \pi, \theta(\pi))\, d\pi$$

where $R(x)$ is upwelling radiation at frequency x, $\theta(\pi)$ is the temperature at pressure level π, and k is a nonlinear function. There are available only "discrete noisy data" $(x_1, Y_1), \cdots, (x_n, Y_n)$ where

$$R(x_i) = \mathrm{E}[Y_i \mid X = x_i], \quad 1 \leq i \leq n \quad .$$

Even under the best of conditions of perfect observations, this problem is difficult to solve because of instabilities that result from numerical errors. The addition of discretization and random noise makes matters even worse.

This lack of stability in the evaluation of T is a form of "ill-posedness", a notion introduced by Tikhonov[27]. Over the last fifty years applied mathematicians have come to appreciate that a vast collection of the problems encountered by engineers and scientists in such areas as Geophysics[2, 6, 7, 4], Meteorology[25, 24, 30] and Tomography[5, 29] are ill-posed in this sense. Given the fact that real data are nearly always subject to random variation (not just rounding error) it seems natural that statisticians should play a more active role in these problems.

An early technique for obtaining approximate solutions to "ill-posed" problems of the above type was proposed by Tikhonov[26]. Let θ be the parameter of interest. Tikhonov's method, which is known as regularization, has two components: a functional l which measures how well θ predicts (matches) the observed data and a functional J which assesses the physical plausibility of θ. These functionals are set up so that smaller values generally correspond to more desirable

values for θ. Given these functionals the method of regularization (MOR) chooses a parameter value θ which minimizes an aggregate

$$l(\theta \mid \text{data}) + \lambda J(\theta) \qquad \lambda > 0.$$

Good and Gaskins[13] introduced the method of regularization to statisticians under the name of "penalized likelihood estimation". The functional l which measures quality of predictions is the "likelihood" term and J is the prior penalty or roughness component. Such estimators have also been proposed in a Bayesian context by Leonard[15, 16].

The purpose of this paper is to provide a framework in which it will be possible to understand the asymptotic behavior of penalized likelihood or method of regularization estimators from a statistical point of view. Our theory is in the tradition of Cramer's analysis of the method of maximum likelihood. This theory is illustrated in the context of generalized smoothing. We begin by giving a more precise mathematical specification of the penalized likelihood method.

1.2. The Penalized Likelihood Method

Estimation Methodology

We consider three types of measurement models corresponding to density, hazard, and regression function estimation. In each case, the model is at least partially parameterized by the corresponding function, which is taken to be an unknown element of a Banach space Θ with norm $\|\cdot\|$. Fix the unknown true parameter θ_0 and suppose $(Z_n : n = 1,2, \cdots)$ denotes the observations in a sequence of statistical experiments. Then a penalized likelihood type estimator is obtained by minimization over Θ of

$$l_{n\lambda}(\theta) = l_n(Z_n, \theta) + \lambda J(\theta) \ .$$

Here, $l_n(Z_n, \theta)$ is referred to as the likelihood term. This could be a negative log likelihood, in general the only requirement is that smaller values of $l_n(\theta)$ correspond to "models", θ, which better fit the observed Z_n. $J(\theta)$ is called the penalty functional ($J : \Theta \to \mathbb{R}^+$), and smaller values of J correspond to more plausible values of θ, or, to a Bayesian, values of θ with higher

prior likelihood. The smoothing parameter λ controls the tradeoff of sample versus prior information.

We next consider the specification of the likelihood term for three types of observational models, and afterwards give examples of the penalty functionals.

Measurement Model and Likelihoods

(a) **Density Estimation**

We observe a random sample $X_1, X_2, \cdots, X_n$ from a density f, where $f = \log \theta_0$. Following Silverman[19], the "likelihood" component of the penalized likelihood is

$$\int_{\mathcal{X}} e^{\theta(x)} dx - \int_{\mathcal{X}} \theta(x) P_X^{(n)}(dx) \quad .$$

where $P_X^{(n)}$ is the empirical distribution of $X_1, X_2, \cdots, X_n$. It is necessary to assume $J(\theta) = 0$ whenever θ is a constant function (Theorem 3.1,[19]).

(b) **Hazard Estimation**

Assume $X_1, X_2, \cdots X_n$ are positive random variables and we observe $\min(X_i, 1)$, the minimum of X_i and 1. This corresponds to a survival experiment with a set time on the length of the experiment. The target parameter is $\theta_0 = \log\lambda$ where $\lambda = \frac{f}{1-F}$ is the hazard function and F is the cumulative distribution function. Partially following Anderson and Senthilselvan[3], the "likelihood" component of the penalized likelihood is given by

$$\int e^{\theta(x)} S_n(x) dx - \int \theta(x) P_X^{(n)}(dx)$$

where S_n is the empirical survival function $(1-F_n)$. The limiting survival function is denoted $S = 1-F$.

(c) **Regression Models**

One observes a sample of n random pairs $(X_{n1}, Y_{n1}), (X_{n2}, Y_{n2}), \cdots, (X_{nn}, Y_{nn})$ where the X_{ni}'s are thought of as covariates and the Y_{ni}'s as responses. The conditional distribution of Y given $X = x$ is denoted

$$Law(Y \mid X=x) = P_{Y|X}(\cdot \mid x) \quad .$$

The covariates $X_{n1}, \cdots, X_{nn}$ need not be truly random, *i.e.* they may be degenerate random elements. In fact, we shall treat them as non-random for the regression model and write x_{ni} for the observed value of X_{ni}. Inferences will proceed conditionally on the observed values of the X_{ni}. If they were random, say *i.i.d*, then since the distribution of X_{ni}

$$Law\ (X_{ni}) = P_X$$

is not of interest (*i.e.* is a nuisance parameter), such conditional inference is reasonable and is justified by the principle of ancillarity, as indicated on pp.33-35 of Cox and Hinkley[10].

Let $P_{XY}^{(n)}$ denote the joint empirical measure of the (x_{ni}, Y_{ni}), *i.e.*

$$P_{XY}^{(n)}(B \times A) = \frac{1}{n}\sum_{i=1}^{n} I_A(Y_{ni})\, I_B(x_{ni}) \quad , \quad A \subseteq \underline{Y} \ , \quad B \subseteq \underline{X} \ ,$$

where I_A denotes the indicator function of the set A, and $\underline{Y}$ ($\underline{X}$) denotes the range of Y_{ni} (X_{ni}). Similarly let

$$P_X^{(n)}(B) = \frac{1}{n}\sum_{i=1}^{n} I_B(x_{ni}) \quad , \quad B \subseteq \underline{X} \ ,$$

denote the marginal empirical of the x_{ni}'s. We will assume $P_X^{(n)}$ approaches a fixed limiting design measure, P_X. Then there is a joint measure P_{XY} with marginal P_X and conditional $P_{Y|X}$. When $(X_{n1}, Y_{n1}), \cdots, (X_{nn}, Y_{nn})$ are *i.i.d.* random pairs from P_{XY}, we shall refer to the *random design model* (RDM). However, the asymptotic theory works better if the x_{ni} are more uniformly distributed than would be obtained from the RDM.

Suppose that the conditional distribution $P_{Y|X}(\cdot \mid x)$ is "partially" specified by a q-dimensional vector $\theta(x)$, and Θ is a space of q-dimensional functions in $\underline{X}$. The likelihood for a single observation at x is $l(y, x, \theta) = \rho(y \mid x, \theta(x))$ The penalized likelihood is

$$l_{n\lambda}(\theta) = \int \rho(y \mid x, \theta(x)) P_{XY}^{(n)}(dxy) + \lambda J(\theta) \quad .$$

The true parameter is defined pointwise at x as the minimizer over $t \in R^q$ of

$$\int \rho(y \mid x, t) P_{Y|X}(dy \mid x) \ .$$

Some choices for l are discussed next.

(i) Consider the normal additive error model:

$$y_{ni} = \mu(x_{ni}) + \epsilon_{ni} \quad ,where$$

where the ϵ's are assumed to be *i.i.d.* normal random random variables with mean zero and constant variance σ^2. Then the natural choices for the partial parameter and the likelihood are

$$\theta(x) = \mu(x) \quad ; \quad l(y,x,\theta) = [y-\theta(x)]^2 \quad .$$

Note that the normality assumption is not really needed. This choice of l is appropriate whenever one wishes to estimate $\theta_0(x) = E[Y \mid X=x]$.

(ii) If in (i) the variance $\sigma^2(x)$ is non-constant but depends smoothly on x as well, then natural choices would be

$$\theta(x) = (\mu(x), -\log\sigma(x))$$
$$l(y,x,\theta) = -\theta_2(x) + \frac{1}{2}e^{2\theta_2(x)}[y-\theta_1(x)]^2 \quad .$$

Here we have chosen the parameterization $\theta_2(x) = -\log\sigma(x)$ to avoid awkward positivity constraints.

(iii) If the errors in (i) are no longer assumed normal but to have density f, then one would naturally use

$$l(y,x,\theta) = -\log f\,[y-\theta(x)]$$

One may wish to replace $-\log f$ by a function ρ as in robust estimation of location; consult Huber[14] for further details. Again one may incorporate scale estimation as well as was done in (ii).

(iv) If the response Y is binary (zero or one) with "pointwise" success probability

$$p(x) = P[Y=1 \mid X=x] \quad ,$$

then a natural choice is

$$\theta(x) = \log[p(x)/(1-p(x))] \quad ,$$

$$l(y,x,\theta) = -y\,\theta(x) + \log(1+e^{\theta(x)}) \quad .$$

The estimator will then be a non-parametric logistic regression estimator. Note that the logit transformation leads to an unconstrained parameter. A probit approach is also acceptable.

(v) If Y is Poisson with mean $\lambda(x)$, then a nonparametric log-linear Poisson regression estimator is obtained by setting

$$\theta(x) = \log\lambda(x)$$
$$l(y,x,\theta) = -y\,\theta(x) + e^{\theta(x)}$$

This and the previous example are both special cases of the generalized linear model as described in Nelder and McCullagh[18]. Any of the cases presented there may be treated in a similar manner.

Parameter Spaces and Penalty Functionals

The penalty functional $J(\theta)$ is often chosen to penalize for "roughness". Suppose θ is q-dimensional valued, let $\alpha = (\alpha_1, \cdots, \alpha_d)$ be a multi-index (d-vector with non-negative integer co-ordinates), and let

$$D_x^{\alpha} = \prod_{j=1}^{d} \frac{\partial^{\alpha_j}}{\partial x_j^{\alpha_j}}$$

be a partial differential operator of order

$$|\alpha| = \sum_{j=1}^{d} \alpha_j \quad .$$

Let $\mathcal{X}$ be a bounded domain in $\mathbb{R}^d$. The *Sobolev space* $W_2^p(\mathcal{X}\,;\mathbb{R}^q)$ is the collection of $\mathbb{R}^q$ valued generalized functions having derivatives of orders $\leq p$ whose components are in $L_2(\mathcal{X})$. The norm is

$$\|\theta \mid W_2^p(\bar{\mathcal{X}};\mathbb{R}^q)\| = \{ \sum_{\substack{|\alpha| \leq p \\ 1 \leq j \leq q}} \|D_x^{\alpha}\theta_j(x) \mid L_2(\mathcal{X})\|^2 \}^{\frac{1}{2}} \quad .$$

One can define W_2^p for any real number $p \geq 0$ using *e.g.* interpolation theory (see Adams[1],

Cox[8], or Triebel[28]), and in all that follows such fractional order Sobolev spaces are allowed, unless explicitly excluded, although the reader may wish to think of p as an integer for simplicity. The spaces W_2^p are Hilbert spaces when equipped with the inner product (for p an integer)

$$<\theta,\xi \mid W_2^p> = \sum_{\substack{|\alpha| \le p \\ 1 \le j \le q}} \int_X [D_x^\alpha \theta_j(x)]\,[D_x^\alpha \xi_j(x)]\,dx$$

If $\Theta = W_2^m(\underline{X}\ ;\ \boldsymbol{R}^q)$, then a natural roughness penalty is

$$J(\theta) = \sum_{j=1}^{q} \sum_{|\alpha|=m} \int_X [D_x^\alpha \theta_j(x)]^2 dx \quad , \tag{1.1}$$

where θ_i is the i'th component of θ and the order m is prechosen. When $q = 1$, this corresponds to the roughness penalty of Cox[9]. The extension to $q > 1$ presents no difficulty. When $d = 1$, it is irrelevant whether one takes the integral over $\underline{X}$ (which is assumed to be a bounded set) or over all of $\boldsymbol{R}$ as the solution of the minimization problem is the same. For multivariate $\underline{X}$, integration over all of $\boldsymbol{R}^d$ in the penalty is also possible, but we have not yet been able to adapt the asymptotic theory to this situation.

In general when Θ is Hilbertian, it is natural to consider penalty functionals of the form

$$J(\theta) = <\theta, W\theta>$$

where W is a positive operator, see Cox[8]. The previously displayed penalty on W_2^m can be put in this form.

1.3. Main Asymptotic Results

Our theory relates to the large sample behavior of roots of the penalized likelihood equations. That is we look at the *score operator*, $Z_{n\lambda}$

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

and discuss the properties of roots of the equations, $Z_{n\lambda} = 0$, as $n \to \infty$. The limiting version of the score functional is also of interest. This is defined in the regression case as

$$Z_\lambda(\theta) = \int Dl(y,x,\theta) P_{XY}(dxdy) + \lambda DJ(\theta) \quad .$$

with analogous expressions for density and hazard estimation.

(Note that both Z_λ and $Z_{n\lambda}$ are mappings from Θ into the dual space Θ^*). The first step in the analysis is linearization. Conditions are given under which Z_λ and $Z_{n\lambda}$ have locally unique roots and it is described how such roots can be approximated by simpler linearized "estimates". The linearized estimates are defined in terms of the Hessian of the "continuous" penalized likelihood. It may be helpful to think of this as a generalized "information" operator. For $\theta \in \Theta$, the Hessian $G_\lambda(\theta)$ of the penalized likelihood at θ is

$$G_\lambda(\theta) = \int D^2 l(y,x,\theta) P_{XY}(dxdy) + \lambda D^2 J(\theta)$$

The first is a *continuous* linearization. It says that for all λ sufficiently small there is a unique root, θ_λ, of $Z_\lambda = 0$ in a neighborhood of θ_0. Moreover, if $d(\theta_\lambda,\theta_0)$ measures the distance between θ_λ and θ_0, then we give conditions under which

$$d(\theta_\lambda,\theta_0) = d(\bar{\theta}_\lambda,\theta_0)\,(1 + o(1))$$

where $\bar{\theta}_\lambda$, the linearized "estimate", is obtain by a one step linearization of $Z_\lambda(\cdot)$ about the true value θ_0

$$\bar{\theta}_\lambda = \theta_0 - G_\lambda^{-1}(\theta_0) Z_\lambda(\theta_0) \tag{1.2}$$

There is a corresponding *discrete* linearization result. It can be shown that for all $\lambda \in [\lambda_n, \lambda_0]$, λ_0 sufficiently small, and for all n sufficiently large, with arbitrarily high probability there is a unique solution, $\hat{\theta}_{n\lambda}$, to $Z_{n\lambda} = 0$, in a neighborhood of θ_λ satisfying

$$d(\hat{\theta}_{n\lambda},\theta_\lambda) = d(\bar{\theta}_{n\lambda},\theta_\lambda)\,(1 + o_p(1))$$

where the linearized "estimate", $\bar{\theta}_{n\lambda}$ is now given by

$$\bar{\theta}_{n\lambda} = \theta_\lambda - G_\lambda^{-1}(\theta_\lambda) Z_{n\lambda}(\theta_\lambda) \quad . \tag{1.3}$$

Since

$$\hat{\theta}_{n\lambda} - \theta_0 = (\hat{\theta}_{n\lambda} - \theta_\lambda) + (\theta_\lambda - \theta_0) \doteq (\hat{\theta}_{n\lambda} - \theta_\lambda) + (\bar{\theta}_\lambda - \theta_0) \quad ,$$

the linearizations allow one understand the asymptotic behavior of roots of the penalized

likelihood equations by studying the the asymptotic behavior of much simpler linear "estimates". The continuous linearization provides information on the asymptotic bias of the estimator while the discrete linearization gives information on its asymptotic variability. The linearization theorems are presented in section 4. In a Hilbert space setting, where there is a rich spectral theory, the properties of the linearized "estimates" can be analyzed in detail, and in section 5 this aspect is worked out for a class of penalized likelihoods described in section 2.

1.4. Some Comments

Although the theory in section 4 implies the asymptotic existence of roots of the penalized likelihood equations, it does not say what can happen in finite samples. Along these lines, a few existence results are given in section 3.

The computational aspects of penalized likelihood are not discussed even though this surely warrants some mention. Often in practical situations it is quite feasible to implement Newton-type minimization algorithms for this purpose, see Cooley[6], and Neuman and Yakowitz[19]. However, further work needs to be done in this area before we can get a good understanding of the issues and subtleties involved.

The choice of the smoothing parameter λ is not discussed here either. It seems that in some situations it may be possible to come up with suitable versions of cross validation or unbiased risk estimates, see O'Sullivan et. al.[21] and O'Sullivan and Wahba[20]. However, a proper asymptotic analysis of this problem is beyond the scope of the present paper.

1.5. Acknowledgment

The authors are indebted to Professors T. Leonard and G. Wahba for bringing this class of estimators to their attention.

2. Results for Generalized Smoothing in Sobolev Spaces

In this section, we state the assumptions and asymptotics for the regression, density and hazard function estimation problems when Θ is a Sobolev space, and the penalty has the form $J(\theta) = <\theta, W\theta>$.

2.1. Assumptions

Any assumption on the x's applies to all three observational models, but assumptions about y's only pertain to the regression model.

Assumption A. (i) $\{x_{n1}, x_{n2}, \cdots, x_{nn}\} \subseteq \underline{X} \subseteq \mathbb{R}^d$, *where* $\underline{X}$ *is a bounded, open, simply connected, nonempty set with* C^∞*-boundary (Definition 3.2.1.2 of Triebel[28]).*

(ii) $Y_{n1}, Y_{n2}, \cdots, Y_{nn}$ *are random elements taking values in some measurable space* $\underline{Y}$*, and the joint conditional distribution of* $Y_{n1}, Y_{n2}, \cdots, Y_{nn}$ *given* $X_{n1}=x_{n1}, X_{n2}=x_{n2}, \cdots, X_{nn}=x_{nn}$ *factors as the product of marginal conditional distributions of* Y_{ni} *given* $X_{ni}=x_{ni}$*, i.e.*

$$Law\ (Y_{n1}, Y_{n2}, \cdots, Y_{nn} \mid X_{n1}=x_{n1}, X_{n2}=x_{n2}, \cdots, X_{nn}=x_{nn}) = \prod_{i=1}^{n} P_{Y \mid X}(\cdot \mid x_{ni}).$$

Assumption B (i) *There is a probability measure* P_X *on* $\underline{X}$ *such that if* F_n *and* F *denote the distribution functions of* $P_X^{(n)}$ *and* P_X*, respectively, then*

$$k_n = \sup_{x \in \underline{X}} |F_n(x) - F(x)| \to 0 \qquad \text{as} \quad n \to \infty.$$

(ii) *For the density and regression problems,* P_X *has a density* $f(x)$ *which satisfies*

$$0 < K_1 \leq f(x) \leq K_2 < \infty \qquad , \text{ for all } x \in \underline{X}\ , \tag{2.1}$$

for some constants K_1 *and* K_2*. For the hazard function problem, (2.1) holds for* $x \in \underline{X} = [0,1]$ *and* $S(1) > 0$. When $X_{n1}, \cdots, X_{nn}$ are *i.i.d.* as is assumed for density, hazard and the RDM regression models, k_n is random and $k_n = O_p(n^{-\frac{1}{2}})$. In regression with designed x's, the best one can achieve is $k_n = O_p(n^{-1}(\log n)^{(d-1)/2})$. See Davis and Rabinowitz[12], p. 268 ff.

Assumption C. (i) Θ *is a Hilbert space of functions* $\theta : \underline{X} \to \mathbb{R}^q$ *with inner product* $<\cdot,\cdot>$ *and norm* $\|\cdot\|$.

(ii) *For some* $m > 3d/2$, $\Theta = W_2^m(\underline{X}\ ;\ \boldsymbol{R}^q)$, *as sets and they have equivalent norms.*

(iii) *The penalty functional* $J(\theta) = <\theta, W\theta>$ *where* W *is a bounded linear operator on* Θ *which is self adjoint and nonnegative definite.*

(iv) *For some* K_1, K_2,

$$K_1||\theta||^2 \leq <\theta, W\theta> + ||\theta \mid L_2(\underline{X}\ ;\ \boldsymbol{R}^q)||^2 \leq K_2||\theta||^2\ ,$$

for all $\theta \in \Theta$.

(v) *The true function parameter* θ_0 *is in* W_2^s *for some* $s > 3d/2$.

Some further assumptions needed for the regression model, but first some notation. For any normed linear spaces A and B, let $B(A,B)$ denote the class of continuous linear operators equipped with the usual operator norm

$$||T \mid B(A,B)|| = \sup\{||Ta \mid B|| : a \in S(1,A)\}$$

where for $R > 0$

$$S(R,A) = \{\ a \in A : ||a \mid A|| \leq R\ \}$$

is the closed, centered ball of radius R in A.

Recall that in the regression problem, the "likelihood" is determined by an M-estimation functional $\rho : \underline{Y} \times \underline{X} \times \boldsymbol{R}^q \to \boldsymbol{R}^+$. Let $\psi : \underline{Y} \times \underline{X} \times \boldsymbol{R}^q \to \boldsymbol{R}^q$ be the gradient, $\dot{\rho}$, of $\rho(y \mid x,t)$ *w.r.t.* the variable t. The dot $\dot{}$ will be used to denote differentiation *w.r.t.* the variable t. A set of assumptions regarding ψ need to be specified. These "regularity conditions" are similar to (but stronger than) those given in Cramer[11]. Loosely speaking, these conditions hold provided $\rho(y \mid x,t)$ is "sufficiently smooth as a function of x and t with derivatives satisfying moment conditions." In what follows, we shall use K, K_1, K_2, $\cdots$ to denote positive finite constants which may depend on p, d, $\underline{X}$, q, and P_X, and are not necessarily the same in each appearance. Dependence on other variables in the problem will be explicitly indicated by their inclusion in parentheses after the constant. Global constants, which are the same in each appearance, will be denoted by $M_1, M_2, \cdots$, with the same rules for dependence on the quantities at hand. Let

$$\bar{\psi}(x,t) = \int_{\underline{Y}} \psi(y \mid x,t) P_{Y \mid X}(dy \mid x)\ .$$

Assumption D (i) *For all* $x \in \underline{X}$, *and for all* $t \in \mathbb{R}^q$,

$$\int ||\psi(y \mid x,t) \mid \mathbb{R}^q||^2 P_{Y \mid X}(dy \mid x) = M_1(x,t) < \infty$$

(ii) *Let*

$$\begin{aligned}\kappa(x,t_1,t_2) &= \mathrm{Cov}[\psi(Y \mid x,t_1),\psi(Y \mid x,t_2) \mid X=x] \\ &= \int [\psi(y \mid x,t_1) - \bar{\psi}(x,t_1)]\,[\psi(y \mid x,t_2) - \bar{\psi}(x,t_2)]'\; P_{Y \mid X}(dy \mid x) \quad ,\end{aligned}$$

then for all $R > 0$, *the below indicated restriction of* κ *satisfies*

$$\kappa \in C^d(\bar{\underline{X}} \times S(R,\mathbb{R}^q) \times S(R,\mathbb{R}^q)\,;\, \mathbb{R}^{q \times q}).$$

(iii) *For all* $u \in S(1,\mathbb{R}^q)$, $x \in \underline{X}$, $t_1,t_2 \in S(R,\mathbb{R}^q)$,

$$0 < M_2(R) \leq u'\,\kappa(x,t_1,t_2)u \leq M_3(R) < \infty$$

(iv) *For all* $x \in \underline{X}$, $\dot{\psi}(y \mid x,t)$ *exists* $P_{Y \mid X}(\cdot \mid x)$ *almost surely for all* $t \in \mathbb{R}^q$ *and satisfies*

$$\int \dot{\psi}(y \mid x,t) P_{Y \mid X}(dy \mid x) = \dot{\bar{\psi}}(x,t)$$

(v) *For all* $x \in \underline{X}$, *for all* $R > 0$, *for all* $t \in S(R,\mathbb{R}^q)$, *and for all* $u \in S(1,\mathbb{R}^q)$,

$$M_4(R) \leq u'\,\dot{\bar{\psi}}(x,t)u \leq M_5(R)$$

(vi) *Let*

$$\tau(x,t) = \mathrm{Cov}[\dot{\psi}(Y \mid x,t) \mid X=x] \quad ,$$

$$\tau_{jj',kk'}(x,t) = \int [\dot{\psi}_{jj'}(y \mid x,t) - \dot{\bar{\psi}}_{jj'}(x,t)]\,[\dot{\psi}_{kk'}(y \mid x,t) - \dot{\bar{\psi}}_{kk'}(x,t)]\; P_{Y \mid X}(dy \mid x) \quad .$$

Then for all $R > 0$, τ *satisfies*

$$\tau \in C^d(\bar{\underline{X}} \times S(R,\mathbb{R}^q)\,;\, \mathbb{R}^{q \times q \times q}).$$

(vii) *For all* $u,v \in S(1,\mathbb{R}^q)$, $x \in \underline{X}$, *and* $t \in S(R,\mathbb{R}^q)$,

$$\sum_{j,k,j',k'=1}^{q} u_j\,u_{j'}\,\tau_{jj',kk'}(x,t)v_k\,v_{k'} \leq M_6(R) < \infty$$

(viii) *For all* $x \in \underline{X}$, $\ddot{\psi}(y \mid x,t)$ *exists* $P_{Y \mid X}(\cdot \mid x)$*-a.s. for all* $t \in \mathbb{R}^q$ *and satisfies*

$$\int \sup_{t \in S(R,\mathbb{R}^q)} ||\ddot{\psi}(y \mid x,t) \mid \mathbb{R}^{q \times q \times q}|| P_{Y \mid X}(dy \mid x) \leq M_7(R) \quad , \quad \textit{for all } x.$$

(ix) *For all* $R > 0$, *the below indicated restriction of* $\bar{\psi}$ *satisfies*

$$\bar{\psi} \in C^d(\underline{X} \times S(R,\mathbb{R}^q) ; \mathbb{R}^q) \quad .$$

Remarks. Only D(ii), D(vii) and D(ix) represent strong departures from Cramer's assumptions. These are used to deal with some of the problems that arise from the infinite dimensional parameter. Note that all boundedness requirements on ρ or ψ only involve the point-wise parameter t in bounded subsets of $\mathbb{R}^q$.

2.2. Derivatives

In this section, we give representations for the derivatives of the penalized likelihood $l_{n\lambda}$ and related functionals. Only directional (Gateaux) derivatives are used. One of the most useful facts from the theory of function spaces is **Sobolev's Imbedding Theorem :** $W_2^p(\bar{X};\mathbb{R}^q)$ *is a subset of* $C^r(\bar{X};\mathbb{R}^q)$ *if* $p > r + d/2$, *and the injection mapping is in* $B(W_2^p, C^r)$, *i.e.*

$$||\theta \mid C^r|| \leq K \; ||\theta \mid W_2^p||$$

for some constant K, $0 < K < \infty$. Under these conditions, we say W_2^p is *continuously imbedded* in C^r and write

$$W_2^p \subseteq C^r \quad .$$

See Theorem 5.4 of Adams[1] (noting that our assumption of C^∞ boundary in A (i) implies any of the cone conditions), or Theorem 4.6.1 (e) of Triebel[28] (note that W_2^s is equal to B_{22}^s by Remark 2.3.3/4 and Definition 4.2.1 of[28]).

Remarks. (a) From C(ii) and Sobolev's imbedding theorem, we have

$$\Theta \subseteq C^d \tag{2.2}$$

It follows from this that each of the real valued maps $\theta \to \theta_j(x)$ (evaluation of the j'th component of $\theta \in \Theta$ at $x \in \underline{X}$) is a bounded linear functional on Θ. By the Riesz representation theorem and C(i), there is for each j, $1 \leq j \leq q$, and each $x \in \underline{X}$ an element $\xi_j(x) \in \Theta$ such that

$$\theta_j(x) = <\theta, \xi_j(x)> \quad , \quad \textit{for all } \theta \in \Theta \ .$$

When $q = 1$, this is equivalent to the property of being a reproducing kernel Hilbert space. We will write

$$\xi(x) = (\xi_1(x), \cdots, \xi_q(x))'$$

as a column vector with components in Θ, for each $x \in \mathcal{X}$. The transpose is denoted $\xi'(x)$. This notation is merely for algebraic convenience.

(b) The penalty functional, $J(\theta) = <\theta, W\theta>$, could be given by

$$J(\theta) = \int_{\mathcal{X}} ||L\,\theta(x) \mid \mathbb{R}^q||^2 dx$$

where $L : W_2^m(\mathcal{X}; \mathbb{R}^q) \to L_2(\mathcal{X}; \mathbb{R}^q)$ is a system of q linear differential operators of order m, in which case $W = L^*L$ is a boundary value operator of order $2m$ obtained from Green's formula. See Proposition 2.2(ii) of Cox[9].

(c) The true parameter in the log density and hazard estimation problems is the underlying log density or hazard of the X_i's. Thus the data are assumed to derive from a model determined by θ_0. However for the regression model, the "true" parameter θ_0 is determined by ρ and $P_{Y \mid X}$ as indicated above. If ρ is obtained by taking the negative log of a (point-wise) likelihood, then we are not assuming that $P_{Y \mid X}(\cdot \mid x)$ is in the given parametric model. In this case, $\theta_0(x)$ is the value of the point-wise parameter which minimizes Kullbak-Leibler "distance" between $P_{Y \mid X}(\cdot \mid x)$ and the model.

(d) We do not require that $\theta_0 \in \Theta$. Indeed, Θ is dense in many function spaces with weaker norms, so our estimates (which are in Θ) can converge to something which is not as smooth as the elements of Θ.

The penalized likelihood type estimator, $\hat{\theta}_{n\lambda}$, is obtained by minimization of the penalized likelihood functional $l_{n\lambda}$ over $\theta \in \Theta$. Since Θ is a Hilbert space, it is isomorphic to Θ^* in a canonical way, and we identify Θ with Θ^* for purposes of calculating $Z_{n\lambda}$. The forms of $Z_{n\lambda}$ for the examples we discuss are as follows:

Log Density: $Z_{n\lambda}(\theta) = \int \xi(x) e^{\theta(x)} dx - \int \xi(x) P_X^{(n)}(dx) + 2\lambda W\theta$

Log Hazard: $Z_{n\lambda}(\theta) = \int \xi(x) e^{\theta(x)} S_n(x) dx - \int \xi(x) P_X^{(n)}(dx) + 2\lambda W\theta$

Generalized Linear Regression: $Z_{n\lambda}(\theta) = \int \xi(x) \psi(y \mid x, \theta(x)) P_{XY}^{(n)}(dxdy) + 2\lambda W\theta$ (2.2)

2.3. Asymptotic Notation

The following asymptotic notation will be used. If f and g are real valued functions on a metric space U and $u_0 \in U$, then

$$f(u) \lesssim g(u) \qquad \text{as} \quad u \to u_0$$

means for some K and some neighborhood N of u_0,

$$\left| \frac{f(u)}{g(u)} \right| \leq K \quad , \quad \text{for all } u \in N,$$

where the numerator is required to be zero whenever the denominator is zero. If there is an additional variable v and $V(u)$ is a set of values of v for each u, then

$$f(u,v) \lesssim g(u,v) \qquad \text{as} \quad u \to u_0$$

uniformly in $v \in V(u)$ means

$$\sup_{v \in V(u)} \left\{ \frac{f(u,v)}{g(u,v)} \right\} \leq 1 \qquad \text{as} \quad u \to u_0 \quad .$$

If $f(u,v,\omega)$ and $g(u,v,\omega)$ are random variables on a probability space Ω, then the above means that for all $\epsilon > 0$ $\exists K \in (0,\infty)$ and $\exists N$, a neighborhood of u_0, such that

$$P\{ \omega \in \Omega : \sup_{v \in V(u)} \left| \frac{f(u,v,\omega)}{g(u,v,\omega)} \right| > K \} < \epsilon \quad , \quad \text{for all } u \in N \quad .$$

The notation

$$f(u) \approx g(u) \qquad \text{as} \quad u \to u_0$$

means $f(u) \lesssim g(u)$ and $g(u) \lesssim f(u)$.

Finally,

$$f(u,v) << g(u,v) \qquad \text{as} \quad u \to u_0 \quad ,$$

uniformly in $v \in V(u)$, is taken to mean that for all $K > 0$

$$P\left\{ \omega \in \Omega : \sup_{v \in V(u)} \left| \frac{f(u,v,\omega)}{g(u,v,\omega)} \right| > K \right\} \to 0$$

as $u \to u_0$.

2.4. Linearized "Estimates" and a Summary of Asymptotic Results

The continuous versions of the $Z_{n\lambda}$'s become important in the asymptotic analysis. For the examples we discuss these are given by

$$\text{Log Density:} \quad Z_\lambda(\theta) = \int \xi(x) e^{\theta(x)} dx - \int \xi(x) P_X(dx) + 2\lambda W\theta$$

$$\text{Log Hazard:} \quad Z_\lambda(\theta) = \int \xi(x) e^{\theta(x)} S(x) dx - \int \xi(x) P_X(dx) + 2\lambda W\theta$$

$$\text{Generalized Linear Regression:} \quad Z_\lambda(\theta) = \int \xi(x) \bar{\psi}(x, \theta(x)) P_X(dx) + 2\lambda W\theta \qquad (2.4)$$

Note that $Z_\lambda(\theta) = EZ_{n\lambda}(\theta)$ under the RDM, and more generally, $Z_{n\lambda}(\theta) \to Z_\lambda(\theta)$ as $n \to \infty$ for fixed λ, θ. Hence, for large samples, $\hat{\theta}_{n\lambda}$ should be approximately unbiased for θ_λ where

$$Z_\lambda(\theta_\lambda) = 0$$

The existence of θ_λ (for λ sufficiently small) is established in section 4. Our approximant, $\bar{\theta}_{n\lambda}$, to $\hat{\theta}_{n\lambda}$ is obtained by linearization of $Z_{n\lambda}(\theta)$ about θ_λ, *i.e.* by setting

$$Z_{n\lambda}(\theta_\lambda) + G_\lambda(\theta_\lambda)(\theta_\lambda - \theta) = 0$$

where

$$G_\lambda(\theta) = DZ_\lambda(\theta) = U(\theta) + 2\lambda W \quad ,$$

$G_\lambda : \Theta \to B(\Theta,\Theta)$ and $U(\theta) \in B(\Theta,\Theta)$ are given by

$$\text{Log Density:} \quad U(\theta)\varsigma = \int \xi(x) e^{\theta(x)} \varsigma(x) dx$$

$$\text{Log Hazard:} \quad U(\theta)\varsigma = \int \xi(x)e^{\theta(x)}\varsigma(x)S(x)dx$$

$$\text{Generalized Linear Regression:} \quad U(\theta)\varsigma = \int \xi(x)\dot{\psi}(x,\theta(x))\varsigma(x)P_X(dx) \tag{2.5}$$

for all $\varsigma \in \Theta$. Thus, the linear approximant is given by

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

Of course in order to study convergence it is necessary to define "closeness". It is natural to use norms for this purpose, and there is a convenient parameterized family of norms (and associated Hilbert spaces) which is determined by the structure of the problem. For the U operators in (2.5), Section 1 of the Appendix shows that for each λ sufficiently small there are sequences $\{\phi_{\lambda\nu} : \nu=1,2,\cdots\}$ of eigenfunctions and $\{\gamma_{\lambda\nu}\}$ of eigenvalues which satisfy

$$<\phi_{\lambda\nu}, U(\theta_\lambda)\phi_{\lambda\mu}> = \delta_{\nu\mu} \tag{2.6}$$

$$<\phi_{\lambda\nu}, W\phi_{\lambda\mu}> = \gamma_{\lambda\nu}\,\delta_{\nu\mu}$$

for all pairs ν,μ of positive integers, where $\delta_{\nu\mu}$ is Kronecker's delta. For $b \geq 0$ let

$$||\theta||_{\lambda b} = \{\sum_{\nu=1}^{\infty}(1+\gamma_{\lambda\nu}^b)<\theta, U(\theta_\lambda)\phi_{\lambda\nu}>^2\}^{\frac{1}{2}} \quad ,$$

and let $\Theta_{\lambda b}$ denote the associated Hilbert space obtained by completing $\{\theta \in \Theta : ||\theta||_{\lambda b} < \infty\}$ in $||\cdot||_{\lambda b}$ norm, with inner product

$$<\theta,\varsigma> = \sum_{\nu=1}^{\infty}(1+\gamma_{\lambda\nu}^b)<\theta, U(\theta_\lambda)\phi_{\lambda\nu}><\varsigma, U(\theta_\lambda)\phi_{\lambda\nu}> \quad .$$

According to Section 1 of the Appendix, the $\Theta_{\lambda b}$ norms are uniformly equivalent, provided $0 \leq b \leq 1$ (Corollary A1.4), so it suffices to consider a fixed λ, say $\lambda = 0$. We write

$$\Theta_{0b} = \Theta_b \quad , \quad ||\cdot||_{0b} = ||\cdot||_b \quad , \quad <\cdot,\cdot>_{0b} = <\cdot,\cdot>_b \quad , \quad \gamma_\nu = \gamma_{0\nu} \quad , \quad \phi_\nu = \phi_{0\nu} \tag{2.7}$$

Furthermore, if $0 \leq b \leq 1$, then $||\cdot||_b$ is equivalent to W_2^{bm}-norm (Lemma A1.2). If the penalty is of the form (1.1), then for $1 \leq b \leq 2$ and $bm-1/2$ not an integer, the $\Theta_{\lambda b}$ norms are uniformly equivalent and Θ_b can be identified as a closed subspace of a Sobolev space W_2^{bm} satisfying certain homogeneous boundary conditions, and with a norm equivalent to the Sobolev norm. The

parameterized collection $\{\Theta_b : b \geq 0\}$ of Hilbert spaces, determined by W and U, is referred to as the *penalty-information* (PI) *scale* of Hilbert spaces. We now state our main results. For the density and hazard estimators we have:

Theorem 2.1 *Suppose Assumptions A, B, and C hold. Let* $p = s/m$ *where* m *is given in C(ii) and* s *in* $C(v)$. *There is some* λ_0 *such that for any* b *satisfying*

$$0 \leq b < \min\{ 2-d/2m \ , \ p \ , \ (p-d/2m)/2 \} \ ,$$

$$||\bar{\theta}_{n\lambda} - \theta_\lambda||_b^2 = O_p(n^{-1}\lambda^{-(b+3d/2m)}) \tag{2.8}$$

$$||\theta_\lambda - \theta_0||_b^2 \lesssim \lambda^{(p-b)} \quad . \tag{2.9}$$

uniformly for $\lambda \in (0,\lambda_0]$. *Moreover if, for some* $\epsilon > \min(0, b-d/2m)$, λ_n *is a sequence such that*

$$n^{(\epsilon-m/3d)} << \lambda_n << 1$$

then for $\lambda = \lambda_n$,

$$||\bar{\theta}_{n\lambda} - \dot{\theta}_{n\lambda}||_b^2 << n^{-1}\lambda^{-(b+d/2m)} \ . \tag{2.10}$$

Proof: Theorem A2.1 of the appendix gives (2.8). Let

$$c^* = \epsilon + d/2m$$

for some ϵ such that

$$b \leq c^* < p.$$

The existence of c^* is guaranteed by C(v). From Theorems 5.1 and 4.1, λ_0 exists and we have

$$||\theta_\lambda - \theta_0||_b^2 \approx \lambda^{(c^*-b)}\lambda^{(p-c^*)} = \lambda^{(p-b)}$$

proving (2.9). Corollary 5.3 and Theorem 4.2 imply

$$\begin{aligned} ||\bar{\theta}_{n\lambda} - \dot{\theta}_{n\lambda}||_b^2 &\approx \{n^{-1}\lambda^{-2(c^*+d/m)}\lambda^{(c^*-b)}\} \cdot n^{-1}\lambda^{-(c^*+d/2m)} \\ &= \{n^{-1}\lambda^{-2(\epsilon+3d/2m)}\} \cdot n^{-1}\lambda^{-(b+d/2m)} \\ &<< n^{-1}\lambda^{-(b+d/2m)} \end{aligned}$$

provided $\lambda \in [\lambda_n, \lambda_0]$. ***Q.E.D.***

The optimal upper bound on the rate of convergence is obtained by equating the asymptotic orders of the variance and bias. The asymptotic behavior of these quantities is given in (2.8) and (2.9). Equating, we find that the optimal rate of convergence applies if

$$\lambda_n^* = n^{-2m/(2mp+d)}$$

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

$$||\hat{\theta}_{n\lambda} - \theta_0||_b^2 = O_p(n^{-2m(p-b)/(2mp+d)})$$

The conditions under which λ_n^* satisfies $n^{(\epsilon-3d/m)} << \lambda_n^* << 1$ for $b \leq d/2m + \epsilon < p$, can be worked out on a case by case basis. For example if $b < d/2m$ then we can let ϵ be arbitrarily small, and the optimal rate is covered provided

$$d < mp$$

which is guaranteed by C(v).

For the regression case we have:

Theorem 2.2 *Suppose Assumptions A, B, C, and D hold. Let* $p = s/m$ *where* m *is given in C(ii) and* s *in* $C(v)$. *Let* b *satisfy*

$$0 \leq b < \min\{2-d/2m, p, (p-d/2m)/2\} .$$

Suppose $\lambda = \lambda_n$ *is a deterministic sequence such that for some* $\epsilon > 0$,

$$\max\{k_n^{m/d}, n^{m/d} k_n^{2m/d}, n^{(\epsilon - m/3d)}\} << \lambda_n << 1 . \tag{2.11}$$

where k_n *is given in assumption B. Then*

$$||\bar{\theta}_{n\lambda} - \hat{\theta}_{n\lambda}||_b^2 << n^{-1}\lambda^{-(b+d/2m)} \tag{2.12}$$

$$E||\bar{\theta}_{n\lambda} - \theta_\lambda||_b^2 \approx n^{-1}\lambda^{-(b+d/2m)} \tag{2.13}$$

$$||\theta_\lambda - \theta_0||_b^2 \leq \lambda^{-(p-b)} . \tag{2.14}$$

Proof: Theorem A2.2 of the appendix gives (2.13). For the other two relations, we need to intro-

duce an auxiliary norm. Let

$$c^* = \epsilon + d/m$$

where $\epsilon > 0$ is chosen so that

$$b \leq c^* < p \quad .$$

Note that such a c^* exists by Assumption C(v) and our requirements on b. It follows from Theorem 4.1 and 5.1 that

$$||\theta_\lambda - \theta_0||_b^2 \leq \lambda^{(c^*-b)} \lambda^{(p-c^*)} = \lambda^{(p-b)} \quad ,$$

which proves (2.14). Theorem 4.2 and Corollary 5.5 imply

$$\begin{aligned} ||\bar{\theta}_{n\lambda} - \hat{\theta}_{n\lambda}||_b^2 &<< \{n^{-1}\lambda^{-2(c^*+d/2m)}\lambda^{(c^*-b)}\} \cdot \{n^{-1}\lambda^{-(c^*+d/2m)}\} \\ &<< \{n^{-1}\lambda^{-2(c+3d/2m)}\} \cdot \{n^{-1}\lambda^{-(b+d/2m)}\} \\ &<< n^{-1}\lambda^{-(b+d/2m)} \end{aligned}$$

which is (2.12). *Q.E.D.*

Again, using these results, we can obtain information about convergence rates. Note that (2.13) gives the order of the asymptotic variance and (2.14) gives an upper bound on the order of the asymptotic bias. The optimal upper bound on the convergence rate is obtained by setting these equal. The value of λ so obtained is

$$\lambda_n^* \approx n^{-2m/(2mp+d)} \tag{2.15}$$

which results in

$$||\hat{\theta}_{n\lambda} - \theta_0||_b^2 \leq n^{-2m(p-b)/(2mp+d)} \quad .$$

It remains to be seen if this convergence rate satisfies (2.11). To this end, we must specify k_n. In what follows, ϵ denotes an arbitrarily small positive quantity, not necessarily the same in each appearance. According to results in numerical integration (e.g. Davis and Rabinowitz[12], p. 268 ff.), we have

$$k_n \gtrsim (\log n)^{(d-1)/2} n^{-1} \gtrsim n^{\epsilon-1}$$

and the lower bound is achievable. With this estimate on the "discrepancy" k_n, one obtains that $n^{(\epsilon - m/4d)}$ is the dominant term on the l.h.s. of (2.11), and that the λ sequence in (2.15) satisfies (2.11) provided

$$mp > 5d/2 \quad . \tag{2.16}$$

One will typically believe $p \geq 1$, so we see that a rather stringent lower bound on m is required for our theory to cover the optimal convergence rate. We conjecture that this lower bound can be reduced considerably.

3. Existence of Minimizers of the General Penalized Likelihoods

The general penalized likelihood, defined on a Banach space Θ with norm $\|\cdot\|$, is written as

$$l_{n\lambda}(\theta) = \int l(y,x,\theta) P_{XY}^{(n)}(dxdy) + \lambda J(\theta) \tag{3.1}$$

where $l : Y \times X \times \Theta \to \mathbb{R}^+$ and $J : \Theta \to \mathbb{R}^+$. For arbitrary "likelihood" functional, l, and penalty functional, J, the existence of a minimizer of (3.1) is difficult to check. In large samples and for λ not too big the results in section 4 demonstrate that, with increasing probability, there exist locally unique roots of the penalized likelihood equations ($Dl_{n\lambda} \equiv 0$). If the penalized likelihood is convex then in large samples, with increasing probability, it will have a unique minimizer. However, this is still an asymptotic result which does not provide useful information for what might happen in small samples. We now present some results on the existence of minimizers in the finite sample situation.

From classical optimization theory Luenberger[17] we know that if f is a *weakly lower semi-continuous* functional and for some K the set $\{f \leq K\}$ is non-empty and bounded, then one is guaranteed, the existence of at least one minimizer of f. Unfortunately, for penalized likelihood functionals, the boundedness condition is difficult to check in practice and a simpler condition is desirable. Intuitively, one would hope that the effect of the penalty/prior ought to be to improve the identifibility of the parameter. Thus if the estimator existed for some λ_0 then for any $\lambda > \lambda_0$ the corresponding penalized likelihood estimator should also exist. Along these lines we have the following result.

Theorem 3.1 *Suppose that for all λ, $l_{n\lambda}$ is weakly lower semi-continuous and that for some λ_0 the sets, $\{l_{n\lambda_0} \leq K\}$, are bounded for all K. Then for all $\lambda > \lambda_0$, $l_{n\lambda}$ has a minimizer.*

Proof: Given $\lambda > \lambda_0$, choose $\theta_1 \in \Theta$ and K such that $l_{n\lambda}(\theta_1) \leq K$ and $\{l_{n\lambda_0} \leq K\}$, is non-empty.

Since $\lambda > \lambda_0$ and $J(\theta) \geq 0$, $\{l_{n\lambda} \leq K\}$, $\subseteq$ $\{l_{n\lambda_0} \leq K\}$, and as the latter set is bounded the existence of the minimizer of $l_{n\lambda}$ follows. *Q.E.D.*

A more interesting possibility, raised by Silverman[23], is the following. Consider

$$\Theta_0 = \{\, \theta : J(\theta) = 0 \,\} \quad ,$$

noting that the minimization of the "likelihood" part of $l_{n\lambda}$, $\int l(y,x,\theta)P_{XY}^{(n)}(dxdy)$, over Θ_0 corresponds to minimizing $l_{n\lambda}$ with $\lambda = \infty$, one asks, when does the existence of a minimizer of $\int l(y,x,\theta)P_{XY}^{(n)}(dxdy)$ over Θ_0 guarantee the existence of a minimizer of $l_{n\lambda}$ for $0 < \lambda < \infty$. The answer to this question in, at least, a Hilbert space setting is provided by the following result.

Theorem 3.2 *Let Θ be a Hilbert space and suppose Θ_0 is a linear subspace of Θ with P being the projection map onto the orthogonal complement of Θ_0 - Θ_0^{perp} say. Suppose*

(i) *J is weakly lower semi-continuous and $J(\theta) \geq$ constant $||P\theta||$ for all $\theta \in \Theta_0^{perp}$*

(ii) *$\int l(y,x,\theta)P_{XY}^{(n)}(dxdy)$ is weakly continuous and convex on Θ.*

Then whenever $\exists !$ minimizer, $\hat{\theta}_0$ of $\int l(y,x,\theta)P_{XY}^{(n)}(dxdy)$ in Θ_0, the penalized likelihood estimator exists for $0 < \lambda < \infty$.

Proof: Let $0 < \lambda < \infty$ be given and let $\theta_1 \in \Theta$. If $B = \{\, l_{n\lambda} \leq l_{n\lambda}(\theta_1) \,\}$ is bounded then we are done (by the weak lower semi-continuity of $l_{n\lambda}$). Suppose B is unbounded, then $\exists \{\theta_k\} \subset B$ such that $||\theta_k|| \to \infty$ and $\{\, l_{n\lambda}(\theta_k) \,\}$ is bounded.

Obviously, by (i), $\{\, ||P(\theta_k)|| \,\}$ must be bounded so it must be that $||(I-P)\theta_k|| \to \infty$ as $k \to \infty$. However

$$\theta_k^0 \equiv \frac{1}{2}(I-P)\theta_k = \frac{1}{2}\theta_k + \frac{1}{2}(-P\theta_k)$$

and $\int l(y,x,\theta)P_{XY}^{(n)}(dxdy)$ is convex so we have

$$\int l(y,x,\theta_k^0)P_{XY}^{(n)}(dxdy) \leq \frac{1}{2}\int l(y,x,\theta_k)P_{XY}^{(n)}(dxdy) + \frac{1}{2}\int l(y,x,(-P\theta_k))P_{XY}^{(n)}(dxdy)$$

and since $\int l(y,x,\theta)P_{XY}^{(n)}(dxdy)$ is convex and has a unique minimizer in Θ_0, $\int l(y,x,\theta_k^0)P_{XY}^{(n)}(dxdy) \to \infty$ as $k \to \infty$. But this implies, since $\int l(y,x,\theta)P_{XY}^{(n)}(dxdy)$ is weakly continuous and $\{\, ||P(\theta_k)|| \,\}$ is bounded, that $\int l(y,x,\theta_k)P_{XY}^{(n)}(dxdy) \to \infty$ as $k \to \infty$. Contradicting the definition of $\{\, \theta_k \,\}$. *Q.E.D.*

Remark: One can relax the convexity condition to quasi-convexity, provided the existence of a unique minimizer of $\int l(y,x,\theta)P_{XY}^{(n)}(dxdy)$ over Θ_0 guarantees that the sets {

$\theta \in \Theta_0 : \int l(y, x, \theta) P_{XY}^{(n)}(dxdy) < K \}$ are bounded for all K.

4. Linearization of Roots of Penalized Likelihood-Type Equations

Recall, from equation (1.2), that in the general case

$$\theta_0 = \arg\min_{\Theta}\{\int l(y,x,\theta)P_{XY}(dxdy)\}$$

So with differentiability,

$$Z_0(\theta_0) \equiv 0 \quad .$$

As we indicated in section 1 there are two general linearization theorems; a continuous one and a discrete one. To describe these results in detail we need to introduce some further technical machinery. Firstly, let $\|\cdot\|_{\lambda c}$ be a family of norms, indexed by λ and c, on Θ. These norms are used to measure convergence of estimators. The Banach space obtained by completing Θ under the norm $\|\cdot\|_{\lambda c}$ is denoted $\Theta_{\lambda c}$. For notational convenience we will often drop the λ subscript. The linearization theorems below depend on applying a particular fixed-point theorem. Now it will turn out that special conditions have to be placed on the norm in order that this fixed-point property be obtained - see section 5. However, once the fixed-point property holds in a particular norm information about the behavior of the linearization in a variety of (typically weaker) norms can be read off. For this reason, the results in this section are always stated for a pair of norms, $\|\cdot\|_{\lambda c^*}$ and $\|\cdot\|_{\lambda c}$. The fixed-point property is established in the c^*-norm while the c-norm results give the behavior of the linearization in norms of more direct interest.

Continuous Linearization

Let

$$b(\lambda) \approx \|\bar{\theta}_\lambda - \theta_0\|_{c^*} = \|G_\lambda^{-1}(\theta_0)(Z_\lambda(\theta_0) - Z_0(\theta_0))\|_{c^*} = \|G_\lambda^{-1}(\theta_0)(\lambda DJ(\theta_0))\|_{c^*} \qquad (4.1)$$

and for some admissible c and c^* and $R > 0$, define constants

$$K_2^*(\lambda,R) = \sup_{\substack{\phi_3 \in S(1,\Theta_{\lambda c^*}) \\ \phi_1,\phi_2 \in S(R,\Theta_{\lambda c^*})}} \|G_\lambda^{-1}(\theta_0)D^3 l_\lambda(\theta_0+\phi_1)\phi_2\phi_3\|_{c^*} \quad ,$$

and

$$K_2(\lambda,R) = \sup_{\substack{\phi_3 \in S(1,\Theta_{\lambda c}\cdot) \\ \phi_1,\phi_2 \in S(R,\Theta_{\lambda c}\cdot)}} ||G_\lambda^{-1}(\theta_0)D^3 l_\lambda(\theta_0+\phi_1)\phi_2\phi_3||_c \quad ,$$

where $D^3 l_{n\lambda}(\theta) = \int D^3 l(y,x,\theta)P_{XY}^{(n)}(dxdy) + \lambda D^3 J(\theta)$. (We assume that l and J are 3 times differentiable w.r.t θ).

The hypothesis for the continuous linearization theorem concern the behavior of these constants as λ approaches zero. Let $r^*(\lambda)$, and $r(\lambda)$ be sequences such that

$$K_2^*(\lambda,x(\lambda)) \approx r^*(\lambda)$$

$$K_2(\lambda,x(\lambda)) \approx r(\lambda) \tag{4.2}$$

for $x(\lambda) \approx b(\lambda)$.

Theorem 4.1 *Let c and c^* be given. Suppose $\lim_{\lambda\to 0} r^*(\lambda) \to 0$, then we can find constants K_0, K and λ_0 such that, for all $\lambda \in [0,\lambda_0]$, $\exists !\ \phi_\lambda \in S(\frac{1}{2}K_0 b(\lambda),\Theta_{\lambda c}\cdot)$ such that if $\theta_\lambda = \phi_\lambda + \theta_0$ then $Z_\lambda(\theta_\lambda) = 0$, and*

$$||\bar{\theta}_\lambda - \theta_\lambda||_{c^*} \leq K\ r^*(\lambda)\ b(\lambda)$$

$$||\bar{\theta}_\lambda - \theta_\lambda||_c \leq K\ r(\lambda)\ b(\lambda)$$

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

Proof. Consider the mapping, F_λ, on Θ given by

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

By definition of $b(\lambda)$, $\exists K_0$ such that for all $\lambda \leq \lambda_0$, $||\bar{\theta}_\lambda - \theta_0||_{c^*} < \frac{1}{2}K_0\, b(\lambda)$. The conclusions of the theorem will follow once we have established that F_λ is a contraction on the closed ball $S(\frac{1}{2}K_0 b(\lambda),\Theta_{\lambda c}\cdot)$ for all λ sufficiently small.

To show that F_λ maps the ball into itself, we consider $||F_\lambda(\phi)||_{c^*}$ for $\phi \in S(\frac{1}{2}K_0 b(\lambda),\Theta_{\lambda c}\cdot)$.

$$||F_\lambda(\phi)||_{c^\bullet} = ||F_\lambda(\phi) - (\bar{\theta}_\lambda - \theta_0) + (\bar{\theta}_\lambda - \theta_0)||_{c^\bullet}$$
$$\leq ||\phi - G_\lambda^{-1}(\theta_0) Z_\lambda(\theta_0+\phi) + G_\lambda^{-1}(\theta_0) Z_\lambda(\theta_0)||_{c^\bullet} + ||\bar{\theta}_\lambda - \theta_0||_{c^\bullet} \quad ,$$

while for the contraction property, we look at $||F_\lambda(\phi_1) - F_\lambda(\phi_2)||_{c^\bullet}$ for $\phi_1, \phi_2 \in S(\frac{1}{2} K_0 b(\lambda), \Theta_{\lambda c^\bullet})$,

$$||F_\lambda(\phi_1) - F_\lambda(\phi_2)||_{c^\bullet} = ||\phi_1 - \phi_2 - G_\lambda^{-1}(\theta_0)[Z_\lambda(\theta_0+\phi_1) - Z_\lambda(\theta_0+\phi_2)]||_{c^\bullet} \quad .$$

Applying the mean value theorem, for any continuous linear functional f, $\exists\ \phi^*$ between ϕ and the origin such that

$$f(\phi - G_\lambda^{-1}(\theta_0)[Z_\lambda(\theta_0+\phi) - Z_\lambda(\theta_0)]) = f(\phi - G_\lambda^{-1}(\theta_0)[DZ_\lambda(\theta_0)\phi + \frac{1}{2} D^2 Z_\lambda(\theta_0+\phi^*)\phi\phi])$$
$$= f(\frac{1}{2} G_\lambda^{-1}(\theta_0) D^3 l_\lambda(\theta_0+\phi^*)\phi\phi) \quad \text{since} \quad DZ_\lambda = G_\lambda$$

Thus

$$|f(\phi - G_\lambda^{-1}(\theta_0)[Z_\lambda(\theta_0+\phi) - Z_\lambda(\theta_0)])| = |f(\frac{1}{2} G_\lambda^{-1}(\theta_0) D^3 l_\lambda(\theta_0+\phi^*)\phi\phi)|$$

Taking the supremum over functionals of unit norm we have

$$||\phi - G_\lambda^{-1}(\theta_0)[Z_\lambda(\theta_0+\phi) - Z_\lambda(\theta_0)]||_{c^\bullet} = \sup_{\phi^* \in L[0,\phi]} ||\frac{1}{2} G_\lambda^{-1}(\theta_0) D^3 l_\lambda(\theta_0+\phi^*)\phi\phi)||_{c^\bullet}$$

where $L[0,\phi] = \{ t\phi \mid t \in [0,1] \}$. Hence, by definition of $K_2^\bullet$

$$||F_\lambda(\phi)||_{c^\bullet} \leq \{\frac{1}{2} K_2^\bullet(\lambda, K_0 b(\lambda)) + \frac{1}{2}\} K_0\, b(\lambda)$$
$$\leq \{K\ r^\bullet(\lambda) + \frac{1}{2}\} K_0\, b(\lambda)$$

Expanding $Z_\lambda(\theta_0+\phi_1) - Z_\lambda(\theta_0+\phi_2)$, an analogous argument gives

$$||F_\lambda(\phi_1) - F_\lambda(\phi_2)||_{c^\bullet} \leq \sup_{\substack{\phi^{**} \in L[\phi_1,\phi_2] \\ \phi^{***} \in L[0,\phi^{**}]}} ||G_\lambda^{-1}(\theta_\lambda) D^3 l_\lambda(\theta_0+\phi^{***})\phi^{**}(\phi_1-\phi_2)||_{c^\bullet}$$
$$\leq \{K_2^\bullet(\lambda, K_0\, b(\lambda))\}\ ||\phi_1 - \phi_2||_{c^\bullet}$$

$$\leq \{K\ r^\bullet(\lambda)\}\ ||\phi_1 - \phi_2||_{c^\bullet}$$

Thus, since $r^\bullet(\lambda) \to 0$, there is some λ_0 such that the terms in brackets are less than one for all $\lambda < \lambda_0$. Hence, for all $\lambda < \lambda_0$, F_λ is a contraction on $S(\frac{1}{2} K_0 b(\lambda), \Theta_{\lambda c^\bullet})$. and so, by Theorem 9.23 of Rudin[22], F_λ has a unique fixed point, ϕ_λ, in $S(\frac{1}{2} K_0 b(\lambda), \Theta_{\lambda c^\bullet})$. It follows that

$\theta_\lambda = \theta_0 + \phi_\lambda$ is the unique solution to $Z_\lambda \equiv 0$ in $S(\frac{1}{2}K_0 b(\lambda), \Theta_{\lambda c}^{\bullet})$. By definition of $\bar{\theta}_\lambda$

$$\begin{aligned}\bar{\theta}_\lambda - \theta_\lambda &= F_\lambda(\phi_\lambda) - G_\lambda^{-1}(\theta_0) Z_\lambda(\theta_0) \\ &= F_\lambda(\phi_\lambda) - F_\lambda(0)\end{aligned}$$

Thus

$$\begin{aligned}||\bar{\theta}_\lambda - \theta_\lambda||_c^{\bullet} &\leq ||F_\lambda(\phi_\lambda) - F_\lambda(0)||_c^{\bullet} \\ &\leq K\ r^{\bullet}(\lambda)\ b(\lambda) \quad .\end{aligned}$$

Using the definition of K_2 and $r(\lambda)$, possibly altering the choice of the generic constant K and reducing λ_0, we can also guarantee that

$$||\bar{\theta}_\lambda - \theta_\lambda||_c \leq K\ r(\lambda)\ b(\lambda) \quad ,$$

for all $\lambda \leq \lambda_0$. *Q.E.D.*

Discrete Linearization

The existence of θ_λ for all $\lambda \in [0, \lambda_0]$ allows us to describe a discrete analogue of Theorem 4.1. For $\lambda < \lambda_0$, let the sample Hessian operator at θ_λ be denoted:

$$I_{n\lambda} = \int D^2 l(y, x, \theta_\lambda) P_{XY}^{(n)}(dxdy) + \lambda D^2 J(\theta_\lambda)$$

For some admissible c and $c^{\bullet}$, let

$$K_1^{\bullet}(n, \lambda) = \sup_{\phi \in S(1, \Theta_{\lambda c}^{\bullet})} ||G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_\lambda(\theta_\lambda))\phi||_c^{\bullet}$$

$$K_2^{\bullet}(n, \lambda, R) = \sup_{\substack{\phi_3 \in S(1, \Theta_{\lambda c}^{\bullet}) \\ \phi_1, \phi_2 \in S(R, \Theta_{\lambda c}^{\bullet})}} ||G_\lambda^{-1}(\theta_\lambda) D^3 l_{n\lambda}(\theta_\lambda + \phi_1)\phi_2\phi_3||_c^{\bullet} \quad ,$$

and

$$K_1(n, \lambda) = \sup_{\phi \in S(1, \Theta_{\lambda c}^{\bullet})} ||G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_\lambda(\theta_\lambda))\phi||_c$$

$$K_2(n, \lambda, R) = \sup_{\substack{\phi_3 \in S(1, \Theta_{\lambda c}^{\bullet}) \\ \phi_1, \phi_2 \in S(R, \Theta_{\lambda c}^{\bullet})}} ||G_\lambda^{-1}(\theta_\lambda) D^3 l_{n\lambda}(\theta_\lambda + \phi_1)\phi_2\phi_3||_c \quad ,$$

where $D^3 l_{n\lambda}(\theta) = \int D^3 l(y, x, \theta) P_{XY}^{(n)}(dxdy) + \lambda D^3 J(\theta)$.

The assumption for the discrete linearization depends on the asymptotic behavior of these constants. Roughly speaking, we require that the sample Hessian operator, $I_{n\lambda}$, converges to the limiting Hessian operator, $G_\lambda(\theta_\lambda)$, in an appropriate manner and that l and J are sufficiently

smooth. Let $d(n,\lambda)$, $r^*(n,\lambda)$, and $r(n,\lambda)$ be (non-random) sequences such that *for all* $\lambda \in [\lambda_n, \lambda_0]$

$$\begin{aligned} ||\bar{\theta}_{n\lambda} - \theta_\lambda||_{c^*} &= O_p(d(n,\lambda)) \\ K_1^*(n,\lambda) + K_2^*(n,\lambda,x(n,\lambda)) &= O_p(r^*(n,\lambda)) \\ K_1(n,\lambda) + K_2(n,\lambda,x(n,\lambda)) &= O_p(r(n,\lambda)) \quad \text{whenever } x(n,\lambda) \approx d(n,\lambda). \end{aligned} \tag{4.3}$$

With conditions on the behavior of these quantities, we have a discrete linearization theorem.

Theorem 4.2 *Let c and c^* be given. Suppose $\{\lambda_n\}$ is a such that $r^*(n,\lambda) \to 0$ as $n \to \infty$ for any sequence of λ's in $[\lambda_n, \lambda_0]$. Consider the event $E(n,\lambda)$ given by*

$$E(n,\lambda): \exists!\ \textit{solution to } Z_{n\lambda} \equiv 0\ ,\dot{\theta}_{n\lambda} = \theta_\lambda + \dot{\phi}_{n\lambda}\ ,\ \dot{\phi}_{n\lambda} \in S(\tfrac{1}{2}K_0 d(n,\lambda), \Theta_{\lambda c^*}),$$

$$\textit{satisfying } ||\dot{\theta}_{n\lambda} - \bar{\theta}_{n\lambda}||_{c^*} \le K\ r^*(n,\lambda)\ d(n,\lambda)\ ,$$

$$\textit{and } ||\dot{\theta}_{n\lambda} - \bar{\theta}_{n\lambda}||_{c} \le K\ r(n,\lambda)\ d(n,\lambda)\ ,$$

then for all $\delta > 0$ we can find n_0 and constants K_0 and K such that this event occurs with probability $> 1-\delta$, for all $n \ge n_0$ and $\lambda \in [\lambda_n, \lambda_0]$.

Proof. The argument amounts to a probabilistic version of the proof of Theorem 4.1. Since $||\bar{\theta}_{n\lambda} - \theta_\lambda||_{c^*} = O_p(d(n,\lambda))$, for some K_0 the event $E'(n,\lambda): ||\bar{\theta}_{n\lambda} - \theta_\lambda||_{c^*} < \frac{1}{2}K_0\, d(n,\lambda)$ occurs with probability $> 1-\delta/3$ *for all* $n > n_0$ and $\lambda \in [\lambda_n, \lambda_0]$.

Consider the mapping $F_{n\lambda}: \Theta \to \Theta$ given by

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

The existence and uniqueness of $\dot{\theta}_{n\lambda}$ will follow once we have established that, for n sufficiently large and $\lambda \in [\lambda_n, \lambda_0]$, with arbitrarily high probability, $F_{n\lambda}$ is a contraction on the closed ball $S(\frac{1}{2}K_0 d(n,\lambda), \Theta_{\lambda c^*})$.

To show that $F_{n\lambda}$ maps the ball into itself, we consider $||F_{n\lambda}(\phi)||_{c^*}$ for $\phi \in S(\frac{1}{2}K_0 d(n,\lambda), \Theta_{\lambda c^*})$,

$$\begin{aligned} ||F_{n\lambda}(\phi)||_{c^*} &= ||F_{n\lambda}(\phi) - (\bar{\theta}_{n\lambda} - \theta_\lambda) + (\bar{\theta}_{n\lambda} - \theta_\lambda)||_{c^*} \\ &\le ||\phi - G_\lambda^{-1}(\theta_\lambda) Z_{n\lambda}(\theta_\lambda + \phi) + G_\lambda^{-1}(\theta_\lambda) Z_{n\lambda}(\theta_\lambda)||_{c^*} + ||\bar{\theta}_{n\lambda} - \theta_\lambda||_{c^*} \quad . \end{aligned}$$

While for the contraction property, we look at $||F_{n\lambda}(\phi_1) - F_{n\lambda}(\phi_2)||_{c^\bullet}$ for $\phi_1, \phi_2 \in S(\frac{1}{2}K_0 d(n,\lambda), \Theta_{\lambda c^\bullet})$,

$$||F_{n\lambda}(\phi_1) - F_{n\lambda}(\phi_2)||_{c^\bullet} = ||\phi_1 - \phi_2 - G_\lambda^{-1}(\theta_\lambda)[Z_{n\lambda}(\theta_\lambda+\phi_1) - Z_{n\lambda}(\theta_\lambda+\phi_2)]||_{c^\bullet} .$$

Applying the mean value theorem as in Theorem 4.1, for any continuous linear functional f, $\exists$ ϕ^* between ϕ and the origin such that

$$f(\phi - G_\lambda^{-1}(\theta_\lambda)[Z_{n\lambda}(\theta_\lambda+\phi) - Z_{n\lambda}(\theta_\lambda)]) = f(\phi - G_\lambda^{-1}(\theta_\lambda)[DZ_{n\lambda}(\theta_\lambda)\phi + \frac{1}{2}D^2 Z_{n\lambda}(\theta_\lambda+\phi^*)\phi\phi])$$

Since $Z_{n\lambda} \equiv Dl_{n\lambda}$ and $DZ_{n\lambda}(\theta_\lambda) \equiv I_{n\lambda}$,

$$= f(G_\lambda^{-1}(\theta_\lambda)[G_\lambda(\theta_\lambda) - I_{n\lambda}]\phi) - f(\frac{1}{2}G_\lambda^{-1}(\theta_\lambda)D^3 l_{n\lambda}(\theta_\lambda+\phi^*)\phi\phi)$$

i.e.

$$|f(\phi - G_\lambda^{-1}(\theta_\lambda)[Z_{n\lambda}(\theta_\lambda+\phi) - Z_{n\lambda}(\theta_\lambda)])| \le |f(G_\lambda^{-1}(\theta_\lambda)[I_{n\lambda} - G_\lambda(\theta_\lambda)]\phi)| + |f(\frac{1}{2}G_\lambda^{-1}(\theta_\lambda)D^3 l_{n\lambda}(\theta_\lambda+\phi^*)\phi\phi)|$$

Taking the supremum over functionals of unit norm, we obtain

$$||F_{n\lambda}(\phi)||_{c^\bullet} \le ||G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_\lambda^{-1}(\theta_\lambda))\phi||_{c^\bullet} + \frac{1}{2}\sup_{\phi^* \in L[0,\phi]} ||G_\lambda^{-1}(\theta_\lambda)D^3 l_{n\lambda}(\theta_\lambda+\phi^*)\phi\phi||_{c^\bullet} + \frac{1}{2}K_0\, d(n,\lambda)$$

A similar expansion of $Z_{n\lambda}(\theta_\lambda+\phi_1) - Z_{n\lambda}(\theta_\lambda+\phi_2)$ yields

$$||F_{n\lambda}(\phi_1) - F_{n\lambda}(\phi_2)||_{c^\bullet} \le ||G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_\lambda(\theta_\lambda))(\phi_1-\phi_2)||_{c^\bullet} + \sup_{\substack{\phi^{**} \in L[\phi_1,\phi_2] \\ \phi^{***} \in L[0,\phi^{**}]}} ||G_\lambda^{-1}(\theta_\lambda)D^3 l_{n\lambda}(\theta_\lambda+\phi^{***})\phi^{**}(\phi_1-\phi_2)||_{c^\bullet}$$

Hence by definition of K_1^* and K_2^* on the event $E'(n,\lambda)$ we have

$$||F_{n\lambda}(\phi)||_{c^\bullet} \le |\{K_1^*(n,\lambda) + \frac{1}{2}K_2^*(n,\lambda,K_0\, d(n,\lambda))\} + \frac{1}{2}|\, K_0\, d(n,\lambda)$$

and

$$||F_{n\lambda}(\phi_1) - F_{n\lambda}(\phi_2)||_{c^\bullet} \le \{K_1^*(n,\lambda) + K_2^*(n,\lambda,K_0\, d(n,\lambda))\}\, ||\phi_1 - \phi_2||_{c^\bullet}$$

The terms in brackets can be made less than a constant times $r^*(n,\lambda)$ plus $\frac{1}{2}$ and a constant times $r^*(n,\lambda)$, respectively, with probability $> 1-\delta/3$ for n sufficiently large. Thus, since $r^*(n\lambda) \to 0$, there is some n_0 such that *for all* $n > n_0$ and $\lambda \in [\lambda_n, \lambda_0]$ the event:

$$||F_{n\lambda}(\phi)||_{c^\bullet} < K_0\, d(n,\lambda) \quad \textit{for all } \phi \in S(\tfrac{1}{2}K_0 d(n,\lambda), \Theta_{\lambda c^\bullet}) \tag{4.4}$$

$$||F_{n\lambda}(\phi_1) - F_{n\lambda}(\phi_2)||_{c^\bullet} < K\, r^*(n,\lambda)\, ||\phi_1 - \phi_2||_{c^\bullet} \quad \textit{for all } \phi_1, \phi_2 \in S(\tfrac{1}{2}K_0 d(n,\lambda), \Theta_{\lambda c^\bullet})$$

where $K\ r^*(n,\lambda) < \frac{1}{2}$,

occurs with probability $> 1-2\delta/3$. On these events, $F_{n\lambda}$ is a contraction mapping on the ball $S(\frac{1}{2}K_0 d(n,\lambda),\Theta_{\lambda c}^*)$, and so, again by Theorem 9.23 of Rudin[22], $F_{n\lambda}$ has a unique fixed point, $\dot{\phi}_{n\lambda}$ in $S(\frac{1}{2}K_0 d(n,\lambda),\Theta_{\lambda c}^*)$. Letting $\dot{\theta}_{n\lambda} = \theta_\lambda + \dot{\phi}_{n\lambda}$, we have $Z_{n\lambda}(\dot{\theta}_{n\lambda}) \equiv 0$. Moreover, since $\bar{\theta}_{n\lambda} = \theta_\lambda - G_\lambda^{-1}(\theta_\lambda)Z_{n\lambda}(\theta_\lambda)$, we have

$$\begin{aligned}
\|\dot{\theta}_{n\lambda} - \bar{\theta}_{n\lambda}\|_c^* &= \|G_\lambda^{-1}(\theta_\lambda)G_\lambda(\theta_\lambda)(\dot{\theta}_{n\lambda} - \bar{\theta}_{n\lambda})\|_c^* \\
&= \|G_\lambda^{-1}(\theta_\lambda)[G_\lambda(\theta_\lambda)\dot{\theta}_{n\lambda} - G_\lambda(\theta_\lambda)\theta_\lambda + Z_{n\lambda}(\theta_\lambda)]\|_c^* \\
&= \|F_{n\lambda}(\dot{\phi}_{n\lambda}) - F_{n\lambda}(0)\|_c^* \\
&\leq K\ r^*(n,\lambda)\ K_0\ d(n,\lambda) \quad .
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|\dot{\theta}_{n\lambda} - \bar{\theta}_{n\lambda}\|_c &= \|F_{n\lambda}(\dot{\phi}_{n\lambda}) - F_{n\lambda}(0)\|_c \\
&\leq \{K_1(n,\lambda) + \frac{1}{2}K_2(n,\lambda,K_0\ d(n,\lambda))\}\|\dot{\phi}_{n\lambda}\|_c^* \quad ,
\end{aligned}$$

and we can choose n_0 larger, if necessary, and guarantee that, for $n > n_0$ and $\lambda \in [\lambda_n,\lambda_0]$, the term in brackets is less than a constant times $r(n,\lambda)$ with probability $> 1-\delta/3$ i.e.

$$\leq K\ r(n,\lambda)\ K_0\ d(n,\lambda)$$

Combining the event upon which this occurs with the event in (4.2), we have that $\exists\, n_0$ and constants K_0 and K such that for all $n > n_0$ and $\lambda \in [\lambda_n,\lambda_0]$ the event $E(n,\lambda)$ occurs with probability $> 1-\delta$. *Q.E.D.*

Straightforward modification to the argument in Theorem 4.2 can be used to prove the following:

Extensions.

(a) If the asymptotic orders of magnitude in (4.3) are uniform for $\lambda \in [\lambda_n,\lambda_0]$, then Theorem 4.2 can be strengthened accordingly. In this case, $\bigcup_{\lambda \in [\lambda_n,\lambda_0]} E(n,\lambda)$ has arbitrarily high probability for n large enough.

(b) If the asymptotic orders of magnitude in (4.3) are almost sure, not just stochastic, then it follows that $\bigcup_{m \geq n} E(m,\lambda)$ will have arbitrarily high probability for n large enough. This, in turn, can be strengthened, as in (a), provided the asymptotic orders of magnitude are uniform for $\lambda \in [\lambda_n, \lambda_0]$.

5. Application to Generalized Smoothing

The hypotheses of linearization theorems in section 4 will now be shown to hold in the generalized smoothing context of section 2. Norms, $\|\cdot\|_{\lambda c}$, associated with the penalty-information scale of Hilbert spaces, given in (2.6) are used. Appendix 1 describes various technical properties of these norms and we will make repeated use of these properties in the sequel. Having established the existence of a valid linearization in some norm $\|\cdot\|_{\lambda c^*}$, the results will be used to derive information on the limiting behavior of the linearization in (weaker) norms, $\|\cdot\|_{\lambda c}$, for $0 \leq c \leq c^*$. Assumption A through D are in force throughout this section.

We begin with the continuous linearization. For any of the generalized smoothing estimators discussed in section 2, the limiting behavior of the constants $b(\lambda)$ and $r(\lambda)$ in (4.1) and (4.2), can be easily described. Suppose the continuous version of the penalized likelihood l_λ is such that the U operator ($U(\theta) = D^2 l_0(\theta)$), is given by

$$U(\theta)\varsigma = \int \xi'(x) h(x, \theta(x)) \varsigma(x) dx \tag{5.1}$$

where $h : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^{q \times q}$ satisfies

(h.1) *For all $R > 0$, there are positive constants $M_1(R)$ and $M_2(R)$ such that for all $x \in X$, $t \in S(R, \mathbb{R}^q)$, and $u \in \mathbb{R}^q$ with $\|u \mid \mathbb{R}^q\| = 1$,*

$$M_1(R) \leq u' h(x,t) u \leq M_2(R)$$

(h.2) *h is differentiable with respect to t and let $\dot{h}(x,t)$ denote the tensor $\frac{\partial h(x,t)}{\partial t}$. For each $R > 0$, there is a positive constant $M_3(R)$ such that for $t \in S(R, \mathbb{R}^q)$,*

$$\sup_{i,j,k,x} |\dot{h}_{ijk}(x,t)| < M_3(R)$$

A quick examination shows that for the generalized smoothing models h becomes:

Log Density Estimation: $h(x,t) = e^t$; Log Hazard Estimation: $h(x,t) = e^t S(x)$; while for Generalized Linear Regression

$$h(x,t) = \dot{\psi}(x,t) f(x) \text{ , and } \dot{h}(x,t) = \ddot{\psi}(x,t) f(x) . \tag{5.2}$$

In all cases h satisfies conditions (h.1) and (h.2): The density case is trivial, use B(ii) for the hazard case, and in the regression situation B(ii) and D(v) imply (h.1) while B(ii) and D(viii) give (h.2). Using (h.1) Appendix 1 gives results concerning the norms associated with the *P.I.* scale of Hilbert spaces derived from U. Utilizing these norms we have the following theorem.

Theorem 5.1 *Let l_λ be such that U is given in* (5.1). *If $d/2m < c^* < (2s-d)/4m$ where s is given in* C(v), *then for all $0 \leq c \leq c^*$ the constants in* (4.1) and (4.2) *have the following behavior as $\lambda \to 0$.*

(i) $\quad b(\lambda) \approx \lambda^{(p-c^*)/2}$ where $p = s/m$.

(ii) $\quad r(\lambda) << \lambda^{(c^*-c)/2}$, as $\lambda \to 0$.

Remark: By C(v) $s/m > 3d/2m$, so there exists c^* satisfying the hypothesis of the theorem.

Proof:

$$b^2(\lambda) = \|\bar{\theta}_\lambda - \theta_0\|^2_{\lambda c^*} = \|\theta_0 - G_\lambda^{-1}(\theta_0)U(\theta_0)\theta_0\|^2_{\lambda c^*} \tag{5.3}$$

so it follows directly from Theorem 2.3 (c) of Cox[8] that

$$b^2(\lambda) \leq K\lambda^{(p-c^*)}$$

for some K and all λ sufficiently small.

For part(ii), we have $D^3 l_\lambda(\theta_0+u)vw = \dot{U}(\theta_0+u)vw$ where

$$\dot{U}(\theta) = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\int \xi_i(x)\dot{h}_{ijk}(x,\theta(x))\xi_j(x)\xi_k(x)dx. \tag{5.4}$$

But the fact that $c^* > d/2m$ implies $\Theta_{\lambda c^*} \subseteq C(\bar{X};R^t)$, by Sobolev's imbedding theorem and Lemma A1.2. Hence, if λ is sufficiently small, $\phi \in S(R,\Theta_{\lambda c^*})$ implies

$$\|\phi \mid C(\bar{X};R^t)\| \leq K\ R \quad .$$

So, for any $\bar{\theta} = \theta_0+u$ with $\|u\|_{c^*} \leq R$ and $x \in X$, we have, by (h.2),

$$\sup_{i,j,k,x} |\dot{h}_{ijk}(x,\bar{\theta}(x))| \leq M_7(R^t) \tag{5.5}$$

where R' depends on θ_0 and R. It follows that

$$\begin{aligned}
&\| G_\lambda^{-1}(\theta_0)\dot{U}(\tilde{\theta})vw \|_{\lambda c}^2 \qquad (5.6)\\
&= \sum_\nu (1+\gamma_\nu^c)\cdot\{\sum_{i=1}^q\sum_{j=1}^q\sum_{k=1}^q < G_\lambda^{-1}(\theta_0)\int \xi_i(x)\dot{h}_{ijk}(x,\tilde{\theta}(x))\, v_j(x)w_k(x)dx, U(\theta_0)\phi_\nu > \}^2\\
&= \sum_\nu (1+\gamma_\nu^c)(1+2\lambda\gamma_\nu)^{-2}\cdot\{\sum_i\sum_j\sum_k \int \phi_{\nu i}(x)\dot{h}_{ijk}(x,\tilde{\theta}(x))v_j(x)w_k(x)dx\}^2\\
&\le \sum_\nu (1+\gamma_\nu^c)(1+2\lambda\gamma_\nu)^{-2}\cdot\{q^2\int \|\phi_\nu(x) \mid \mathbb{R}^q\|^2 dx\}\\
&\qquad\cdot\{M_7^2(R')K\,\|v \mid C(\underline{X},\mathbb{R}^q)\|^2\,\|w \mid C(\underline{X},\mathbb{R}^q)\|^2\}\\
&\le K(R)\,\|v \mid C(\underline{X},\mathbb{R}^q)\|^2\,\|w \mid C(\underline{X},\mathbb{R}^q)\|^2 \sum_\nu (1+\gamma_\nu^c)(1+2\lambda\gamma_\nu)^{-2}\ .
\end{aligned}$$

Here, the second equality follows from Lemma A1.3 (ii), the third relation from Cauchy-Schwartz inequality (5.5), and the fourth from Lemma A1.2 with $b = 0$, and the fact that $\|\phi_\nu\|_0^2 = 2$ for all $\nu=1,2,\cdots$. Thus, if $v \in S(R,\Theta_{\lambda c^*})$ and $w \in S(1,\Theta_{\lambda c^*})$, by Sobolev's imbedding theorem and the fact that $c^* > d/2m$, the sup-norms of v and w can be bounded by constant multiples of R and 1 respectively. Thus for such u, v and w, by Lemma A1.1(c)

$$\| G_\lambda^{-1}(\theta_0)D^3 l_\lambda(\theta_0+u)vw \|_{\lambda c}^2 \le K(R)\,\lambda^{-(c+d/2m)}$$

and so for $\lambda < \lambda_0$, from part(i),

$$|r(\lambda)|^2 \le K\ \lambda^{(p-c^*)}\lambda^{-(c+d/2m)}$$

where K depends on θ_0 and λ_0. However, since $2c^* < s/m - d/2m$, $\lambda^{p-2c^*-d/2m} << 1$ and so $|r(\lambda)|^2 << \lambda^{(c^*-c)}$. *Q.E.D.*

For the discrete linearization we must analyze the constants given in (4.3). Results for the density and hazard estimators are given in the next theorem. Theorem 5.4 tackles the regression case.

Theorem 5.2. *Consider the log density and hazard estimators. Let λ_0 be such that Theorem 4.1 holds. If $d/2m < c^* < 2-3d/2m$ then for $0 \le c \le c^*$ we have*

(a) $\sup_{\|u\|_{\lambda c^*}^2 = 1} \| G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_\lambda(\theta_\lambda))u \|_{\lambda c}^2 \le T_1(n,\lambda)$ and

(b) $$\sup_{\substack{\|u\|^2_{\lambda c^*} \le R^2, \|v\|^2_{\lambda c^*} \le R^2 \\ \|w\|^2_{\lambda c^*} \le 1}} \|G_\lambda^{-1}(\theta_\lambda) D^3 l_{n\lambda}(\theta_\lambda + u)vw\|^2_{\lambda c} \le T_2(n,\lambda) R^2$$

$$\text{where } T_1(n,\lambda) = \begin{cases} 0 & \text{density case} \\ O_p(k_n^2)\lambda^{-(c+d/2m)} & \text{hazard case} \end{cases},$$

$$\text{and } T_2(n,\lambda) = \begin{cases} O(\lambda^{-(c+d/2m)}) & \text{density case} \\ O_p(\lambda^{-(c+d/2m)}) & \text{hazard case} \end{cases}.$$

All relations being uniform in λ for $\lambda \in [0,\lambda_0]$.

Remark: Assumption C(ii) guarantees $d/2m < 2-3d/2m$, so the assumption on c^* is not vacuous.

Proof: With $u \in S(1,\Theta_{\lambda c^*})$, we have

$$G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_\lambda(\theta_\lambda))u = \begin{cases} 0 & \text{density case} \\ G_\lambda^{-1}(\theta_\lambda)\int \xi(x) e^{\theta_\lambda(x)} u(x)[S_n(x) - S(x)]dx & \text{hazard case} \end{cases}.$$

In the hazard case,

$$\begin{aligned} &\|G_\lambda^{-1}(\theta_\lambda)\int \xi(x) e^{\theta_\lambda(x)} u(x)[S_n(x) - S(x)]dx\|^2_{\lambda c} \\ &\qquad = \sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu}^c] < G_\lambda^{-1}(\theta_\lambda)\int \xi(x) e^{\theta_\lambda(x)} u(x)[S_n(x) - S(x)]dx, U(\theta_\lambda)\phi_{\lambda\nu} >^2 \\ &\qquad = \sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu}^c][1+2\lambda\gamma_{\lambda\nu}]^{-2} \{\int \phi_{\lambda\nu}(x) e^{\theta_\lambda(x)} u(x)[S_n(x) - S(x)]dx\}^2 \end{aligned}$$

the latter relation comes from Lemma A1.3(ii). For $\lambda < \lambda_0$, θ_λ is bounded in sup-norm, and $c^* > d/2m$ so by Lemma A1.2, $\Theta_{\lambda c^*} \subseteq C(\underline{X}, \mathbb{R})$. Hence the term in brackets can be bounded as

$$\begin{aligned} \left| \int \phi_{\lambda\nu}(x) e^{\theta_\lambda(x)} u(x)[S_n(x) - S(x)]dx \right|^2 &\le \sup_x |F_n(x) - F(x)|^2 \|u\|_{C(\underline{X},\mathbb{R})}\|^2 \int \phi^2_{\lambda\nu}(x)dx \\ &\approx k_n^2 \|u\|^2_{\lambda c}. \end{aligned}$$

By Lemma A1.3(i), $\|\phi_{\lambda\nu}\|_{L_2(\underline{X},\mathbb{R})}\|^2 \approx 1$ uniformly in λ and ν. It follows that

$$\sup_{\|u\|^2_{\lambda c^*} = 1} \|G_\lambda^{-1}(\theta_\lambda)\int \xi(x) e^{\theta_\lambda(x)} u(x)[S_n(x) - S(x)]dx\|^2_{\lambda c} \le k_n^2 \sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu}^c][1+2\lambda\gamma_{\lambda\nu}]^{-2}$$

$$= k_n^2 \lambda^{-(c+d/2m)} \qquad \text{by A1.1(c)} .$$

Since the relation is uniform in λ for $\lambda \in [0,\lambda_0]$, this proves the first part of the theorem. For the second part we let u, $v \in S(R, \Theta_{\lambda c}\cdot)$ and $w \in S(1, \Theta_{\lambda c}\cdot)$,

$$||G_\lambda^{-1}(\theta_\lambda) D^3 l_{n\lambda}(\theta_\lambda + u) vw||_{\lambda c}^2 = \begin{cases} ||G_\lambda^{-1}(\theta_\lambda) \int \xi' (x) e^{\theta_\lambda(x)+u(x)} v(x) w(x) dx||_{\lambda c}^2 & \textit{density case} \\ ||G_\lambda^{-1}(\theta_\lambda) \int \xi' (x) e^{\theta_\lambda(x)+u(x)} v(x) w(x) S_n(x) dx||_{\lambda c}^2 & \textit{hazard case.} \end{cases}$$

The density case is independent of n. Indeed, the analysis in Theorem 5.1 gives $O(\lambda^{-(c+d/2m)}) R^2$ for this term uniformly in λ. For the hazard case, we have

$$\begin{aligned} &||G_\lambda^{-1}(\theta_\lambda) \int \xi' (x) e^{\theta_\lambda(x)+u(x)} v(x) w(x) S_n(x) dx||_{\lambda c}^2 \\ &\qquad \le ||G_\lambda^{-1}(\theta_\lambda) \int \xi' (x) e^{\theta_\lambda(x)+u(x)} v(x) w(x) S(x) dx||_{\lambda c}^2 \\ &\qquad\quad + ||G_\lambda^{-1}(\theta_\lambda) \int \xi' (x) e^{\theta_\lambda(x)+u(x)} v(x) w(x) |S_n(x) - S(x)| dx||_{\lambda c}^2 . \end{aligned}$$

Again the argument of Theorem 5.1 gives $O(\lambda^{-(c+d/2m)}) R^2$ for the first term. Writing out the second term, we have, by manipulations identical to those above

$$\begin{aligned} &||G_\lambda^{-1}(\theta_\lambda) \int \xi' (x) e^{\theta_\lambda(x)+u(x)} v(x) w(x) |S_n(x) - S(x)| dx||_{\lambda c}^2 \\ &\qquad = \sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu}^c][1+2\lambda\gamma_{\lambda\nu}]^{-2} \{\int \phi_{\lambda\nu}(x) e^{\theta_\lambda(x)+u(x)} v(x) w(x) |S_n(x) - S(x)| dx\}^2 \end{aligned}$$

An analysis of this term gives in $k_n \lambda^{-(c+d/2m)} R^2$ uniformly in λ. Since $k_n \xrightarrow{p} 0$ by B(i), the second term is negligible by comparison to the first and the result follows. *Q.E.D.*

Combining Theorem 5.2 with Theorem A2.1 (which gives stochastic bounds on $||\overline{\theta}_{n\lambda} - \theta_\lambda||_{\lambda c}^2$) we obtain the following Corollary.

Corollary 5.3 *Under the hypotheses of Theorem 5.2, the sequences* $d(n,\lambda)$, $r^*(n,\lambda)$ *and* $r(n,\lambda)$ *have the following behavior uniformly for* $\lambda \in (0,\lambda_0]$

(i) $\qquad d^2(n,\lambda) \approx n^{-1} \lambda^{-(c^*+d/2m)}$

(ii) $\qquad r^{*2}(n,\lambda) \approx n^{-1} \lambda^{-2(c^*+d/m)}$

(iii) $\qquad r^2(n,\lambda) \approx n^{-1} \lambda^{-2(c^*+d/m)} \lambda^{c^*-c}$

Thus if $n^{-1}\lambda^{-2(c^*+d/m)} \to 0$ *the asymptotic behavior of* $\hat{\theta}_{n\lambda}$ *and* $\bar{\theta}_{n\lambda}$ *is the same.*

Proof Theorem A2.1 gives (i) while (ii) and (iii) follow from Theorem 5.2. *Q.E.D.*

Finally we turn to the generalized linear regression model. Much of the technique used in our next theorem is taken from Theorem A2.2. For this reason and in order to keep the proof more brief, we shall omit details of arguments which are very similar to those used in Theorem A2.2.

Theorem 5.4. *Let* λ_0 *be such that Theorem 4.1 holds, and suppose* $\{\lambda_n : n \geq 1\}$ *is such that*

$$k_n \lambda_n^{-d/m} << 1 \tag{5.7}$$

$$k_n^2 \lambda_n^{-d/m} << n^{-1} \quad . \tag{5.8}$$

Let $d/m < c^* < 2\text{–}3d/2m$, *and* $0 \leq c \leq c^*$, *then*

(a) $\sup_{\|u\|^2_{\lambda c^*}=1} \|G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_\lambda(\theta_\lambda))u\|^2_{\lambda c} \leq T_1(n,\lambda)$ where $E|T_1(n,\lambda)| \leq n^{-1}\lambda^{-(c+d/2m)}$.

(b) $\sup_{\substack{\|u\|^2_{\lambda c^*} \leq R^2, \|v\|^2_{\lambda c^*} \leq R^2 \\ \|w\|^2_{\lambda c^*} \leq 1}} \|G_\lambda^{-1}(\theta_\lambda) D^3 l_{n\lambda}(\theta_\lambda + u)vw\|^2_{\lambda c} \leq T_2(n,\lambda)R^2$ where

$E|T_2(n,\lambda)| \leq \lambda^{-(c+d/2m)}$.

The bounds on the expectations being uniform for $\lambda \in [\lambda_n, \lambda_0]$.

Remark: Assumption C(ii) guarantees $3d/2m < 2\text{–}3d/2m$, so the assumption on c^* is not vacuous.

Proof: We begin by sketching the proof of part(a).

Letting $G_{n\lambda}(\theta_\lambda)\varsigma = EDZ_{n\lambda}(\theta_\lambda)\varsigma$

$$= \int_X \xi'(x)\dot{\bar{\psi}}(x, \theta_\lambda(x))\varsigma(x) P_X^{(n)}(dx) + 2\lambda W\varsigma \qquad \text{for } \varsigma \in \Theta_{\lambda c^*},$$

for any $u \in \Theta_{\lambda c^*}$ we have

$$\|G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_\lambda(\theta_\lambda))u\|^2_{\lambda c} \leq \|G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_{n\lambda}(\theta_\lambda))u\|^2_{\lambda c} + \|G_\lambda^{-1}(\theta_\lambda)(G_{n\lambda}(\theta_\lambda) - G_\lambda(\theta_\lambda))u\|^2_{\lambda c} \quad . \tag{5.9}$$

Here the second term can be analyzed, as $\|G_\lambda^{-1}(\theta_\lambda)(\bar{Z}_{n0}(\theta_\lambda) - Z_0(\theta_\lambda)\|^2_{\lambda c}$ in Theorem A2.2, to give

$$\| G_\lambda^{-1}(\theta_\lambda)(G_{n\lambda}(\theta_\lambda) - G_\lambda(\theta_\lambda))u \|_{\lambda c}^2 = \| \sum_{j=1}^{q} \sum_{k=1}^{q} \int_X G_\lambda^{-1}(\theta_\lambda)\xi_j(x)\dot{\bar{\psi}}_{jk}(x, \theta_\lambda(x))u_k(x)[P_X^{(n)} - P_X](dx) \|_{\lambda c}^2$$

$$= \sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu}^c][1+2\lambda\gamma_{\lambda\nu}]^{-2}$$

$$\cdot \{\int \phi_{\lambda\nu}'(x)\dot{\bar{\psi}}(x, \theta_\lambda(x))u(x)[P_X^{(n)} - P_X](dx)\}^2$$

Since $\dot{\bar{\psi}} \in C^d(\bar{X} \times S(R, \mathbf{R}^q); \mathbf{R}^{q \times q})$ and $c^* \geq d/m$ so $u \in W_2^d(\bar{X}, \mathbf{R}^q)$, we obtain the following bound by means of Lemma 4.2 of Cox[9]

$$| \int_X \phi_{\lambda\nu}'(x)\dot{\bar{\psi}}(x, \theta_\lambda(x))u(x)[P_X^{(n)} - P_X](dx) | \leq k_n \| \phi_{\lambda\nu} \mid W_2^d \| \cdot \| u \mid W_2^d \|.$$

But, from Lemma A1.2 and A1.3(i), $\| \phi_{\lambda\nu} \mid W_2^d \|^2 \leq 1 + \gamma_\lambda^{d/m}$, and this implies, by (5.8), that

$$\| G_\lambda^{-1}(\theta_\lambda)(G_{n\lambda}(\theta_\lambda) - G_\lambda(\theta_\lambda))u \|_{\lambda c}^2 \leq k_n^2 \lambda^{-(c+3d/2m)} \| u \|_{\lambda c}^2.$$

$$\leq n^{-1}\lambda^{-(c+d/2m)} \| u \|_{\lambda c}^2. \quad .$$

The first term on the r.h.s. of (5.9) is stochastic and to analyze it we use a technique similar to that in Huber[14], pp. 166-167 . Let u and v be unit vectors in $\Theta_{\lambda c}$ and $\Theta_{\lambda c^*}$ respectively.

$u = \sum_{\nu=1}^{\infty} u_\nu \phi_{\lambda\nu}$, $v = \sum_{\nu^*=1}^{\infty} v_{\nu^*} \phi_{\lambda\nu^*}$, where $\sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu}^c] u_\nu^2 = \sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu^*}^{c^*}] v_{\nu^*}^2 = 1.$

Now

$$< u, G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_{n\lambda}(\theta_\lambda))v >_{\lambda c} = \sum_{\nu=1}^{\infty} \sum_{\nu^*=1}^{\infty} u_\nu [1+\gamma_{\lambda\nu}^c] < \phi_{\lambda\nu}, U(\theta_\lambda) \int G_\lambda^{-1}(\theta_\lambda)\xi'(x)$$

$$\cdot (\dot{\psi}(y \mid x, \theta_\lambda(x)) - \dot{\bar{\psi}}(x, \theta_\lambda(x)))$$

$$\cdot v_{\nu^*} \phi_{\lambda\nu^*}(x) P_{XY}^{(n)}(dxdy) >$$

$$= \sum_{\nu=1}^{\infty} \sum_{\nu^*=1}^{\infty} u_\nu [1+\gamma_{\lambda\nu}^c]^{\frac{1}{2}} \Delta_{\nu\nu^*} [1+\gamma_{\lambda\nu^*}^{c^*}]^{\frac{1}{2}} v_{\nu^*} \qquad (5.10)$$

where $\Delta_{\nu\nu^*}$ is explicitly given by

$$\Delta_{\nu\nu^*} = \sum_{k=1}^{q} \sum_{j=1}^{q} [1+\gamma_{\lambda\nu}^c]^{\frac{1}{2}} [1+\gamma_{\lambda\nu^*}^{c^*}]^{-\frac{1}{2}} \int < \phi_{\lambda\nu}, U(\theta_\lambda) G_\lambda^{-1}(\theta_\lambda)\xi_j(x) >$$

$$\cdot (\dot{\psi}_{jk}(y \mid x, \theta_\lambda(x)) - \dot{\bar{\psi}}_{jk}(x, \theta_\lambda(x))) \phi_{\lambda\nu^* k}(x) P_{XY}^{(n)}(dxdy)$$

Applying the Cauchy Schwartz inequality to (5.10) gives

$$<u, G_\lambda^{-1}(\theta_\lambda)(I_{n\lambda} - G_{n\lambda}(\theta_\lambda))v>_{\lambda c}^2 \leq \sum_{\nu=1}^{\infty}[1+\gamma_{\lambda\nu}^c]u_\nu^2 \sum_{\nu=1}^{\infty}[1+\gamma_{\lambda\nu^*}^{c^*}]v_{\nu^*}^2 \sum_{\nu\nu^*}\Delta_{\nu\nu^*}^2$$

$$= \sum_{\nu\nu^*}\Delta_{\nu\nu^*}^2 .$$

We will now show that $\sum_{\nu\nu^*}E[\Delta_{\nu\nu^*}^2] \leq n^{-1}\lambda^{-(c+d/2m)}$ uniformly in $\lambda \in [\lambda_n, \lambda_0]$. And so with

$$T_1(n,\lambda) = [\sum_{\nu\nu^*}\Delta_{\nu\nu^*}^2 + k_n^2\lambda^{-(c+3d/2m)}]$$

$$E|T_1(n,\lambda)| \leq [n^{-1}\lambda^{-(c+d/2m)} + k_n^2\lambda^{-(c+3d/2m)}]$$

$$\approx n^{-1}\lambda^{-(c+d/2m)} \quad \text{by (5.8),}$$

part(a) is established.

$$E[\Delta_{\nu\nu^*}^2] = \sum_{j,k,j',k'=1}^{q}[1+\gamma_{\lambda\nu}^c]\,[1+\gamma_{\lambda\nu^*}^{c^*}]^{-1}\,\frac{1}{n}\int <\phi_{\lambda\nu}, U(\theta_\lambda)G_\lambda^{-1}(\theta_\lambda)\xi_j(x)>$$

$$<\phi_{\lambda\nu}, U(\theta_\lambda)G_\lambda^{-1}(\theta_\lambda)\xi_{j'}(x)> \cdot \tau_{jkj'k'}(x,\theta_\lambda(x)) \cdot \phi_{\lambda\nu^*k}(x)\phi_{\lambda\nu^*k'}(x)P_X^{(n)}(dx)$$

Summing over ν and ν^* gives the expression

$$\sum_{\nu\nu^*}E[\Delta_{\nu\nu^*}^2] = \sum_{j,k,j',k'=1}^{q}\frac{1}{n}\int <G_\lambda^{-1}(\theta_\lambda)\xi_j(x), G_\lambda^{-1}(\theta_\lambda)\xi_{j'}(x)>_{\lambda c}$$

$$\tau_{jkj'k'}(x,\theta_\lambda(x)) \cdot \sum_{\nu^*=1}^{\infty}\phi_{\lambda\nu^*k}(x)\phi_{\lambda\nu^*k'}(x)[1+\gamma_{\lambda\nu^*}^{c^*}]^{-1}P_X^{(n)}(dx)$$

The result will follow once it is established that $P_X^{(n)}$ may be replaced by P_X in the r.h.s. We show that the above expression is $\approx n^{-1}\lambda^{-(c+d/2m)}$, with error $<< n^{-1}\lambda^{-(c+d/2m)}$ and this will establish the result. Assuming the replacement of the discrete measure, $P_X^{(n)}$, by the continuous analogue P_X is alright, the $n^{-1}\lambda^{-(c+d/2m)}$ bound is obtained by an argument similar to that used in Theorem A2. The crucial step is to show

$$\int_X \phi_{\lambda\nu j}(x)\phi_{\lambda\nu j'}(x)\tau_{jkj'k'}(x,\theta_\lambda(x))\phi_{\lambda\nu^*k}\phi_{\lambda\nu^*k'}P_X(dx) \leq M_0(R)\|\phi_{\lambda\nu}\|_{\lambda 0}^2\|\phi_{\lambda\nu^*}\|_{\lambda 0}^2 .$$

But this follows from B(ii) and D(vi). Thus using this bound, Lemma A1.3(i) and the analysis in Theorem A2 we can get

$$\sum_{\nu\nu^*}E[\Delta_{\nu\nu^*}^2] \approx \sum_{j,k,j',k'=1}^{q}\frac{1}{n}\int <G_\lambda^{-1}(\theta_\lambda)\xi_j(x), G_\lambda^{-1}(\theta_\lambda)\xi_{j'}(x)>_{\lambda c}$$

$$\tau_{jkj'k'}(x,\theta_\lambda(x)) \cdot \sum_{\nu^*=1}^{\infty}\phi_{\lambda\nu^*k}(x)\phi_{\lambda\nu^*k'}(x)[1+\gamma_{\lambda\nu^*}^{c^*}]^{-1}P_X(dx)$$

$$\leq \sum_{\nu=1}^{\infty}[1+\gamma_{\lambda\nu}^{c}][1+2\lambda\gamma_{\lambda\nu}]^{-2}\cdot\sum_{\nu^{\bullet}=1}^{\infty}[1+\gamma_{\lambda\nu^{\bullet}}^{c^{\bullet}}]^{-1}$$

However, from Lemma A1.1(a), $\gamma_{\lambda\nu}\approx\nu^{2m/d}$, and since $c^{\bullet}>d/2m$, the sum over $\nu^{\bullet}$ is finite. Hence, by Lemma A1.1(c)

$$\sum_{\nu\nu^{\bullet}}E[\Delta_{\nu\nu^{\bullet}}^{2}]\lesssim n^{-1}\lambda^{-(c+d/2m)}\quad.$$

It remains to justify the replacement of $P_X^{(n)}$ by P_X. Employing the integration by parts formula in Lemma 4.2 of Cox[9] and assumption D(vii), we obtain, by arguments similar to those used for equation (A2.8) in Theorem A2, that

$$\begin{aligned}&\Big|\int<G_\lambda^{-1}(\theta_\lambda)\xi_j(x),G_\lambda^{-1}(\theta_\lambda)\xi_{j'}(x)>_{\lambda c}\\&\quad\cdot\tau_{jkj'k'}(x,\theta_\lambda(x))\cdot\sum_{\nu^{\bullet}=1}^{\infty}\phi_{\lambda\nu^{\bullet}k}(x)\phi_{\lambda\nu^{\bullet}k'}(x)[1+\gamma_{\lambda\nu^{\bullet}}^{c^{\bullet}}]^{-1}[P_X^{(n)}-P_X](dx)\Big|\end{aligned}\tag{5.11}$$

$$\leq K\{\sup_x|F_n(x)-F(x)|\}\,\|\tau\mid C^d(\bar{X}\times S(R,\mathbb{R}^q);\mathbb{R}^{q\times q\times q}\|\cdot\|\theta_\lambda\mid C^d(\bar{X};\mathbb{R}^q)\|\cdot$$

$$\begin{aligned}&\Big\{\sum_{\beta\in\{0,1\}^d}\int|D_x^{\beta}|<G_\lambda^{-1}(\theta_\lambda)\xi_j(\cdot),G_\lambda^{-1}(\theta_\lambda)\xi_k(\cdot)>_{\lambda c}\\&\quad\cdot\sum_{\nu^{\bullet}=1}^{\infty}\phi_{\lambda\nu^{\bullet}k}(\cdot)\phi_{\lambda\nu^{\bullet}k'}(\cdot)[1+\gamma_{\lambda\nu^{\bullet}}^{c^{\bullet}}]^{-1}](x[\beta])\mid dx\Big\}\end{aligned}$$

where $x[\beta]$ has components

$$x_i[\beta]=\begin{cases}x_i & \text{if}\quad\beta_i=1\\ A & \text{if}\quad\beta_i=0\end{cases},$$

and $A\in(0,\infty)$ is chosen so that $X\subseteq[-A,A]^d$; see Assumption A(i). The integrand in the last term is written as

$$\begin{aligned}&D_x^{\beta}|<G_\lambda^{-1}(\theta_\lambda)\xi_j(x),G_\lambda^{-1}(\theta_\lambda)\xi_k(x)>_{\lambda c}\sum_{\nu^{\bullet}=1}^{\infty}\phi_{\lambda\nu^{\bullet}k}(x)\phi_{\lambda\nu^{\bullet}k'}(x)[1+\gamma_{\lambda\nu^{\bullet}}^{c^{\bullet}}]^{-1}|\\&\quad=D_x^{\beta}\sum_{\nu=1}^{\infty}\sum_{\nu^{\bullet}=1}^{\infty}[1+\gamma_{\lambda\nu}^{c}][1+2\lambda\gamma_{\lambda\nu}]^{-2}[1+\gamma_{\lambda\nu^{\bullet}}^{c^{\bullet}}]^{-1}\phi_{\lambda\nu j}(x)\phi_{\lambda\nu j'}(x)\phi_{\lambda\nu^{\bullet}k}(x)\phi_{\lambda\nu^{\bullet}k'}(x)\\&\quad=\sum_{\nu=1}^{\infty}\sum_{\nu^{\bullet}=1}^{\infty}[1+\gamma_{\lambda\nu}^{c}][1+2\lambda\gamma_{\lambda\nu}]^{-2}[1+\gamma_{\lambda\nu^{\bullet}}^{c^{\bullet}}]^{-1}D_x^{\beta}[\phi_{\lambda\nu j}(x)\phi_{\lambda\nu j'}(x)\phi_{\lambda\nu^{\bullet}k}(x)\phi_{\lambda\nu^{\bullet}k'}(x)]\end{aligned}$$

(Again the argument justifying the interchange of D_x^{β} and $\sum_{\nu=1}^{\infty}\sum_{\nu^{\bullet}=1}^{\infty}$ parallels that used in Theorem A2). Utilizing the product rule for differentiation, we have

$$\int | D_x^{\beta}[\phi_{\lambda\nu_j}(x)\phi_{\lambda\nu_j'}(x)\phi_{\lambda\nu^*k}(x)\phi_{\lambda\nu^*k'}(x)](x\,[\beta]) | \, dx$$

$$= \sum_{\substack{\beta_1,\beta_2,\beta_3,\beta_4 \\ \beta_1+\beta_2+\beta_3+\beta_4=\beta}} \int | [(D_x^{\beta_1}\phi_{\lambda\nu_j})(D_x^{\beta_2}\phi_{\lambda\nu_j'})(D_x^{\beta_3}\phi_{\lambda\nu^*k})(D_x^{\beta_4}\phi_{\lambda\nu^*k'})](x\,[\beta]) | \, dx$$

$$\leq K \sum_{\beta_1+\beta_2+\beta_3+\beta_4=\beta} \| \phi_{\lambda\nu_j} \,|\, C^{|\beta_1|} \| \cdot \| \phi_{\lambda\nu_j'} \,|\, C^{|\beta_2|} \|$$

$$\cdot \int | (D_x^{\beta_3}\phi_{\lambda\nu^*k})(D_x^{\beta_4}\phi_{\lambda\nu^*k'})(x\,[\beta]) | \, dx$$

$$\leq K(\epsilon) \sum_{\beta_1+\beta_2+\beta_3+\beta_4=\beta} \| \phi_{\lambda\nu} \,|\, W_2^{|\beta_1|+d/2+\epsilon} \| \| \phi_{\lambda\nu} \,|\, W_2^{|\beta_2|+d/2+\epsilon} \|$$

$$\cdot \| \phi_{\lambda\nu^*} \,|\, W_2^{|\beta_3|+(d-|\beta|)/2+\epsilon} \| \cdot \| \phi_{\lambda\nu^*} \,|\, W_2^{|\beta_4|+(d-|\beta|)/2+\epsilon} \| \,,$$

where $\epsilon > 0$ is arbitrary. The bounds on $\phi_{\lambda\nu}$ in this last in equality follow from Sobolev's imbedding theorem, and the bounds on $\phi_{\lambda\nu^*}$ follow from the Cauchy-Schwartz inequality and Sobolev's theorem on traces (Eqn. (2), p 97 of Adams[1] or Theorem 4.7.2 of Triebel[28]). Replacing the Sobolev norms by equivalent Θ_{sb} norms and using the fact that $\beta \in \{0,1\}^d$, we see that the last displayed quantity is

$$\lesssim \sum_{l=0}^{|\beta|} [1+\gamma_{\lambda\nu}^{(d+l+\epsilon)/2m}] \, [1+\gamma_{\lambda\nu^*}^{(d-l+\epsilon)/2m}]$$

Now returning to Eqn. (5.11) and using $|\beta| \leq d$ and bounds on $\|\tau \,|\, C^d\|$ and $\|\theta_\lambda \,|\, C^d\|$, we obtain that the l.h.s of (5.11) is

$$\lesssim k_n \sum_{\nu} [1+\gamma_{\lambda\nu}^{c}][1+\lambda\gamma_{\lambda\nu}]^{-2}[1+\gamma_{\lambda\nu}^{(2d+\epsilon)/2m}] \cdot \sum_{\nu^*} [1+\gamma_{\lambda\nu^*}^{c^*}]^{-1}[1+\gamma_{\lambda\nu^*}^{(d+\epsilon)/2m}]$$

Now $\gamma_{\lambda\nu^*} \approx (\nu^*)^{2m/d}$, so with $c^* > d/m$, the $\sum$ over ν^* is finite for some $\epsilon > 0$. Moreover, since $c + d/m < 2 - d/2m$, Lemma A1.1(c) implies that the sum over ν is $\approx \lambda^{-c-3d/2m}$. Hence the bound in the last equation becomes

$$\leq k_n \lambda^{-c-3d/2m} = k_n \lambda^{-d/m} \cdot \lambda^{-(c+d/2m)} << \lambda^{-(c+d/2m)} \quad \text{by (5.7)} \,.$$

Thus replacing $P_X^{(n)}$ by P_X is an $o(n^{-1}\lambda^{-(c+d/2m)})$ approximation, establishing part(a) of the Theorem.

We now outline the proof of part(b). For this let u, v, $w \in \Theta_{\lambda c}$. with $\|u\|^2_{\lambda c} \leq R$, $\|v\|^2_{\lambda c} \leq R$, and $\|w\|^2_{\lambda c} \leq 1$.

$$\|G_\lambda^{-1}(\theta_\lambda)D^3 l_{n\lambda}(\theta_\lambda+u)vw\|^2_{\lambda c} = \| \sum_{j,k,l=1}^{q} \int G_\lambda^{-1}(\theta_\lambda)\xi_j(x)\ddot{\psi}_{jkl}(y \mid x, \theta_\lambda(x)+u(x)) \cdot v_k(x)w_l(x)P_{XY}^{(n)}(dxdy)\|^2_{\lambda c} \quad .$$

For convenience let $\bar{\theta}_\lambda = \theta_\lambda + u$. By straightforward algebra, the Cauchy-Schwartz inequality, B(ii), and Lemma A1.3(ii), we have that the last quantity is

$$\leq K \{\sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu}^c][1+2\lambda\gamma_{\lambda\nu}]^{-2} \int \|\phi_{\lambda\nu}(x) \mid \mathbb{R}^q\|^2 P_X^{(n)}dx\}$$
$$\cdot \int \|\ddot{\psi}(y \mid x, \bar{\theta}_\lambda(x)) \mid \mathbb{R}^{q\times q\times q}\|^2 P_{XY}^{(n)}(dxdy)$$
$$\cdot \|v \mid C(\underline{X}, \mathbb{R}^q)\|^2 \|w \mid C(\underline{X}, \mathbb{R}^q)\|^2 \quad .$$

But $\phi_{\lambda\nu}$, v, and $w \in \Theta_{\lambda c}$. and since $c' > d/m$, $\Theta_{\lambda c} \subseteq W_2^d(\bar{\underline{X}}, \mathbb{R}^q)$, the replacement of $P_X^{(n)}$ by P_X can be carried out as before to give

$$\leq K \{\sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu}^c][1+2\lambda\gamma_{\lambda\nu}]^{-2} \int \|\phi_{\lambda\nu}(x) \mid \mathbb{R}^q\|^2 P_X(dx)\}$$
$$\cdot \int \|\ddot{\psi}(y \mid x, \bar{\theta}_\lambda(x)) \mid \mathbb{R}^{q\times q\times q}\|^2 P_{XY}^{(n)}(dxdy)$$
$$\cdot \|v \mid C(\underline{X}, \mathbb{R}^q)\|^2 \|w \mid C(\underline{X}, \mathbb{R}^q)\|^2 \quad .$$

This in turn, utilizing the argument in (5.6), is bounded as

$$\leq K \lambda^{-(c+d/2m)} \int \|\ddot{\psi}(y \mid x, \bar{\theta}_\lambda(x)) \mid \mathbb{R}^{q\times q\times q}\|^2 P_{XY}^{(n)}(dxdy)$$
$$\cdot \|v\|^2_{\lambda c} \cdot \|w\|^2_{\lambda c} \cdot$$
$$\leq K \lambda^{-(c+d/2m)} \int \|\ddot{\psi}(y \mid x, \bar{\theta}_\lambda(x)) \mid \mathbb{R}^{q\times q\times q}\|^2 P_{XY}^{(n)}(dxdy) R^2$$

Thus, by D(viii), for some R', depending only on λ_0 and R, with

$$T_2(n,\lambda) = K \lambda^{-(c+d/2m)} \int \sup_{t\in S(R', \mathbb{R}^q)} \|\ddot{\psi}(y \mid x, t) \mid \mathbb{R}^{q\times q\times q}\|^2 P_{XY}^{(n)}(dxdy)$$
$$\mathrm{E}\,|T_2(n,\lambda)| \leq K M_7^2(R') \lambda^{-(c+d/2m)}$$

Which completes the proof of (b). *Q.E.D.*

Combining Theorem 5.4 with Theorem A2.2 (which gives bounds on $E||\bar{\theta}_{n\lambda} - \theta_\lambda||^2_{\lambda c}$.) we obtain the following Corollary.

Corollary 5.5 *Under the hypotheses of Theorem 5.4, the sequences* $d(n,\lambda)$, $r^*(n,\lambda)$ *and* $r(n,\lambda)$ *have the following behavior for* $\lambda \in [\lambda_n, \lambda_0]$

(i) $$d^2(n,\lambda) \approx n^{-1}\lambda^{-(c^*+d/2m)}$$

(ii) $$r^{*2}(n,\lambda) \approx n^{-1}\lambda^{-2(c^*+d/2m)}$$

(iii) $$r^2(n,\lambda) \approx n^{-1}\lambda^{-2(c^*+d/2m)}\lambda^{c^*-c}$$

Thus if $n^{-1}\lambda^{-2(c^*+d/2m)} \to 0$ *the asymptotic behavior of* $\hat{\theta}_{n\lambda}$ *and* $\bar{\theta}_{n\lambda}$ *is the same.*

Proof The results are obtained by Markov's inequality. (i) follows from Theorem A2.2, while (ii) and (iii) follow from Theorem 5.4. *Q.E.D.*

Remark 1. In order for Corollary 5.5 to be uniform in $\lambda \in [\lambda_n, \lambda_0]$ a more careful analysis than that in Theorem 5.4 or Theorem A2 would be required.

A1. The Penalty-Information Scale of Hilbert Spaces

Following (5.1), let U be an operator valued map $U : \Theta \to B(\Theta,\Theta)$ given by

$$U(\theta)\varsigma = \int \xi'(x)h(x,\theta(x))\varsigma(x)dx$$

where $h : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^{q\times q}$ satisfies

For all $R > 0$, there are positive constants $M_1(R)$ and $M_2(R)$ such that for all $x \in X$, $t \in S(R,\mathbb{R}^q)$,

$$M_1(R) \leq u' h(x,t)u \leq M_2(R) \tag{A1.1}$$

In section 5, we indicated that the U operators, $U(\theta) = D^2 l_0$, associated with the the penalized likelihoods of section 2 satisfied the above condition. We now state and prove certain technical results on the P.I. scale of Hilbert spaces associated with U and the generalized smoothing operator W - see (2.7). Assumptions A through C are in force throughout. Note that U is an operator valued map $U : \Theta \to B(\Theta,\Theta)$. The existence of the Θ valued integral which defines $U(\theta)\xi$ is easy to establish.

Lemma A1.1 *Fix $R > 0$.*

(a) *For all $\theta_\bullet \in S(R,C(X;\mathbb{R}^q))$, there exists $\{\phi_{\bullet\nu} : \nu = 1,2,\cdots\} \subseteq \Theta$ and $\{\gamma_{\bullet\nu} : \nu = 1,2,\cdots\} \subseteq [0,\infty)$ such that*

$$<\phi_{\bullet\nu}, U(\theta_\bullet)\phi_{\bullet\mu}> = \delta_{\nu\mu}$$
$$<\phi_{\bullet\nu}, W\phi_{\bullet\mu}> = \gamma_{\bullet\nu}\delta_{\nu\mu}$$

(b) *The eigenvalues in part (a) satisfy*

$$\gamma_{\bullet\nu} \approx \nu^{2m/d}$$

as $\nu \to \infty$ uniformly in $\theta_\bullet \in S(R,C)$.

(c) *If $b \geq 0$ and $c \geq 0$ are such that*

$$b + c < 2 - d/2m$$

then

$$\sum_{\nu}(1+\gamma_{\theta\nu}^{b})(1+\gamma_{\theta\nu}^{c})(1+2\lambda\gamma_{\theta\nu})^{-2} \approx \lambda^{-(b+c+d/2m)}$$

as $\lambda \to 0$ *uniformly in* $\theta_\theta \in S(\mathrm{R},\mathrm{C})$.

Proof. (a) Let $\Theta_{\theta 0}$ denote $L_2(\bar{X})$ equipped with the norm

$$\begin{aligned}||\theta \mid \Theta_{\theta 0}|| &= \{\int \theta'(x)h(x,\theta_\theta(x))\theta(x)P_x(dx)\}^{\frac{1}{2}} \\ &= <\theta, U(\theta_\theta)\theta>^{\frac{1}{2}}.\end{aligned}$$

It follows from (A1.1) and the definition of $U(\theta_\theta)$ in (2.4) that $\exists\, K_1(R)$ and $K_2(R)$ such that *for all* $\theta_\theta \in S(R,C)$,

$$\begin{aligned}K_1(R)||\theta \mid L_2(\underline{X};\mathbb{R}^q)|| &\le ||\theta \mid \Theta_{\theta 0}|| \\ &\le K_2(R)||\theta \mid L_2(\underline{X};\mathbb{R}^q)|| \quad , \text{ for all } \theta \in L_2(\underline{X}).\end{aligned} \tag{A1.2}$$

Now let T denote the imbedding (injection operator) of Θ into $\Theta_{\theta 0}$. Note that T is the composition of the imbedding of Θ into W_2^m (see C(ii)) and the imbedding of W_2^m into $\Theta_{\theta 0}$, and the latter is compact by the norm equivalence (A1.2) and the easily checked compactness of the imbedding of W_2^m into L_2. Hence, T is compact, and so T^*T is compact.

One easily checks that the adjoint operator $T^* : \Theta_{\theta 0} \to \Theta$ is given by

$$T^*\theta = \int \xi'(x)h(x,\theta_\theta(x))\theta(x)P_x(dx)$$

and so $T^*T = U(\theta_\theta)$. The existence of the eigensystem $\phi_{\theta\nu}$, $\gamma_{\theta\nu}$ follows from section 3.3 of Weinberger[31] and the construction in Proposition 2.2 of Cox[8].

For part (b), let $\{\bar{\phi}_\nu\}$ and $\{\bar{\gamma}_\nu\}$ be the eigenfunctions and eigenvalues satisfying

$$\begin{aligned}\int_X \bar{\phi}'_\nu(x)\bar{\phi}_\mu(x)dx &= \delta_{\nu\mu} \\ \sum_{|\alpha|=m}\binom{m}{\alpha}\int_X D_x^\alpha \bar{\phi}'_\nu(x) D_x^\alpha \bar{\phi}_\mu(x)dx &= \bar{\gamma}_\nu\delta_{\nu\mu}\end{aligned} \tag{A1.3}$$

It follows from (A1.1) that the quadratic forms

$$\begin{aligned}B_\theta(\theta,\theta) &= <\theta, U(\theta_\theta)\theta> \\ \bar{B}(\theta,\theta) &= <\theta,\theta \mid L_2(\bar{X};\mathbb{R}^q)>\end{aligned}$$

$\{B_i : 0 \le i \le m\}$ is a normal system (Definition 4.3.3.1 of Triebel[28]); (iii) if $d = 1$ then $B_i = (d/dx)^{m+i}$; (iv) Assuming $bm - 1/2$ is not an integer, Θ_{*b} is the closed subspace of W_2^{bm} given by $\{\theta \in W_2^{bm} : B_i\theta \equiv 0$ on ∂X for all $i < (b-1)m - 1/2\}$, and (A1.4) holds.

Proof. The result for $b = 0$ is immediate from (A1.1). For $b = 1$, it follows from the assumptions together with C(ii) and C(iv). For $0 < b < 1$, we apply the K-method of interpolation as defined in Triebel[28] ; see also Cox[8]. Let

$$K_*^2(u,\theta) = inf\{u^2 <\theta_1, W\theta_1> + <\theta-\theta_1, U(\theta_*)(\theta-\theta_1)> : \theta \in \Theta\}$$

$$K^2(u,\theta) = inf\{u^2 <\theta_1, W\theta_1> + ||\theta-\theta_1 \mid L_2||^2 : \theta \in \Theta\} .$$

A straightforward calculation with the expansion in terms of $\phi_{*\nu}$'s yields

$$||\theta||_{*b}^2 = c(b)\int_0^\infty K_*^2(u,\theta)u^{-(2b+1)}du + <\theta, U(\theta_*)\theta> \qquad \text{(A1.5)}$$

for $0 < b < 1$, where

$$c(b) = \{\int_0^\infty u^{1-2b}(1+u^2)^{-1}du\}^{-1}$$

Let $|||\theta|||_b^2$ be given by

$$|||\theta|||_b^2 = c(b)\int_0^\infty K(u,\theta)u^{-(2b+1)}du + ||\theta \mid L_2||^2 . \qquad \text{(A1.6)}$$

Now $|||\cdot|||_b$ doesn't depend on θ_*, and is equivalent to W_2^{bm} norm when $b = 0,1$ by the same argument as before. For $b \in (0,1)$, the $|||\cdot|||_b$ - norm is obtained by the K-method of interpolation (or equivalent to one so obtained), and so also to W_2^{bm} norm, as it can be obtained by applying the K-method to $W_2^0 = L_2$ and W_2^m (see Theorem 4.3.1.1 of Triebel[28]). From the $b = 0$ case of the lemma, we have *for all* $\theta \in \Theta$

$$K_1(R)K(K_1(R)^{-1}u,\theta) \le K_*(u,\theta) \le K_2(R)K(K_1(R)^{-1}u,\theta)$$

Substituting this into (A1.5) and (A1.6) gives

$$min\{1, K_1(R)^{2(1-b)}\}|||\theta|||_b^2 \le ||\theta||_{*b}^2 \le max\{1, K_2(R)^{2(1-b)}\}|||\theta|||_b^2$$

which completes the proof of (A1.4).

We only sketch the proof of part (b). One shows first that $\Theta_{\bullet 2}$ has the indicated form by an integration by parts (Green's formula) argument combined with duality theory as in Theorem 3.2 of Cox[8]. The boundary operators arise from the multivariate Green's formula as in the proof of Proposition 2.2 (ii) of Cox[9]. One fills in $1<b<2$ using the interpolatory theory of Besov spaces with boundary conditions as in Proposition 3.1 of Cox[9] or Theorem 3.4 (b) of Cox[8]. ***Q.E.D.***

In the next result, we collect some useful facts.

Lemma A1.3. *Let $\theta_\bullet \in S(R,C)$ and $c > 0$.*

(i) $\|\phi_{\bullet\nu}\|_c^2 = 1 + \gamma_{\bullet\nu}^c$ *for* $\nu=1,2,\cdots$

(ii) $G_\lambda^{-1}(\theta_\bullet)U(\theta_\bullet)\phi_{\bullet\nu} = (1+2\lambda\gamma_{\bullet\nu})^{-1}\phi_{\bullet\nu}$ *for* $\nu=1,2,\cdots$ *and* $\lambda > 0$.

(iii) *Let* $\lambda_0 > 0$ *and* $c \leq 1$, *then for all* $x \in X$, $\epsilon > 0$, $1 \leq j \leq q$, *and* $\lambda \in (0,\lambda_0]$,

$$\|G_\lambda^{-1}(\theta_\bullet)\xi_j(x)\|_c^2 \leq K(R,\lambda_0)\lambda^{-(c+(1+\epsilon)d/m)}.$$

Proof. Part (i) is immediate from the definitions. The calculation needed to verify (ii) is given in Cox[8], equation (3.12). For (iii), we have

$$\begin{aligned}\|G_\lambda^{-1}(\theta_\bullet)\xi_j(x)\|_c^2 &= \sum_\nu (1+\gamma_{\bullet\nu}^c)<G_\lambda^{-1}(\theta_\bullet)\xi_j(x), U(\theta_\bullet)\phi_{\bullet\nu}>^2 \\ &= \sum_\nu (1+\gamma_{\bullet\nu}^c)<\xi_j(x), G_\lambda^{-1}(\theta_\bullet)U(\theta_\bullet)\phi_{\bullet\nu}>^2 \\ &= \sum_\nu (1+\gamma_{\bullet\nu}^c)(1+2\lambda\gamma_{\bullet\nu})^{-2}<\xi_j(x),\phi_{\bullet\nu}>^2 \\ &= \sum_\nu (1+\gamma_{\bullet\nu}^c)(1+2\lambda\gamma_{\bullet\nu})^{-2}\phi_{\bullet\nu j}^2(x)\end{aligned} \tag{A1.7}$$

The first equation follows from the definition of $\|\cdot\|_{\bullet c}$, the second from self adjointness of $G_\lambda^{-1}(\theta_\bullet)$ as an element of $B(\Theta,\Theta)$ (note the W is self adjoint by C(iii) and an easy calculation shows $U(\theta_\bullet)$ is self adjoint), the third from part (ii) above, and the fourth from the definition of $\xi_j(x)$ as the Riesz representer of evaluation at x of the j'th component of an element of Θ.

Using Sobolev's imbedding theorem, Lemma A1.2, and part (i), we have for any $\epsilon > 0$

$$\begin{aligned}\phi^2_{\cdot\nu_j}(x) &\leq ||\phi_{\cdot\nu} \mid W_2^{(1+\epsilon)d/2}||^2 \\ &\leq K(R)||\phi_{\cdot\nu}||^2_{\cdot(1+\epsilon)d/2m} \\ &= K(R)(1+\gamma_{\cdot\nu}^{(1+\epsilon)d/2m})\end{aligned}$$

which gives

$$||G_\lambda^{-1}(\theta_\cdot)\xi_j(x)||_c^2 \leq K(R)\sum_\nu (1+\gamma_{\cdot\nu}^c)(1+\gamma_{\cdot\nu}^{(1+\epsilon)d/2m})(1+2\lambda\gamma_{\cdot\nu})^{-2} \quad .$$

Letting $\lambda \to 0$, the last quantity is

$$\leq \lambda^{-(c+(1+\epsilon)d/2m+d/2m)}$$

uniformly in $\theta_\cdot \in \dot{S}(R,C)$, by Lemma A1.1(c). Note that $m > 3d/2$ in C(ii) guarantees that $c+(1+\epsilon)d/2m < 2-d/2m$ for some $\epsilon > 0$ and all $c \in [0,1]$. The result follows from this. *Q.E.D.*

Corollary A1.4. *The norms* $||\cdot||_{\lambda c}$ *for* $0 \leq c \leq 1$ *are uniformly equivalent for* $\lambda \in [0,\lambda_0]$, *i.e. if* $\lambda,\lambda' \in [0,\lambda_0]$, *then* $\Theta_{\lambda c} = \Theta_{\lambda' c}$ *and*

$$K_1||\theta||_{\lambda c} \leq ||\theta||_{\lambda' c} \leq K_2||\theta||_{\lambda c}$$

for all $\theta \in \Theta_{\lambda c}$, *where the constants* K_1, *and* K_2 *do not depend on* λ, c, *or* θ.

A2. Rates of Convergence of the Linearized Estimators

The linearized estimators in section 2

$$\bar{\theta}_{n\lambda} = \theta_\lambda - G_\lambda^{-1}(\theta_\lambda) Z_{n\lambda}(\theta_\lambda) \tag{A2.1}$$

tends to θ_λ as $n \to \infty$. Here we derive stochastic bounds on $||\theta_{n\lambda} - \theta_\lambda||_c^2$ for the generalized smoothing estimators discussed in section 2. Our bounds for the linearized log density and log hazard estimators say $||\bar{\theta}_{n\lambda} - \theta_\lambda||_c^2 \approx k_n^2 \lambda^{-(c+d/2m)}$ uniformly for $\lambda \in [0, \lambda_0]$. The bound obtained for the regression case gives $E||\bar{\theta}_{n\lambda} - \theta_\lambda||_c^2 \approx n^{-1}\lambda^{-(c+d/2m)}$ which is uniform for $\lambda \in [\lambda_n, \lambda_0]$, with some restrictions on the asymptotic behavior of λ_n. Here again norms employed are those associated with *P.I.* scale of Hilbert spaces.

Theorem A2.1 *Let c be given with*

$$0 \leq c < 2 - d/2m$$

where $p = s/m$ is given in C(v). *Then as $n \to \infty$, the linearized log density and hazard estimators introduced in section 2 satisfy*

$$E||\bar{\theta}_{n\lambda} - \theta_\lambda||_c^2 \lesssim k_n^2 \lambda^{-(c+d/2m)}$$

uniformly in $\lambda \in [0, \lambda_0]$. ($d = 1$ for the hazard case).

Proof. For the linearized log density estimator we have

$$\begin{aligned} \bar{\theta}_{n\lambda} - \theta_\lambda = G_\lambda^{-1}(\theta_\lambda) Z_{n\lambda}(\theta_\lambda) &= G_\lambda^{-1}(\theta_\lambda)[Z_{n\lambda}(\theta_\lambda) - Z_\lambda(\theta_\lambda)] \\ &= G_\lambda^{-1}(\theta_\lambda) \int \xi(x)\, [P_X^{(n)} - P_X](dx) \quad . \end{aligned}$$

Taking norms and using Lemma A1.3 (ii) gives

$$||\bar{\theta}_{n\lambda} - \theta_\lambda||_c^2 = \sum_{\nu=1}^{\infty} [1+\gamma_{\lambda\nu}^c][1+2\lambda\gamma_{\lambda\nu}]^{-2} \{\int \phi_{\lambda\nu}(x)[P_X^{(n)} - P_X](dx)\}^2 \quad .$$

Note that $E\{\int \phi_{\lambda\nu}(x)[P_X^{(n)} - P_X](dx)\}^2 = n^{-1}\mathrm{Var}\phi_{\lambda\nu}(X) \leq n^{-1}E\phi_{\lambda\nu}^2(X) \approx n^{-1}||\phi_{\lambda\nu}||_0^2 = n^{-1}$. The result follows from this and Lemma A1.1 (c). An almost identical argument is used for the log hazard estimator. Here

$$\begin{aligned}\bar{\theta}_{n\lambda} - \theta_\lambda &= -G_\lambda^{-1}(\theta_\lambda)[Z_{n0}(\theta_\lambda) - Z_0(\theta_\lambda)] \\ &= -G_\lambda^{-1}(\theta_\lambda)\{\int \xi(x)e^{\theta_\lambda(x)}[S_n(x)-S(x)]dx - \int \xi(x)(P_X^{(n)}-P_X)(dx)\} \\ &= G_\lambda^{-1}(\theta_\lambda)\int \xi(x)(P_X^{(n)}-P_X)(dx) - G_\lambda^{-1}(\theta_\lambda)\int \xi(x)e^{\theta_\lambda(x)}[S_n(x)-S(x)]dx\end{aligned}$$

Thus

$$\begin{aligned}&E||\bar{\theta}_{n\lambda} - \theta_\lambda||_{\lambda c}^2 \\ &\quad = \sum_{\nu=1}^{\infty}[1+\gamma_{\lambda\nu}^c][1+2\lambda\gamma_{\lambda\nu}]^{-2} \\ &\qquad \cdot \mathrm{Var}[n^{-1}\sum_{i=1}^{n}\phi_{\lambda\nu}(X_i)I_{[0,1]}(X_i) + n^{-1}\sum_{i=1}^{n}\int_0^1 \phi_{\lambda\nu}(x)e^{\theta_\lambda(x)}I_{[x,\infty]}(X_i)dx]\end{aligned}$$

The variance in the last expression is

$$\lesssim n^{-1}E[(\phi_{\lambda\nu}(X)I_{[0,1]}(X))^2] + n^{-1}(\int_0^1 \phi_{\lambda\nu}(x)e^{\theta_\lambda(x)}dx)^2 \lesssim n^{-1} \ ,$$

uniformly in λ and ν by a familiar argument. *Q.E.D.*

Finally, for the linearized regression estimator we have the following result.

Theorem A2.2 *Consider the linearized regression estimator. Suppose that the sequence* $\{\lambda_n : n \geq 1\}$ *is such that*

$$k_n\lambda_n^{-d/m} << 1 \tag{A2.2}$$

$$k_n^2\lambda_n^{-d/m} << n^{-1} \ . \tag{A2.3}$$

Let c *be given with*

$$0 \leq c < p$$

where $p = s/m$ *is given in* C(v). *Then as* $n \to \infty$,

$$E||\bar{\theta}_{n\lambda}-\theta_\lambda||_{\lambda c}^2 \approx n^{-1}\lambda^{-(c+d/2m)}$$

uniformly in $\lambda \in [\lambda_n, \lambda_0]$.

Proof. From the definitions,

$$E||\bar{\theta}_{n\lambda}-\theta_\lambda||_{\lambda c}^2 = E||G_\lambda^{-1}(\theta_\lambda)Z_{n\lambda}(\theta_\lambda)||_{\lambda c}^2$$

$$= \mathrm{E}||G_\lambda^{-1}(\theta_\lambda)[Z_{n0}(\theta_\lambda) - Z_0(\theta_\lambda)]||_{\lambda c}^2$$
$$= \mathrm{E}||G_\lambda^{-1}(\theta_\lambda)[(Z_{n0}(\theta_\lambda) - \bar{Z}_{n0}(\theta_\lambda)) + (\bar{Z}_{n0}(\theta_\lambda) - Z_0(\theta_\lambda))]||_{\lambda c}^2$$

where $Z_\lambda(\theta_\lambda) = 0$ was used at the second step, and

$$\bar{Z}_{n0}(\theta) = \mathrm{E}Z_{n\lambda}(\theta) = \int \xi(x)\bar{\psi}(x,\theta(x))P_X^{(n)}(dx) + 2\lambda W\theta \quad .$$

We will show

$$||G_\lambda^{-1}(\theta_\lambda)[\bar{Z}_{n0}(\theta_\lambda) - Z_0(\theta_\lambda)]||_{\lambda c}^2 << \mathrm{E}||G_\lambda^{-1}(\theta_\lambda)[Z_{n0}(\theta_\lambda) - \bar{Z}_{n0}(\theta_\lambda)]||_{\lambda c}^2 \qquad \text{(A2.4)}$$
$$\lesssim \frac{1}{n}\sum_{j=1}^{q}\sum_{k=1}^{q}\int < G_\lambda^{-1}(\theta_\lambda)\xi_j(x), G_\lambda^{-1}(\theta_\lambda)\xi_k(x) >_{\lambda c}$$
$$\kappa_{jk}(x,\theta_\lambda(x),\theta_\lambda(x))P_X(dx)$$
$$\approx n^{-1}\lambda^{-(c+d/2m)}$$

uniformly in $\lambda \in [\lambda_n, \lambda_0]$. This will prove the theorem.

Working backwards through (A2.4), we first show the third relation. The calculation used in (A1.7) can be modified to show that

$$< G_\lambda^{-1}(\theta_\lambda)\xi_j(x), G_\lambda^{-1}(\theta_\lambda)\xi_k(x) >_{\lambda c} = \sum_\nu (1+\gamma_{\lambda\nu}^c)(1+2\lambda\gamma_{\lambda\nu})^{-2}\phi_{\lambda\nu j}(x)\phi_{\lambda\nu k}(x) \quad ,$$

so the l.h.s. of the third relation in (A2.4) is equal to

$$n^{-1}\sum_\nu (1+\gamma_{\lambda\nu}^c)(1+2\lambda\gamma_{\lambda\nu})^{-2}\int \phi'_{\lambda\nu}(x)\kappa(x,\theta_\lambda(x),\theta_\lambda(x))\phi_{\lambda\nu}(x)P_X(dx) \qquad \text{(A2.5)}$$

Assumption D(iii) guarantees that the integrals are bounded above and below by constant multiples of $||\phi_{\lambda\nu} \mid L_2(\bar{X};\mathbb{R}^q)||^2$ with constants that are independent of $\lambda \in (0,\lambda_0]$ and ν. Further, by D(v), we may replace $L_2(\bar{X};\mathbb{R}^q)$ norm by $||\cdot||_{\lambda 0}$ and the statement is still valid, including the uniformity of the constants, as the norms $||\cdot||_{\lambda 0}$ are uniformly equivalent for $\lambda \in [0,\lambda_0]$ by Corollary A1.4. Since $||\phi_{\lambda\nu}||_{\lambda 0}^2 = 2$, the quantity in (A2.5) is

$$\approx n^{-1}\sum_\nu (1+\gamma_{\lambda\nu}^c)(1+2\lambda\gamma_{\lambda\nu})^{-2} \approx n^{-1}\lambda^{-(c+d/2m)} \quad ,$$

where the latter relation follows from Lemma A1.1 (c). This proves the last relation in (A2.4).

Turning now to the middle relation in (A2.4), a straightforward calculation shows

$$E||G_\lambda^{-1}(\theta_\lambda)[Z_{n0}(\theta_\lambda) - \bar{Z}_{n0}(\theta_\lambda)]||_{\lambda c}^2 = \sum_{j=1}^{q}\sum_{k=1}^{q}\int <G_\lambda^{-1}(\theta_\lambda)\xi_j(x), G_\lambda^{-1}(\theta_\lambda)\xi_k(x)>_{\lambda c} \kappa(x,\theta_\lambda(x),\theta_\lambda(x))P_X^{(n)}(dx) ,$$

so it suffices to show that as $n \to \infty$

$$|\int <G_\lambda^{-1}(\theta_\lambda)\xi_j(x), G_\lambda^{-1}(\theta_\lambda)\xi_k(x)>_{\lambda c} \kappa_{jk}(x,\theta_\lambda(x),\theta_\lambda(x))[P_X^{(n)} - P_X](dx)| \ll \lambda^{-(c+d/2m)} \tag{A2.6}$$

uniformly in $\lambda \in [\lambda_n, \lambda_0]$. By the integration by parts formula in the proof of Lemma 4.2(i) of Cox[9], the l.h.s. of (A2.6) is

$$\leq \sum_{\beta \in \{0,1\}^d} |\int D_x^\beta [<G_\lambda^{-1}(\theta_\lambda)\xi_j(\cdot), G_\lambda^{-1}(\theta_\lambda)\xi_k(\cdot)>_{\lambda c} \kappa_{jk}(\cdot,\theta_\lambda(\cdot),\theta_\lambda(\cdot))](x[\beta]) \{F_n(x[\beta]) - F(x[\beta])\}dx| \tag{A2.7}$$

where β is a multi-index with only zeros and ones, D denotes differentiation w.r.t. x (the dot appearing in 5 places), and the derivative is evaluated at $x[\beta]$ with components

$$x_i[\beta] = \begin{cases} x_i & \text{if } \beta_i = 1 \\ A & \text{if } \beta_i = 0 \end{cases}$$

where $A \in (0,\infty)$ is chosen so that $\mathcal{X} \subseteq [-A, A]^d$; see Assumption A(i). The distribution functions F_n and F are defined in Assumption B(i). By applying the product rule and the chain rule for differentiation, one can see that the quantity in (A2.7) is

$$\leq K \{\sup_x |F_n(x) - F(x)|\} ||\kappa | C^d(\bar{\mathcal{X}} \times S(R,\mathbb{R}^q) \times S(R,\mathbb{R}^q); R^{q \times q})|| \cdot ||\theta_\lambda | C^d(\bar{\mathcal{X}}; \mathbb{R}^q)|| \{\max_{\beta \in \{0,1\}^d} \int |D_x^\beta [<G_\lambda^{-1}(\theta_\lambda)\xi_j(x), G_\lambda^{-1}(\theta_\lambda)\xi_k(x)>_{\lambda c}](x[\beta])| dx\} \tag{A2.8}$$

where $K \in (0,\infty)$ is independent of λ, and $R > 0$ is chosen so that $||\theta_\lambda | C|| \leq R$, *for all* $\lambda \in [0,\lambda_0]$. Now

$$D_x^\beta <G_\lambda^{-1}(\theta_\lambda)\xi_j(x), G_\lambda^{-1}(\theta_\lambda)\xi_k(x)>_{\lambda c} = D_x^\beta \sum_\nu (1+\gamma_{\lambda\nu}^c)(1+2\lambda\gamma_{\lambda\nu})^{-2}\phi_{\lambda\nu j}(x)\phi_{\lambda\nu k}(x)$$
$$= \sum_\nu (1+\gamma_{\lambda\nu}^c)(1+2\lambda\gamma_{\lambda\nu})^{-2} D_x^\beta [\phi_{\lambda\nu j}(x)\phi_{\lambda\nu k}(x)] ,$$

where the interchange of D_x^{β} and $\sum_{\nu}$ will be justified below. Utilizing the proof of Lemma 4.2 in Cox[9] again, we have

$$\int_{\bar{X}} | D_x^{\beta}[\phi_{\lambda\nu_j}(\cdot)\phi_{\lambda\nu_k}(\cdot)](x\,[\beta]) | \, dx \leq K \, \|\phi_{\lambda\nu} \mid W_2^d(\bar{X};\mathbb{R}^q)\|^2$$
$$\leq K \, \|\phi_{\lambda\nu}\|_{\lambda b}^2 \quad , \quad b = d/m \quad ,$$

where Assumption C(ii) and Corollary A1.4 are used at the last step. Now by Assumptions B(i) and D(ii), Theorem 4.3 (i) with $c > 3d/2m$ (which implies $\|\theta_\lambda \mid C^d(\bar{X};\mathbb{R}^q)\|$ is uniformly bounded), and Lemmas A1.3 (i) and A1.1 (c), the quantity in (A2.8) is

$$\lesssim K \; k_n \sum_{\nu}(1+\gamma_{\lambda\nu}^c)(1+2\lambda\gamma_{\lambda\nu})^{-2}(1+\gamma_{\lambda\nu}^{d/m}) \qquad \text{(A2.9)}$$
$$\approx k_n \lambda^{-c-3d/2m}$$

In view of (A2.2), this establishes (A2.6), once the claim about interchanging differentiation and summation is proved. In fact, we have shown that the series of derivatives (the last expression in (A2.8)) is absolutely convergent in $L^1(\bar{X};\mathbb{R}^q)$, and thus converges in $L^1(\bar{X};\mathbb{R}^q)$ to something. A standard argument using the mean value theorem and Lebesgue's dominated convergence theorem can now be applied. (The series of absolute values of the derivatives is the dominating function. Note that each $\phi_{\lambda\nu}$ is in $C(\bar{X};\mathbb{R}^q)$ since $\phi_{\lambda\nu} \in \Theta \subseteq W_2^m \subseteq C^d$ by C(ii) and Sobolev's imbedding theorem.)

Finally we consider the first relation in (A2.4). From Lemma A1.3(ii),

$$G_\lambda^{-1}(\theta_\lambda)\xi_j(x) = \sum_{\nu} < G_\lambda^{-1}(\theta_\lambda)\xi_j(x), U(\theta_\lambda)\phi_{\lambda\nu} > \phi_{\lambda\nu}$$
$$= \sum_{\nu}(1+2\lambda\gamma_{\lambda\nu})^{-1}\phi_{\lambda\nu_j}(x)\phi_{\lambda\nu}$$

so that

$$\|G_\lambda^{-1}(\theta_\lambda)[\bar{Z}_{n\alpha}(\theta_\lambda) - Z_0(\theta_\lambda)]\|_{\lambda c}^2 = \|\sum_{j=1}^{q}\int G_\lambda^{-1}(\theta_\lambda)\xi_j(x)\bar{\psi}_j(x,\theta_\lambda(x))[P_X^{(n)} - P_X](dx)\|_{\lambda c}^2$$
$$= \sum_{\nu}(1+\gamma_{\lambda\nu}^c)(1+2\lambda\gamma_{\lambda\nu})^{-2}\{\int \phi'_{\lambda\nu}(x)\bar{\psi}(x,\theta_\lambda(x))[P_X^{(n)} - P_X](dx)\}^2 \; .$$

The argument used in deriving the bound (A2.6) can be adapted to show

$$| \int \phi'_{\lambda\nu}(x)\bar{\psi}(x,\theta_\lambda(x))[P_X^{(n)}-P_X](dx) | \leq k_n \, ||\phi_{\lambda\nu}| \, W_2^d(\bar{X};R^q)|| \ .$$

Assumption D(vii) is used here, and the relation is uniform in λ,ν. When this is substituted into the previous relation and use is made of C(ii) and Lemma A1.1(c), one obtains

$$\begin{aligned} ||G_\lambda^{-1}(\theta_\lambda)[\bar{Z}_{n\alpha}(\theta_\lambda) - Z_0(\theta_\lambda)]||_{\lambda c}^2 &\lesssim k_n^2 \sum_\nu (1+\gamma_{\lambda\nu}^c)(1+2\lambda\gamma_{\lambda\nu})^{-2}(1+\gamma_{\lambda\nu}^{d/m}) \\ &\approx k_n^2 \lambda^{-(c+3d/2m)} \ . \end{aligned}$$

In view of (A2.3), this shows the first expression in (A2.4) is $<< n^{-1}\lambda^{-(c+d/2m)}$, which establishes (A2.4) in entirety. *Q.E.D.*

References

1. Adams, R., *Sobolev Spaces,* Academic Press, New York, 1975.

2. Aki, K. and Richards, G., *Quantitative Seismology: Theory and Methods,* W. H. Freeman, San Francisco, 1980.

3. Anderson, J. A. and Senthilselvan, A., "Smooth estimates for the hazard function," *J. R. Statist. Soc. B.*, vol. 42, no. 3, pp. 322-327, 1980.

4. Bolt, B. A., "What can inverse problems do for applied mathematics and the sciences," *Search*, vol. 11, no. 6, 1980.

5. Budinger, T. F., "Physical attributes of single-photon tomography," *J. Nucl. Med.*, vol. 21, no. 6, 1980.

6. Cooley, R. L., "Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 1 Theory," *Water Resour. Res.*, vol. 18, pp. 965-976, 1982.

7. Cooley, R. L., "Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 2 Applications," *Water Resour. Res.*, vol. 19, pp. 662-676, 1983.

8. Cox, D. D., "Approximation of the method of regularization estimators," Tech. Rep. No. 723, Statistics Dept., University of Wisconsin-Madison, 1983.

9. Cox, D. D., "Multivariate smoothing spline functions," *SIAM J. Numer. Anal.*, vol. 21, pp. 789-813, 1984.

10. Cox, D. R. and Hinkley, D. V., *Theoretical Statistics,* Chapman and Hall, London, 1974.

11. Cramer, H., *Mathematical Methods of Statistics,* Princeton University Press, Princeton, N.J., 1946.

12. Davis, P. J. and Rabinowitz, P., *Methods of Numerical Integration,* Academic Press, New York, 1975.

13. Good, I. J. and Gaskins, R. A., "Non-parametric roughness penalties for probability densities," *Biometrika*, vol. 58, pp. 255-277, 1971.

14. Huber, P. J., *Robust Statistics,* John Wiley & Sons, New York, 1981.

15. Leonard, T., "Density estimation, stochastic processes and prior information," *J. R. Statist. Soc. B.*, vol. B, no. 40, pp. 113-146, 1978.

16. Leonard, T., "An empirical Bayesian approach to the smooth estimation of unknown functions," Tech. Rep. No. 2339, Math. Research Center, University of Wisconsin-Madison, 1982.

17. Luenberger, D. G., *Optimization by Vector Space Methods,* Wiley, New York, 1969.

18. McCullagh, P. and Nelder, J. A., *Generalized Linear Models,* Chapman and Hall, London, 1983.

19. Neuman, S. P. and Yakowitz, S., "A statistical approach to the inverse problem of aquifer hydrology 1. Theory," *Water Resour. Res.*, vol. 15, pp. 845-860, 1979.

20. O'Sullivan, F. and Wahba, G., "A cross validated Bayesian retrieval algorithm for nonlinear remote sensing experiments," *Journal of Computat. Physics*, vol. 59, no. 3, pp. 441-455, 1985.

21. O'Sullivan, F., Yandell, B., and Raynor, W. J., "Automatic smoothing of regression functions in generalized linear models," *Amer. Statist. Assoc.*, 1986 (in press).

22. Rudin, W., *Principles of Mathematical Analysis,* McGraw-Hill, New York, 1976.

23. Silverman, B. W., "On the estimation of a probability density function by the maximum penalised likelihood method," *Ann. Statist.*, vol. 10, no. 3, pp. 795-810, 1982.

24. Smith, W., "The retrieval of atmospheric profiles from VAS geostationary radiance observations," *J. Atmospheric Sciences*, vol. 40, pp. 2025-2035, 1983.

25. Smith, W. L., Woolf, H. M., Hayden, C. M., Wark, D. Q., and McMillin, L. M., "The TIROS-N operational vertical sounder," *Bull. Amer. Meteor. Soc.*, vol. 50, no. 10, pp. 1177-1187, October 1979.

26. Tikhonov, A., "Solution of incorrectly formulated problems and the regularization method," *Soviet Math Dokl*, vol. 5, pp. 1035-1038, 1963.

27. Tikhonov, A. and Arsenin, V., *Solutions of Ill-Posed Problems*, Wiley, New York, 1977.

28. Triebel, H., *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, New York, 1978.

29. Vardi, Y., Shepp, L. A., and Kaufman, L., "A statistical model for positron emission tomography," *J. Amer Statist. Assoc.*, vol. 80, no. 389, 1985.

30. Wahba, G. and Wendelberger, J., "Some new mathematical methods for varitional objective analysis using splines and cross validation," *Monthly Weather Review*, vol. 108, no. 8, pp. 1122-1143, 1980.

31. Weinberger, H., "Variational Methods for Eigenvalue Approximation," in *CBMS Regional conference series in applied mathematics*, SIAM, Philadelphia, 1974.

TECHNICAL REPORTS
Statistics Department
University of California, Berkeley

1 BREIMAN, L. and FREEDMAN, D. (Nov. 1981, Revised Feb. 1982). How many variables should be entered in a regression equation? Jour. Amer. Statist. Assoc., March 1983, 78, No. 381, 131-136.

2 BRILLINGER, D. R. (Jan. 1982). Some contrasting examples of the time and frequency domain approaches to time series analysis. Time Series Methods in Hydrosciences, (A. H. El-Shaarawi and S. R. Esterby, eds.) Elsevier Scientific Publishing Co., Amsterdam, 1982.

3 DOKSUM, K. A. (Jan. 1982). On the performance of estimates in proportional hazard and log-linear models. Survival Analysis, (John Crowley and Richard A. Johnson, eds.) IMS Lecture Notes - Monograph Series, (Shanti S. Gupta, series ed.) 1982, 74-84.

4 BICKEL, P. J. and BREIMAN, L. (Feb. 1982). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. Ann. Prob., Feb. 1982, 11, No. 1, 185-214.

5 BRILLINGER, D. R. and TUKEY, J. W. (March 1982). Spectrum estimation and system identification relying on a Fourier transform. To appear in Collected Works of J. W. Tukey, vol. 2, Wadsworth, 1985, 1001-1141.

6 BERAN, R. (May 1982). Jackknife approximation to bootstrap estimates. Ann. Statist., March 1984, 12 No. 1, 101-118.

7 BICKEL, P. J. and FREEDMAN, D. A. (June 1982). Bootstrapping regression models with many parameters. Lehmann Festschrift, (P. J. Bickel, K. Doksum and J. L. Hodges, Jr., eds.) Wadsworth Press, Belmont, 1983, 28-48.

8 BICKEL, P. J. and COLLINS, J. (March 1982). Minimizing Fisher information over mixtures of distributions. Sankhya, 1983, 45, Series A, Pt. 1, 1-19.

9 BREIMAN, L. and FRIEDMAN, J. (July 1982). Estimating optimal transformations for multiple regression and correlation.

10 FREEDMAN, D. A. and PETERS, S. (July 1982, Revised Aug. 1983). Bootstrapping a regression equation: some empirical results. JASA, 1984, 79, 97-106.

11 EATON, M. L. and FREEDMAN, D. A. (Sept. 1982). A remark on adjusting for covariates in multiple regression.

12 BICKEL, P. J. (April 1982). Minimax estimation of the mean of a mean of a normal distribution subject to doing well at a point. Recent Advances in Statistics, 1980 Wald Lectures, (W. Chernoff, ed.) Academic Press, 1983.

14 FREEDMAN, D. A., ROTHENBERG, T. and SUTCH, R. (Oct. 1982). A review of a residential energy end use model.

15 BRILLINGER, D. and PREISLER, H. (Nov. 1982). Maximum likelihood estimation in a latent variable problem. Studies in Econometrics, Time Series, and Multivariate Statistics, Academic Press, New York, 1983.

16 BICKEL, P. J. (Nov. 1982). Robust regression based on infinitesimal neighborhoods. Ann. Statist., Dec. 1984, 12, 1349-1368.

17 DRAPER, D. C. (Feb. 1983). Rank-based robust analysis of linear models. I. Exposition and review.

18 DRAPER, D. C. (Feb. 1983). Rank-based robust inference in regression models with several observations per cell.

19 FREEDMAN, D. A. and FIENBERG, S. (Feb. 1983, Revised April 1983). Statistics and the scientific method, Comments on and reactions to Freedman, A rejoinder to Fienberg's comments. To appear in Cohort Analysis in Social Research, (W. M. Mason and S. E. Fienberg, eds.).

20 FREEDMAN, D. A. and PETERS, S. C. (March 1983, Revised Jan. 1984). Using the bootstrap to evaluate forecasting equations. To appear in J. of Forecasting.

21 FREEDMAN, D. A. and PETERS, S. C. (March 1983, Revised Aug. 1983). Bootstrapping an econometric model: some empirical results. JBES, 1985, 2, 150-158.

22 FREEDMAN, D. A. (March 1983). Structural-equation models: a case study.

23 DAGGETT, R. S. and FREEDMAN, D. (April 1983, Revised Sept. 1983). Econometrics and the law: a case study in the proof of antitrust damages. To appear in the Proc. of the Neyman-Kiefer Conference, (L. Le Cam, ed.) Wadsworth, 1984.

24 DOKSUM, K. and YANDELL, B. (April 1983). Tests for exponentiality. Handbook of Statistics, (P. R. Krishnaiah and P. K. Sen, eds.) 4, 1984.

25 FREEDMAN, D. A. (May 1983). Comments on a paper by Markus.

26 FREEDMAN, D. (Oct. 1983, Revised March 1984). On bootstrapping two-stage least-squares estimates in stationary linear models. Ann. Statist., 1984, 12, 827-842.

27 DOKSUM, K. A. (Dec. 1983). Proportional hazards, transformation models, partial likelihood, the order bootstrap, and adaptive inference, I.

28 BICKEL, P. J., GOETZE, F. and VAN ZWET, W.R. (Jan. 1984). A simple analysis of third order efficiency of estimates. To appear in Proc. of the Neyman-Kiefer Conference, (L. Le Cam, ed.) Wadsworth, 1984.

29 BICKEL, P. J. and FREEDMAN, D. A. (Jan. 1984). Asymptotic Normality and the bootstrap in stratified sampling. To appear in Ann. Statist.

30 FREEDMAN, D. A. (Jan. 1984). The mean vs. the median: a case study in 4-R Act litigation. To appear in JBES.

31 STONE, C. J. (Feb. 1984). An asymptotically optimal window selection rule for kernel density estimates. Ann. Statist., Dec. 1984, 12, 1285-1297.

32 BREIMAN, L. (May 1984). Nail finders, edifices, and Oz.

33 STONE, C. J. (Oct. 1984). Additive regression and other nonparametric models. Ann. Statist., 1985, 13, 689-705.

34 STONE, C. J. (June 1984). An asymptotically optimal histogram selection rule. To appear in Proc. of the Neyman-Kiefer Conference, (L. Le Cam, ed.) Wadsworth, 1985.

35 FREEDMAN, D. A. and NAVIDI, W. C. (Sept. 1984, revised Jan. 1985). Regression models for adjusting the 1980 Census.

36 FREEDMAN, D. A. (Sept. 1984, revised Nov. 1984). De Finetti's theorem in continuous time.

37 DIACONIS, P. and FREEDMAN, D. (Oct. 1984). An elementary proof of Stirling's formula.

38 LE CAM, L. (Nov. 1984). Sur l'approximation de familles de mesures par des familles Gaussiennes. Ann. Inst. Henri Poincaré, 1985, 21, 225-287.

39 DIACONIS, P. and FREEDMAN, D. A. (Nov. 1984). A note on weak star uniformities.

40 BREIMAN, L. and IHAKA, R. (Dec. 1984). Nonlinear discriminant analysis via SCALING and ACE.

41 STONE, C. J. (Jan. 1985). The dimensionality reduction principle for generalized additive models.

42 LE CAM, L. (Jan. 1985). On the normal approximation for sums of independent variables.

43 BICKEL, P. J. and YAHAV, J. A. (1985). On estimating the number of unseen species: how many executions were there?

44 BRILLINGER, D. R. (1985). The natural variability of vital rates and associated statistics.

45 BRILLINGER, D. R. (1985). Fourier inference: some methods for the analysis of array and nonGaussian series data. Water Resources Bulletin, 1985.

46 BREIMAN, L. and STONE, C. J. (1985). Broad spectrum estimates and confidence intervals for tail quantiles.

47 DABROWSKA, D. M. and DOKSUM, K. A. (1985). Partial likelihood in transformation models with censored data.

48 HAYCOCK, K. A. and BRILLINGER, D. R. (November 1985). LIBDRB: A subroutine library for elementary time series analysis.

49 BRILLINGER, D. R. (October 1985). Fitting cosines: some procedures and some physical examples. Joshi Festschrift, 1986.

50 BRILLINGER, D. R. (November 1985). What do seismology and neurophysiology have in common? - Statistics! Comptes Rendus Math. Rep. Acad. Sci. Canada.

51 O'SULLIVAN, F. and COX, D. D. (October 1985). Analysis of penalized likelihood-type estimators with application to generalized smoothing in Sobolev Spaces.

52 O'SULLIVAN, F. (November 1985). A practical perspective on ill-posed inverse problems: A review with some new developments. To appear in Journal of Statistical Science.

53 LE CAM, L. and YANG, G. L. (November 1985). On the preservation of local asymptotic normality under information loss.

54 BLACKWELL, D. (November 1985). Approximate normality of large products.

Copies of these Reports plus the most recent additions to the Technical Report series are available from the Statistics Department technical typist in room 379 Evans Hall or may be requested by mail from:

Department of Statistics
Technical Reports
University of California
Berkeley, California 94720

Cost: $1 per copy.