A GENERAL APPROACH FOR ASYMPTOTICS OF PENALIZED SPLINE ESTIMATION IN EXTENDED LINEAR MODELS

A Dissertation

by

SUYA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Jianhua Huang
Co-Chair of Committee,	Lan Zhou
Committee Members,	Suhasini Subba Rao
	Jim Ji
Head of Department,	Valen Johnson

August 2016

Major Subject: Statistics

ABSTRACT

The penalized spline estimator has been formally introduced in the context of the nonparametric regression model. Despite the wide range of its application, the theory of penalized spline estimator has fallen behind. In this dissertation, we first look into the existing theoretical results about the penalized spline estimator with some scrutiny and point out the room left for improvement, that is, they are built upon certain asymptotic scenarios for one specific model. Then we state and prove a unified theory, that is, the convergence rate of the penalized spline estimator is established for the set of extended linear models, which holds under various asymptotic scenarios. The application of the main theory to a list of extended linear models including nonparametric regression, generalized regression, counting process, density estimation, spectral density estimation, diffusion process and nonparametric M-regression is also provided for completeness.

ACKNOWLEDGEMENTS

This dissertation is part of my PhD research. I greatly appreciate my advisor, Dr Jianhua Huang, for introducing me to the field of my research. He leads the guidance whenever I needed help and will always encourage creative and critical thinking. His advice on my research and career development will be beneficial to me in the future.

I would also like to thank my co chair, Dr Lan Zhou and other committee members Dr Suhasini, Subba Rao and Dr Jim Ji for their support and suggestions all the way toward the end of my PhD. They have been so helpful in openning my mind by providing opinions on my research from their point of view. They are always there when in need.

Last but not least, I want to express my special thanks to my family. I have spent the last five years abroad, leaving my parents alone at home. They have always been supportive and respected my decisions. My beloved fiancée, Yanjun Qian, is accompanying me through my PhD, always willing to hear what I went through. Thank you for being my soul mate.

TABLE OF CONTENTS

Page

1. INTRODUCTION

1.1 Background and terminology

Extended linear modeling (Hansen (1994),Stone et al. (1997)) provides a flexible framework regarding function estimation problems with either one covariate or multiple covariates, including ordinary and generalized regression, density and conditional density estimation, hazard regression or more generally counting process regression, spectral density estimation and diffusion process, nonparametric M-estimation and etc. The extended linear modeling has already drawn people's attention because the models within have been found to share some common property regardless of its impressive capacity.

The penalized spline estimator has been formally introduced in Eilers and Marx (1996) in the context of the nonparametric regression model. Since then it has become a popular smoothing technique and a proper tool as long as the computational cost is taken into consideration. From the application aspect, Ruppert et al. (2003) provides a nice coverage of modeling situations suitable for this technique. Despite the wide range of applications, the theory of penalized spline estimate has fallen behind.

We will first briefly summerize the existing theoretical results about penalized spline smoothing, while later in chapter 2 some details will be added.

Hall and Opsomer (2005) utilizes the equivalence between nonparametric regression and its white noise representation and obtained the mean integrated squared error for the estimator. The white noise representation assumes that both the data points and knots are continuously distributed.

Li and Ruppert (2008) finds an equivalent kernel representation for piecewise

constant and linear B-splines using first or second order difference penalties under regression model, which provides similarity of these penalized splines and the Nadaraya-Watson kernel estimators. It is worth noting that the theoretical results in their paper are derived under the assumption that the degree of the spline space is smaller or equal to the penalty order while the latter is half the smoothness of the true function. Also in their setting the optimal convergence rate would not depend on the penalty parameter as well as the number of knots when certain regularity conditions are fullfilled.

Kauermann et al. (2009) works on the generalized regression model and obtained the mean squared error and the central limit theorem of the penalized spline estimator. The optimal rate of convergence in the theoretical results is built when the penalty parameter is shrinking relatively fast and that the number of knots grows at some fixed rate.

The first work that found a transition in the asymptotic behavior of penalized spline and that of two other techniques, the so-called regression spline and smoothing spline, in terms of the mean squared rate of convergence is provided in Claeskens et al. (2009). The transition from penalized spline to regression spline (smoothing spline) is achieved when few (many) knots are used while the penalty tuning parameter is small (large). Holland (2012) further extends the result in Claeskens et al. (2009) to the partially linear model in the multivariate case.

Wang et al. (2011) compares penalized spline with smoothing spline by treating both as ODE solutions and characterizes the equivalence between the two when the number of knots is large. Their setting is similar to Li and Ruppert (2008) in that the penalty order is half the smoothness of the true function but the degree of spline is said to be not related the penalty order.

On the other hand, as already discovered in the nonparametric regression set-

ting, penalized spline smoothing has an interesting link to mixed effects model, by comprehending the penalty imposed on the spline coefficients as a Gaussian prior, see Wand (2003). The key ingredient is the Laplace approximation to the integral calculation involved in the marginal distribution, where the integral is taken with respect to the random coefficients. In Kauermann et al. (2009), the equivalence in generalized smoothing model and generalized linear mixed models is established and the rates at which the spline basis dimension increases and the precision of the random coefficients shall increase that together garantee the Laplace approximation is also investigated.

The existing works focus on the asymptotic properties for some specific models. The techniques are restrictive and not easy to extend to cases corresponding to estimators that have no explicit form. In our case, we establish a general theory that captures the rate of convergence property of the penalized spline estimator for all extended linear models. The proposed method makes use of the concave property of the objective functional instead of its specific form so that we are able to provide the proofs for various models. Indeed, even under the regression model, the proof is comparatively neat than Claeskens et al. (2009).

One can formulate the penalized spline estimation quite flexibly, leading to different asymptotic scenarios. This is the main reason that demonstrating a unified theory requires a hard work while appears interesting. These asymptotic scenarios depend on the specification of **three components**. First, the design of data could be treated as fixed or random and the placement of the knots can be either equally spaced or not, leading to balanced design or unbalanced design. In reality, it is natural to have random draws for the data and unbalanced design of knots but we could control the positions of knots. Second, we need simultaneously specify the degree of spline, the order of function derivative used in the penalty term and the smoothness

order corresponding to the function class of the parameter of interest. Third, we could allow the number of knots and the penalty parameter vary along with the sample size, whose rates could affect the property of the estimator. By "asymptotic scenario" we mean one combination of certain scenario in each of the three components. As we will see in chapter 2, the existing work mainly covers certain scenarios, in other words, the combination of the three components is quite restricted.

The contribution of the current paper is stated here. First, as mentioned above, the established theory applies to a general class of models with an unknown function to be estimated. To our best knowledge, this is the first paper that provides a unified theory for the penalized spline estimator. Second, our theory is built over a broadly varying asymptotic scenarios, it is quite general in this sense too.

This rest of this paper is organized as follows. In section 1.2 and 1.3, we briefly introduce the extended linear model and formulate the penalized spline estimator for the extended linear model. In chapter 2, we are going to zoom in some of the existing works mentioned above for better illustration of limitation and comparison with the current paper. In chapter 3, we present the convergence rate of the penalized spline estimators together with the conditions that make it complete and rigorous. We provide the detailed proofs in chapter 4. We end this article with chapter 5, which serves as the know-how application of the general theory to a number of selected extended linear models.

1.2 Extended linear model

Let η_0 be a nonzero function of interest with support $\mathcal{U}$. To be concise, we only consider $\mathcal{U}$ to be an one dimensional interval $[a, b]$. The function η_0 is always associated with the distribution of a random variable $\mathbf{W}$, $\mathbf{W}$ is possibly a random vector or a random function or a vector of random functions. For the purpose of estimating

η_0, an *i.i.d.* sample, $\mathbf{W}_1, \ldots, \mathbf{W}_n$, shall be obtained from the distribution of $\mathbf{W}$ (In some cases, we are capable of working with independent but not identical sample, $\mathbf{W}_1, \ldots, \mathbf{W}_n$, one example lies in spectral density estimation, the corresponding section is 5.5). It is typical to impose some smoothness structure on η_0, in other word, we assume η_0 belongs to a function space $\mathbb{H}^m$ containing smooth functions (in some sense) defined on $\mathcal{U}$:

$$\mathbb{H}^m = \{f : f^{(m-1)} \text{ is absolutely continuous and } f^{(m)} \in L_2\}. \tag{1.1}$$

Here $f^{(l)}$ denotes the l-th derivative of f. The function space $\mathbb{H}^m$ is widely spread in the smoothing spline literature. From here and now on, we refer to $\mathbb{H} = \mathbb{H}^m$ as the model space, i.e., $\eta_0 \in \mathbb{H}$.

For a candidate function η and an *i.i.d.* observation $\vec{\mathbf{W}} = \{\mathbf{W}_1, \ldots, \mathbf{W}_n\}$, the scaled log-likelihood is written as $l(\eta, \vec{\mathbf{W}})$ (Here the scaling is given by the factor $1/n$). Furthermore, we could come up with the expected log-likelihood, namely, $\Lambda(\eta) = El(\eta, \vec{\mathbf{W}})$, where the expectation is taken with respect to the true distribution of $\mathbf{W}$. Meanwhile, for simplicity, we would assume η_0 is the maximizer of the expected log-likelihood, that is, $\eta_0 = \arg\max_{\eta \in \mathbb{H}} \Lambda(\eta)$, this is not quite an assumption when $\eta_0 \in \mathbb{H}$ as one can show the maximizer of $\Lambda(\eta)$ agrees with η_0 by the information inequality. $l(\eta, \vec{\mathbf{W}})$ and $\Lambda(\eta)$ could be more general than log-likelihood and its expectation. In some circumstances, η_0 is not involved in the joint probability distribution of $\vec{\mathbf{W}}$, but is only associated with the conditional distribution, or even more complicated psuedo likelihood as well as loss function, the scaled logarithm of which can be treated as $l(\eta, \vec{\mathbf{W}})$. From here and now on, such broad representation of $l(\eta, \vec{\mathbf{W}})$ is allowed in extended linear modeling. Under these circumstances, the previous assumption that "η_0 is the maximizer of $\Lambda(\cdot)$" might not hold, the theorem still holds when replacing

η_0 by the maximizer of $\Lambda(\cdot)$ instead. For simplicity of making statements, we shall still call $l(\eta, \vec{\mathbf{W}})$ and $\Lambda(\eta)$ as the scaled log-likelihood and the expected log-likelihood respectively when the focus is not on a specific model.

We shall call the model an extended (concave) linear model if (i) for each value of $\vec{\mathbf{W}}$, $l(\cdot, \vec{\mathbf{W}})$ is concave in its first argument and (ii) $\Lambda(\cdot)$ is strictly concave. That means, given any $h_1 \in \mathbb{H}$, $h_2 \in \mathbb{H}$ and $0 \leq \alpha \leq 1$, $l(\alpha h_1 + (1-\alpha)h_2, \vec{\mathbf{W}}) \geq \alpha l(h_1, \vec{\mathbf{W}}) + (1-\alpha) l(h_2, \vec{\mathbf{W}})$, while $\Lambda(\alpha h_1 + (1-\alpha)h_2) > \alpha\Lambda(h_1) + (1-\alpha)\Lambda(h_2)$. Note in the above definition, the functions $h_1, h_2 \in \mathbb{H}$ is implicitly assumed to be selected from the set of functions in $\mathbb{H}$ such that both the log-likelihood functions and the expected log-likelihood are well-defined. From here and after this feasible set is supposed to be convex.

Condition 1.2.1. *For each value of* $\vec{\mathbf{W}}$, $l(\eta, \vec{\mathbf{W}})$ *is concave in* η *and* $\Lambda(\eta)$ *is strictly concave in* η.

For simplicity, we will omit $\vec{\mathbf{W}}$ in the argument of $l(\cdot, \cdot)$ and write $l(\eta, \vec{\mathbf{W}})$ as $l(\eta)$.

The class of extended linear models (Hansen (1994),Stone et al. (1997)) is extremely rich, such models include the ordinary and generalized nonparametric regression, density (or conditional density) estimation, counting process regression, spectral density estimation, diffusion process and nonparametric M-regression, readers could refer to Huang (2001), Stone et al. (1997) for a more complete review. In chapter 5 we will elaborate on a list (belonging to the extended linear model framework) to which we will apply the general theory stated in chapter 3. Hence readers will be convinced that the penalized likelihood method considered in this paper would behave similarly when it is applied to the group of models, e.g., the extended linear models.

Notation. Given two sequence of positive numbers a_n and b_n, let $a_n \lesssim b_n$

denotes the ratio a_n/b_n is bounded for all n and $a_n \asymp b_n$ if and only if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Let $\|h\|_\infty$ denotes the sup norm of h, $\|\cdot\|$ be the norm defined on $\mathbb{H}$ such that $\|h\|^2 = V(h) = \int h^2(x)\omega(x)dx$, where $\omega(x)$ is some weight function which might vary under different model. Note that two notations representing the same norm as $\int h^2(x)\omega(x)dx$ are introduced just for convenience and notation uniformity. Using this norm, we could introduce an inner product on $\mathbb{H}$ denoted as $V(h_1, h_2) = \int h_1(x)h_2(x)\omega(x)dx$.

1.3 Penalized spline estimator

Suppose $\eta_0 \in \mathbb{H}$ is the underlying parameter of interest. In practice, to seek an estimator for η_0, we would like to restrict ourselves on a smaller space that is usually of finite dimension. In contrast to the model space, the space where the estimator lies in is refered to as the estimation space. In this paper the estimation space is considered to be the B-spline space.

1.3.1 Basics of spline

We start with a partition or knot sequence, i.e., a nondecreasing sequence $\mathbf{k} \doteq \ldots k_{-2}, k_{-1}, k_0, k_1, k_2, \ldots$. We shall only adopt the "simple knot" setup where the knots have no ties, that is, $k_i < k_{i+1}$. The b-splines of degree 0 or order 1 for this knot sequence are the characteristic functions of the partition given by $\mathbf{k}$, i.e.,

$$b_{i,0}(x) = \begin{cases} 1 & \text{if } k_i \le x < k_{i+1} \\ 0 & \text{if otherwise.} \end{cases} \tag{1.2}$$

From these zero degree b-splines, the higher-order b-splines are defined recursively, that is

$$b_{i,p}(x) := \omega_{ip}(x) b_{i,p-1}(x) + (1 - \omega_{i+1,p}(x)) b_{i+1,p-1}(x) \tag{1.3}$$

with the weight function

$$\omega_{ip}(x) := \frac{x - k_i}{k_{i+p-1} - k_i}.$$

Definition 1.3.1. *A spline of degree p with any knot sequence $\mathbf{k}$ is a linear combination of the b-splines $b_{i,p}$ above.*

In our setup, the interested region for the unknown function is the interval $[a, b]$. The relevant knots are finite. The knot $\tilde{\mathbf{k}}$ could be, as an attempt, selected to be $a = \tilde{k}_0 < \tilde{k}_1 < \cdots < \tilde{k}_{K_n+1} = b$. This, however, could cause some problem of characterizing a function near the boundary, because there is less non zero basis functions near the boundary than those in the interior. To avoid this issue, in practice, we would specify some extra knots outside the interval $[a, b]$ such that the b-splines corresponding to the extra knots are not vanishing on $[a, b]$. The number of knots after including the extra knots is denoted as N_n, which is determined by both the number of interior knots K_n and the degree p. Let $\tilde{\mathbf{k}}$ together with the extra knots be called the extended knot and is denoted as $\mathbf{k} := k_1, \ldots, k_{N_n}$.

The spline space (restricted on $[a, b]$) of degree p is the set of all splines of degree p with the extended knot $\mathbf{k}$,

$$\mathbb{G}_n = \{g : g = \sum_{k=1}^{N_n} g_k b_{k,p}\} \tag{1.4}$$

where $\{b_{k,p}\}_{k=1,\ldots,N_n}$ are the b-spline basis functions of degree p associated with extended knot $\mathbf{k}$. The number of basis functions N_n (or equivalently, the number of

extended knots) and the number of interior knots K_n satisfy that

$$N_n = K_n + p + 1. \tag{1.5}$$

Let δ_n be the largest distance between all the neighboring knots, that is,

$$\delta_n = \max_{1 \le i \le N_n - 1} |k_{i+1} - k_i|. \tag{1.6}$$

The knots are pre-specified for $\mathbb{G}_n$. It is not necessary that they distribute equally between $[a, b]$ (rigorously the range of knot $\mathbf{k}$ is a little wider than $[a, b]$). We do require that the knots are not spreading too unevenly, to be specific, the knots are supposed to have "bounded mesh ratio", that is, the ratio of the maximum and minimum distance between two neighbouring knots is bounded from above and below by two positive numbers, which do not depend on n. Under this assumption, we would have $\delta_n \asymp \frac{1}{K_n}$.

We will state two basic results regarding spline functions, the proof of which can be found in Section 4.4 and Theorem 5.1.2 of DeVore and Lorentz (1993) correspondingly.

Lemma 1.3.1 (L_2 norm of spline functions)**.** *There exist constant $C_1 > 0$ and $C_2 > 0$ which do not depend on n such that for any $g \in \mathbb{G}_n$ having expression (1.4), it holds that*

$$C_1 \delta_n \sum_k g_k^2 \le \int g^2(x) dx \le C_2 \delta_n \sum_k g_k^2.$$

Proposition 1.3.1 (Ratio between L_∞ and L_2 norm of spline function)**.** *Let $A_n = \sup_{g \in \mathbb{G}_n, \|g\| \neq 0} \{\|g\|_\infty / \|g\|_{L_2}\}$. Then $A_n \asymp \delta_n^{-1/2}$, that is, there exist constants $C_3 > 0$ and $C_4 > 0$ such that*

$$C_3 \le A_n \delta_n^{1/2} \le C_4.$$

It is worth noting that the previously defined notations might vary as n increases, for notation simplicity, we will drop the subscript from here and now on. So $\mathbb{G} = \mathbb{G}_n$, $K = K_n$, $\delta = \delta_n$, $\lambda = \lambda_n$.

1.3.2 Penalized spline estimation in extended linear model

With the scaled log-likelihood $l(\eta)$ defined on each spline function, one could seek the estimator of η_0 by maximizing $l(\eta)$ within $\mathbb{G}_n$, leading to the maximum likelihood estimator. Such estimators have been well studied, the convergence rates for extended linear models have been treated in Huang (2001), while Huang (2003) deals with the asymptotic distribution in the nonparametric regression setting. On the other hand, to mitigate overfitting when there are too many knots selected, another type of estimator called penalized spline estimator was introduced in O'Sullivan (1986), O'Sullivan (1988) and later a modified version in Eilers and Marx (1996). Penalized spline fitting is considered a flexible smoothing technique while at the same time computational efficient, readers could refer to Ruppert et al. (2003) and many others for its variety of applications.

In the present work, we are focusing on the penalized spline estimator by minimizing the penalized (negative) log likelihood within the spline space $\mathbb{G}$, to be specific,

$$\hat{\eta} = \arg\min_{\eta \in \mathbb{G}} \{-l(\eta) + \lambda_n J(\eta)\},$$

which is directly estimating the minimizer of its expectation

$$\bar{\eta} = \arg\min_{\eta \in \mathbb{G}} \{-\Lambda(\eta) + \lambda_n J(\eta)\}.$$

Here the penalty term $J(h) = J(h, h)$ is a quadratic functional (defined on $\mathbb{H}$) quantifying the roughness of $h \in \mathbb{H}$, we consider $J(\cdot)$ to be the integral of the squared

derivative of order q, as introduced by O'Sullivan (1986),

$$J(h) = \int \{h^{(q)}(x)\}^2 dx.$$

With the penalty being defined, we can provide more stories regarding the penalized spline estimator. Take a concrete example in the context of nonparametric regression model, $y_i = \eta_0(x_i) + \epsilon_i$, the criterion functional becomes penalized sum of squared loss,

$$\hat{\eta} = \arg\min_{\eta \in \mathbb{G}} \left\{ \frac{1}{n} \sum_i (y_i - \eta(x_i))^2 + \lambda \int \{\eta^{(q)}(x)\}^2 dx \right\}. \tag{1.7}$$

For the validity of the derivative operation on η_0, we assume that $q \leq m$; also, we would require that the estimation space is contained in the model space, which leads to $m \leq p$ (p and m are defined in (1.4) and (1.1) respectively).

Penalized spline estimation can be thought of as an intermediate step between two other approaches in nonparametric estimation. For the ease of illustration, we use the penalized spline in the regression model as an example. Two related methods, regression spline and smoothing spline, are defined as

$$\hat{\eta}_{reg} = \arg\min_{\eta \in \mathbb{G}} \left\{ \frac{1}{n} \sum_i (y_i - \eta(x_i))^2 \right\} \tag{1.8}$$

and

$$\hat{\eta}_{ss} = \arg\min_{\eta \in \mathbb{H}} \left\{ \frac{1}{n} \sum_i (y_i - \eta(x_i))^2 + \lambda \int \eta^{(m)}(x)^2 dx \right\} \tag{1.9}$$

respectively. $\hat{\eta}_{reg}$ is a natural extension of MLE estimator or quasi MLE estimator (where no penalty is added) to the nonparametric estimation problem, it can be

problematic when too many knots are used which causes overfitting. $\hat{\eta}_{ss}$ originates from Wahba (1990) which makes use of the RKHS theory to study the property of the estimator. The criterion function looks similar to (1.7), however, because the estimator is searched in a larger space $\mathbb{H}$ it can get into trouble from the aspect of computation. Both regression spline and smoothing spline have been studied for quite a long period of time. Because of the similarity of $\hat{\eta}$ from $\hat{\eta}_{reg}$ and $\hat{\eta}_{ss}$, the connection of penalized spline and these two related methods remains an open topic and requires further investigation.

In addition to the B-spline basis, we could also adopt polynomial spline basis,

$$\phi_i(x) = \begin{cases} x^i & \text{if } 0 \leq i \leq p \\ (x - \tilde{k}_{i-p})_+^p & \text{if } p+1 \leq i \leq K+p \end{cases} \tag{1.10}$$

as the estimation space used in Ruppert et al. (2003). In this case, it has become standard to add a penalty to all but the first $p+1$ coefficients

$$\min_{\vec{g} \in R^{K+p+1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{k=0}^{K+p} g_k \phi_k(x_i))^2 + \lambda \sum_{j=p+1}^{K+p} g_k^2 \right\}. \tag{1.11}$$

In this approach (1.11), the penalty is not directly connected to the usual measure of roughness of using derivatives (in fact, the penalty in (1.11) is equal to the integral of squared generalized $(p+1)$-th derivative of the spline function, thus involves the derivative of dirac functions). Our theory will not cover this choice of penalty, say, $q = p+1$.

2. LITERATURE REVIEW

2.1 Hall and Opsomer (2005)

Hall and Opsomer (2005) utilizes the white noise model version of the nonparametric regression, $y_t = f_0(t) + \epsilon_t$. The penalized spline can be adapted to the white noise model, assuming that the knots are placed continuously and therefore the coefficients of the basis functions which depend on the knots are formulated as a function. Meanwhile, the data points $\{(t, y_t), t \in [a, b]\}$ are likewise continuous. The criterion (1.11) becomes

$$I(\beta_0, \ldots, \beta_p, \beta(s)) = \int \{y_t - \sum_{k=0}^{p} \beta_k \phi_k(t) - \int \beta(s)\rho(s)\phi(t|s)ds\}^2 h(t)dt + \lambda \int \beta(t)^2 dt, \tag{2.1}$$

where $\phi_0(\cdot), \ldots, \phi_p(\cdot), \phi(\cdot|s)$ can be any basis functions and in the case of polynomial spline, $\phi_k(t) = t^k$, $\phi(t|s) = (t-s)_+^p$, h is the density of the distribution of x_i's and ρ equals the distribution of knots. In this section, $\hat{f} = \sum_{k=0}^{p} \hat{\beta}_k \phi_k(t) - \int \hat{\beta}(s)\rho(s)\phi(t|s)ds$, where $\hat{\beta}_0, \ldots, \hat{\beta}_p, \hat{\beta}(s)$ denotes the minimizer of (2.1).

Under the white noise model, the effect of estimating $\{\beta_k\}_{k=0}^p$ has negligibly small influence on both the bias and variance of the estimator, hence they can be set to zero. Utilizing spectral functional decomposition arguments and under the equally spaced knots design meaning that $\rho(s) \equiv \rho$, the solution to (2.1) can be explicitly worked out. the expressions of the bias, stochastic error and mean integrated squared error of the resulting estimator are given. For the mean integrated squared error, Hall and Opsomer obtained that under the white noise model

$$\int E(\hat{f} - f_0)^2 h = O\{(n\lambda^{1/2(p+1)})^{-1} + \lambda\}, \tag{2.2}$$

where f_0 is assumed to be having $p+1$ well-defined, square integrable derivatives.

We now turn to the limitations of this paper. First of all, all the explicit expressions derived in this work are based on the uniform knots design. Second, there is no freedom in choosing the degree of spline p and smoothness of true function m, which is kept to be $p = m - 1$. More importantly, the white noise model is considered as the limiting case of the discrete model, where the knots and data points are continuous, it is unclear how the results (e.g. MSE, bias and variance) in this paper can be extended to the discrete model (1.11).

2.2 Li and Ruppert (2008)

In this paper, the authors considered a nonparametric regression model $y_i = f_0(x_i) + \epsilon_i$, with heteroscedastic error ϵ_i having mean zero and variance $\sigma^2(x_i)$. They studied the following penalized spline estimator (as a note here, I change the first term by scaling it with $1/n$ for notation conformity, therefore, the λ in this section would need to be adjusted to λ/n to get the original results):

$$\min_{\vec{g}\in\mathbb{R}^N} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{N} g_j b_{j,p}(x_i))^2 + \lambda \sum_{j=q+1}^{N} \{\Delta^q(g_j)\}^2 \right\} \tag{2.3}$$

where the penalty involves the q-th order difference operator on the coefficients, which has been used in Eilers and Marx (1996).

The solution to problem (2.3) can be obtained by solving a linear system of equations, $(\frac{B^TB}{n} + \lambda(D^q)^TD^q)\hat{\vec{g}} = \frac{1}{n}B^TY$, where matrix $B = (B_{ik})_{i=1,\dots,n}^{k=1,\dots,K+p-1} = (b_{k,p}(x_i))_{i=1,\dots,n}^{k=1,\dots,K+p}$ and $D^q\vec{g} = (\Delta^q(g_{q+1}), \cdots, \Delta^q(g_{K+p-1}))^T$. The authors showed for some combination of $p = 0, 1$ and $q = 1, 2$, by cleverly making use of the banded pattern of $\frac{B^TB}{n} + \lambda(D^q)^TD^q$ and choosing λ, the coefficient vector $\hat{\vec{g}}$ can be explicitly written out as a weighted average of binned data, so can $\hat{f}$. As an illustration,

consider zero degree spline with first order penalty, e.g. $p = 0$ and $q = 1$. Under equally spaced knots $0 = k_0 < \ldots < k_K = 1$ and $x_i = i/n$ assumption, $\frac{1}{n}B^TB + \lambda(D^q)^TD^q$ becomes a Toeplitz matrix with modified first and last diagonal elements. It is shown that the first and the last element $\hat{\vec{g}}$ can be solved separately while the middle elements of $\hat{\vec{g}}$ depend on these two. For any $x \in (0, 1)$ and $t = t_n(x)$ such that $t/K \to x$, when λ and K be chosen as

$$\lambda \sim n^{-1}\{Khn^{-1/5}\}^2 \text{ for some } h > 0, \tag{2.4}$$

the estimator is equivalent to a binned Nadaraya-Watson kernel estimator, with equivalent kernel $H(x) = \frac{1}{2}\exp(-|x|)$ and bandwidth of order $n^{-1/5}$

$$\hat{f}(x) = \hat{g}_t \sim \frac{\sum_{j=1}^{K} \rho_n^{|t-j|}\bar{y}_j}{\sum_{j=1}^{K} \rho_n^{|t-j|}}, \tag{2.5}$$

where $\rho_n = \exp\{-h^{-1}n^{1/5}K^{-1}\}$, $\bar{y}_j$ is the average of all y_i such that $k_{j-1} < x_i \leq k_j$.

The bin size is controlled by total number of knots, under the following condition

$$K \sim n^{\gamma} \text{ for } \gamma > 2/5, \tag{2.6}$$

the binning effect by knots negligible, that is,

$$\bar{y}_t = f_0(\bar{x}_t) + \epsilon' + o(n^{-2/5}) \tag{2.7}$$

with $\bar{x}_t$ being the midpoint of the t-th bin, and $\epsilon' \sim N(0, [K/n]\sigma^2(\bar{x}_t))$.

Therefore, asymptotic distribution including bias, variance can be derived based on the kernel representation (2.5) and binned data representation (2.7) using well-

known techniques, such as Wand and Jones (1995)

$$n^{2/5}\{\hat{f}(x) - f_0(x)\} \to N(\mathcal{B}(x), \mathcal{V}(x)), \tag{2.8}$$

where $\mathcal{B} = h^2 f_0^{(2)}(x)$ and $\mathcal{V}(x) = 4^{-1}h^{-1}\sigma^2(x)$. In addition to (2.4), there are other conditions needed to show (2.8). First, f_0 is assumed to be having continuous second derivative. Second, the response has a finite higher-than-second order moment $E(Y^{2+\delta}) < \infty$, for some $\delta > 0$.

Remark 2.2.1. *In order to derive (2.8), it is required that f_0 has continuous second derivative. This is due to the order of the equivalent kernel.*

The above result for $p = 0, q = 1$ and $x \in (0, 1)$ can be extended to some other cases. For other values of p, q, the equivalent kernel expression can be found with different condition than (2.4) and (2.6). When x is at the boundary, it is more complicated than $x \in (0, 1)$, although equivalent kernel may be found but it is not always easy to build the asymptotic for it. To be specific, the authors provided asymptotic normality in cases such as $p = 0, q = 2$ and $x \in (0, 1)$ or x is at the boundary region; $p = 1, q = 1$ and $x \in (0, 1)$. They did an analysis on the equivalent kernel for $p = 1, q = 1$ and x is at the boundary region without giving the asymptotic normality. For the linear spline $p = 1$ they have no discussion on cases other than the one just mentioned.

We will end this section with some remarks on the critical conditions in the paper. The design of data and knots does not apply to the general case, to be specific, the authors say "our theory does not cover the situation where the data points are unequally spaced but the knots are equally spaced", because otherwise the structure of the matrix which formulates the linear system for the solution will be destroyed. Also, it is worthy to point out that the order specification is in a quite different

flavor. Suppose the true function f has continuous m-th derivative. The theoretical results are constrained to be one of the following combinations: $p = 0, q = 1, m = 2$; $p = 0, q = 2, m = 4$; $p = 1, q = 1, m = 2$ for the interior points; $p = 0, q = 2, m = 2$ for the boundary points (see the discussion in the last paragraph). Last but not least, the asymptotic results, e.g.(2.8), is based on the choice of λ and K as in (2.4) and (2.6), which indicates that the relationship between penalty parameter and number of knots is fixed, and one is completely determined by the other. Last but not least, the idea of solving the linear system is very smart but turns out to be restrictive, the whole method will fail in models other than nonparametric regression model.

2.3 Wang et al. (2011)

Wang et al. (2011) shows the equivalence between P-spline and smoothing spline under the large number of knots assumption (there will be further explaination on what this means in the following context). Here P-spline refers to the solution to (2.3). Unlike Li and Ruppert (2008), there is no restriction on the order p and q.

They first investigated the case when $p = q$. The most effort is put into the process of writing P-spline $\hat{f}$ and smoothing spline $\hat{f}_{ss}$ (defined in (1.9)) as the solutions to two differential equations sharing the same Green function ($K_\lambda(x, y)$ below) which differ in terms that are of small estimable order. The whole process depends on matrix operation on the first order equation, where matrices that are discrete anaogs of integration are used. The construction of such matrix is a result of equal design of data points and knots.

Under the above differential equation setup, the difference function $\hat{f} - \hat{f}_{ss}$ can be controlled by the boundedness of the Green function and the small terms. Concisely speaking, the convergence rates under the supreme norm over any compact subset

and over the whole interval are

$$\sup_{x\in[\varrho,1-\varrho]}|\hat{f}(x)-f_{ss}(x)| = O_p\left(\frac{\lambda^{1/2}}{K}\right)+O_p\left(\left(\frac{\log K}{n\lambda K}\right)^{1/2}\right)$$
$$\sup_{x\in[0,1]}|\hat{f}(x)-f_{ss}(x)| = O_p\left(\frac{1}{K}\right)+O_p\left(\left(\frac{\log K}{n\lambda K}\right)^{1/2}\right), \tag{2.9}$$

It is known that the smoothing spline estimator f_{ss} is asymptotically equivalent to the kernel smoothing (Silverman (1984))

$$f_{ss}(x) = \int K_\lambda(x,y)f(y)dy + \frac{1}{n}\sum_{i=1}^{n} K_\lambda(x,x_i)\epsilon_i + \text{higher order terms}, \tag{2.10}$$

where the equivalent kernel $K_\lambda(x,y)$ is the Green's function for an ordinary differential equation with boundary conditions

$$(-1)^q\lambda u^{(2q)}(t)+u(t)=v(t)$$
$$\text{subject to}\quad u^{(q)}(0)=u^{(q)}(1)=0,\quad \text{for}\quad k=q,\ldots,2q-1. \tag{2.11}$$

Based on (2.9) and (2.10), the former characterizes the pointwise distance between $\hat{f}$ and $\hat{f}_{ss}$ while the latter connects $\hat{f}_{ss}$ with Kernel smoothing, of which the asymptotic normality is known, one is able to derive the asymptotic normality of $\hat{f}$ under some conditions indicating that the number of knots is large enough such that the pointwise distance $\hat{f}-\hat{f}_{ss}$ is negligible compared to the asymptotic rate of $\hat{f}_{ss}$, explicitly, this requires the number of knots increases faster than certain (negative) power of λ (readers could refer to Corollary 3.1 in their paper, however, the authors did not explain on the conditions and it's not easy to understand). Meanwhile, the asymptotic distribution result is built upon the assumption that f_0 is $2q$ times continuously differentiable.

Next, the authors considered the cases when $p \neq q$. The idea is that they come up with a bridge $\tilde{f}^{[q]}$ that is a q-th degree spline but sharing the same coefficients as $\hat{f}$, roughly speaking. The author claims that it would be easy to establish the asymptotic normality result for $\tilde{f}^{[q]}$ follwing the similar discussion as they did when $p = q$, while they are also providing the distance between $\hat{f}$ and $\tilde{f}^{[q]}$, thus the asymptotic result could be passed from $\tilde{f}^{[q]}$ to $\hat{f}$.

Next, we are going to point out some concerns. First, when the design is not balanced or the knots are not equally spaced, the paper did not provide the equivalence (2.9) explicitly except some remarks in the last section. Without equal knots, $\Delta^p \hat{b}_{p+j} = \frac{1}{K^p} \frac{d^p}{dx^p} \hat{f}(x), x \in (k_{j-1}, k_j], j = 1, \ldots, K$ would not hold, therefore, the estimator would not be easily written as the solution to the differential equation as in Theorem 2.1. Second, the asymptotic distribution is built under the condition that f_0 is $2q$ continuously differentiable with bounded $2q$-th derivative, or equivalently, $m = 2q$, in practice, this hidden relationship could hardly be achieved when m is unknown. Third, the number of knots has to be large enough to ensure that the approximation error (2.9) is negligible relative to the variance of $\hat{f}_{ss}$. Last but not least, the proposed method may not be extended to models other than the nonparametric regression model.

2.4 Kauermann et al. (2009)

Kauermann et al. (2009) considers the penalized spline estimator for the generalized exponential family model

$$y|x \sim \exp\left[\frac{y\vartheta(x) - a\{\vartheta(x)\}}{\phi} + c(y, \phi)\right], \tag{2.12}$$

with $\vartheta(x) = \vartheta\{\eta_0(x)\}$ as the natural parameter of the underlying exponential family and ϕ as dispersion parameter. The unknown function $\eta(x)$ is assumed to be smooth

in x.

The penalized spline estimator can be formulated as the maximizer of (as a note here, I change the first term by scaling it with $1/n$ for notation conformity, therefore, the λ in this section would need to be adjusted to λ/n to get the original results)

$$l(\vec{\theta}, \lambda) := \frac{1}{n} l(\vec{\theta}) - \lambda u^T u = \frac{1}{n} \sum_{i=1}^{n} \left[y_i \vartheta(P_i \vec{\theta}) - a\{\vartheta(P_i \vec{\theta})\} \right] - \lambda u^T u, \tag{2.13}$$

where P_i is the row vector of polynomial spline basis of degree p (1.10) evaluated at x_i, $\theta = (\beta^T, u^T)^T$ with β and u denotes the coefficients of the polynomial basis and the truncated polynomials. The penalty in (2.13) could also be written as $\vec{\theta}^T D \vec{\theta}$, where D is a diagonal matrix with the upper left diagonals zero and lower right diagonals 1 such that $\vec{\theta}^T D \vec{\theta} = u^T u$.

Changing the polynomial spline basis in (2.13) with b-spline basis, that is, $P_i \vec{\theta} = B_i \vec{g}$, where B_i is the row vector of spline basis evaluated at x_i, $\vec{g} = K^{-p} L^{-1} \vec{\theta}$, for some matrix L

$$l(\vec{g}, \lambda) := \frac{1}{n} l(\vec{g}) - \lambda \vec{g}^T D_{p+1} \vec{g} = \frac{1}{n} \sum_{i=1}^{n} \left[y_i \vartheta(B_i \vec{g}) - a\{\vartheta(B_i \vec{g})\} \right] - \lambda K^{2p} \vec{g}^T \tilde{D} \vec{g} \tag{2.14}$$

with $\tilde{D} = L^T D L$.

We briefly list the key steps in the argument to the main result. For the generalized exponential model, the first order estimating equation is

$$0 = \frac{1}{n} \mathbf{B} l_\eta\{\mathbf{B}^T \hat{\vec{g}}\} - 2\lambda \tilde{D} \hat{\vec{g}}, \tag{2.15}$$

where $l_\eta\{\mathbf{B}^T \hat{\vec{g}}\} := \left(\frac{\partial \vartheta}{\partial \eta(B_i \hat{\vec{g}})} [y_i - a'(\vartheta(B_i \hat{\vec{g}}))] \right)_{i=1,\ldots,n}$, $\mathbf{B} = (B_1^T, \ldots, B_n^T)$.

To solve (2.15), they expand each component as a function of $\vec{g}$ around $\vec{g}_0$, the

coefficient of best spline approximation of the true function η_0 based on Kullback-Leibler measure. The series inversion is then applied to get a series expression of $\hat{g}_l - g_{0l}$, where g_{0l} is the l-th component of $\vec{g}_0$. Writing in matrix form the expression of $\hat{\vec{g}} - \vec{g}_0$ is obtained

$$\hat{\vec{g}} - \vec{g}_0 = F^{-1}(\lambda)(\mathbf{B}l_\eta\{\mathbf{B}^T\vec{g}_0\} - n\lambda K^{2q}\tilde{D}\vec{g}_0) + o_p(K/n) \tag{2.16}$$

where $F(\lambda) = \mathbf{B}\hat{W}\mathbf{B}^T + n\lambda K^{2p}\tilde{D}$ ($\hat{W}$ is the estimator of the diagonal matrix $\mathbf{W}$ with the diagonal elements equals the conditional variances, $var(y_i|u)$), under the condition that $\lambda = O(n^\gamma)$, $\gamma \leq -(2p+1)/(2p+3)$ and $K \asymp n^{1/(2p+3)}$, the order of (j,l)-th entry of $F^{-1}(\lambda)$ is shown to be $\rho^{|j-l|}O[(n/K + n\lambda K^{2p})^{-1}]$, which is, $\rho^{|j-l|}n^{-(2p+2)/(2p+3)}$.

The first and second term in (2.16) can be seen as the shrinkage bias and estimation error. When η_0 is $(p+1)$-times continuously differentiable and $K \asymp n^{1/(2p+3)}$, $\|K^{2p}\tilde{D}\vec{g}_0\|_\infty = O(K^{p-1}) = O(n^{-(p-1)/(2p+3)})$ together with the rate for $F^{-1}(\lambda)$, $E(\hat{\vec{g}} - \vec{g}_0) = -F^{-1}(\lambda)n\lambda K^{2p}\tilde{D}\vec{g}_0\{1 + o(1)\} = O(n^{-(p+1)/(2p+3)})$, $var(\hat{\vec{g}}) = F^{-1}(\lambda)F(0)F^{-1}(\lambda)\{1 + o(1)\} = O(n^{-(p+1)/(2p+3)})$. The mean squared error of $\hat{\vec{g}} - \vec{g}_0$ is of the same order as MSE of $\hat{\vec{g}} - \vec{g}_0$, combining the results above, the latter is $O(n^{-(p+1)/(2p+3)})$.

On the other hand, the approximation bias $\eta_0(x) - B(x)\vec{g}_0$ is of the same rate $O(n^{-(p+1)/(2p+3)})$ when η_0 is $(p+1)$-times continuously differentiable and $K \asymp n^{1/(2p+3)}$.

Hence the mean-squared error for generalized penalized spline estimator $\hat{\eta}(x)$ is shown to be

$$MSE\{\hat{\eta}(x)\} = O(n^{-(2p+2)/(2p+3)}). \tag{2.17}$$

The authors also provide the asymptotic distribution as well as its bias and

variance expressions

$$bias\{\hat{\eta}(x)\} = -n\lambda B(x)F(\lambda)^{-1}K^{2p}\tilde{D}\vec{g}_0 - \delta(x) \tag{2.18}$$

and

$$var\{\hat{\eta}(x)\} = B(x)F^{-1}(\lambda)F(0)F^{-1}(\lambda)B(x)^T, \tag{2.19}$$

where $B(x)$ is the row vector of spline bases evaluated at x, $\delta(x)$ denotes the smallest approximation bias $\delta(x) = \eta_0(x) - B(x)\vec{g}_0$.

The paper also considers 1) the connection of mixed model and penalized approach and 2) the fully Bayesian approach with the mixed model.

Although the results in this paper are relatively complete, the conditions imposed are less satisfied. The knots are required to be equally spaced. Moreover, there is a "coupling of order" phenomenon, that is, $m = p + 1$, the original problem (2.13) is not dealing with the penalty, therefore q does not exist in this setting. Unlike the previous paper, it focuses on proving results under the small number of knots assumption, namely, the smoothing effect is small compared to the modeling bias. Also, to get the MSE (2.17), the rate at which the dimension of the spline basis grows is considered to be fixed $K \asymp n^{1/(2p+3)}$.

2.5 Claeskens et al. (2009)

This paper also considers the nonparametric regression model and the penalized spline estimator (1.7). Compared with other works, Claeskens et al. (2009) is the very first work that found a breakpoint in the (mean squared) rate of convergence as well as the bias and variance of penalized regression spline, either close to regression spline (1.8) or close to smoothing spline (1.9) depending on an explicitly defined

function of the number of knots, the sample size and the penalty parameter $K_q = (K + p + 1 - q)(n\lambda\tilde{c}_1)^{1/(2q)}n^{-1/(2q)}$. In this section, $\vec{h} = (h(x_1), \ldots, h(x_n))^T$ for any $h \in \mathbb{H}$.

The proof requires a lot of mathematical derivation, here we will sketch the outline. By solving the normal equation corresponding to (1.7), The penalized estimator (evaluated at the sample points) takes the form of a ridge regression estimator $\vec{\hat{f}} = \mathbf{B}^T(\mathbf{B}\mathbf{B}^T + n\lambda D_q)^{-1}\mathbf{B}Y$, where $\mathbf{B}$ is the same as defined in the previous section 2.4, D_q is the penalty matrix such that $\int (g^{(q)}(x))^2 dx = \vec{g}^T D_q \vec{g}$ for any $g \in \mathbb{G}$ with $\vec{g}$ as the coefficients of b-spline expansion. This expression can be simplified by making use of the following eigen decomposition

$$(\mathbf{B}\mathbf{B}^T)^{-1/2} D_q (\mathbf{B}\mathbf{B}^T)^{-1/2} = U \mathrm{diag}(s) U^T. \tag{2.20}$$

With (2.20), define $A = \mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1/2}U$, then we can rewrite the penalized spline estimator as

$$\vec{\hat{f}} = A\{I + n\lambda \mathrm{diag}(s)\}^{-1} A^T Y. \tag{2.21}$$

We shall point out that $\vec{\hat{f}}_{reg} = AA^T Y$. And we would denote $\vec{\tilde{f}} = A\{I + n\lambda \mathrm{diag}(s)\}^{-1} A^T \vec{f}_0$ as the conditional mean of $\vec{\hat{f}}$ given $x_1, \ldots, x_n$ and $\vec{\tilde{f}}_{reg} = AA^T f_0$ as the conditional mean of $\vec{\hat{f}}_{reg}$ given $x_1, \ldots, x_n$.

Furthermore the expression (2.21) can be used to obtain the average mean squared

error (AMSE)

$$\begin{aligned}\text{AMSE}(\hat{f}) &= \frac{1}{n}E\{(\vec{\hat{f}}-\vec{f_0})^T(\vec{\hat{f}}-\vec{f_0})\} \\ &= \frac{1}{n}E\{(\vec{\hat{f}}-\vec{\tilde{f}})^T(\vec{\hat{f}}-\vec{\tilde{f}})+\frac{1}{n}E\{(\vec{\tilde{f}}-\vec{\tilde{f}}_{reg})^T(\vec{\tilde{f}}-\vec{\tilde{f}}_{reg})\}\} \\ &\quad + E\{(\vec{\tilde{f}}_{reg}-\vec{f_0})^T(\vec{\tilde{f}}_{reg}-\vec{f_0})\} \\ &= \frac{\sigma^2}{n}\sum_j \frac{1}{(1+\lambda^* s_j)^2} + n\lambda)^2\sum_j \frac{s_j^2 b_j^2}{(1+\lambda^* s_j)^2} + \frac{1}{n}\vec{f_0}^T(I-AA^T)\vec{f_0}.\end{aligned} \tag{2.22}$$

The first, second and third term in the expression of AMSE correspond to the asymptotic variance, shrinkage bias and approximation bias separately.

The approximation bias $K^{-2(p+1)}$ has been constructed elsewhere under the somewhat stronger condition $f \in C^{p+1}$. The main theorem states the approximation bias as K^{-2q} when $f \in W^{2q}$, however, there is no clear evidence that the approximation bias holds for $f \in W^q$. To this point, the first two terms in the rate of convergence measured in AMSE (2.22) rely on the rate of s_j's. The author stated it as lemma A3 in their paper which says that $s_1 = \ldots = s_q = 0$, $s_j = n^{-1}(j-q)^{2q}\tilde{c}_1, j = q+1, \ldots, K+p+1$. The authors mentioned that this result on s_j is adapted from (2.5d) in Speckman (1985), where s_j represents the eigenvalue in the following sense, there exists $\{\varphi_1, \ldots, \varphi_n\}$ lying in the natural spline space of degree $2q-1$ such that

$$\begin{aligned}\textstyle\sum_{l=1}^n \varphi_i(x_l)\varphi_j(x_l) &= \delta_{ij} \\ \textstyle\int \varphi_i^{(q)}(x)\varphi_j^{(q)}(x)dx &= \delta_{ij}s_j \\ 0 = s_1 = \ldots, = s_q \le s_{q+1} \le \ldots \le s_n.&\end{aligned} \tag{2.23}$$

(2.23) differs from (2.20) in that the latter involves the basis in the b-spline space

of degree p that simultaneously diagonalizes the two functional in (2.23) and the number of basis in the latter case is $K + p + 1$ instead of n in the natural spline scenario. It seems that the two problems (2.20) and (2.23) are not the same.

Here are the main results in this paper. If $K_q < 1$ and $f_0 \in C^{p+1}[a, b]$

$$AMSE(\hat{f}) = O\left(\frac{K}{n}\right) + O(\lambda^2 K^{2q}) + O(K^{-2(p+1)}). \tag{2.24}$$

If $K_q \geq 1$ and $f_0 \in W^q[a, b]$

$$AMSE(\hat{f}) = O\left(\frac{1}{(n\lambda)^{1/(2q)}}\right) + O(\lambda) + O(K^{-2q}). \tag{2.25}$$

The expressions of asymptotic bias and variance also take different forms conditioning on the value of K_q. To save space, they will not be printed here.

We found the results in this paper interesting, however, imperfect. The reason lies in the following aspects. First, The design of data points in this paper is the same as that in Speckman (1985), which is quite stringent. It requires the empirical distribution of the deterministic design points to be converging to some distribution function uniformly with rate $O(K^{-1})$. Next, The AMSE of $\hat{f}$ in (2.24) builds upon a stronger smoothness condition on $f \in C^{p+1}$ (wanted W^{p+1} or even W^q), the extension from C^{p+1} to W^{p+1} does not apply by just following the remark in this paper. Specifically, the approximation bias in (2.22), (the expression of which can be found in equation (7) of Claeskens et al. (2009)), is built upon the assumption that $f \in C^{p+1}$. Although the authors mentioned that "according to Barrow and Smith (1978) the expression for the approximation bias (7) holds for $f \in W^{p+1}$ as well", we do not find the evidence from Barrow and Smith (1978). Last but not least, there exists an abuse of result for the eigenvalues (lemma A3), which might be the most

key results needed in deriving (2.24) and (2.25) and in calculating the higher order terms in the bias and variance expression of $\hat{f}$. The references for lemma A3, Utreras (1981), Speckman (1985), are intended for natural spline basis instead of b-spline. Hence the validity of lemma A3 is under question.

3. STATEMENT OF THE MASTER THEOREM

This chapter states the main theorem about the convergence rates of the penalized spline estimators for extended linear model defined in section 1.2. As mentioned in the introduction, Claeskens et al. (2009) considered the convergence rate of penalized spline estimator under the regression setting. Our result generalizes their result in two aspects. First of all, we will show that the same convergence rate holds for a wide set of models known as extended linear models and for arbitrary penalty terms under some straightforward regularity conditions. Second, we prove that the same convergence rate holds under a different norm, which is related to the RKHS space $\mathbb{H}$, and the norm, defined by $V(h) + \lambda J(h)$ or $(V + \lambda J)(h)$ in short, is stronger than the L_2 norm considered in Claeskens et al. (2009).

3.1 Regularity conditions

The main conditions needed for showing the convergence rates can be summarized into two categories. Some are imposed on the penalized log-likelihood, others are relevant to the expected log-likelihood. We begin with two such conditions:

Condition 3.1.1. *For any pair of g_1, $g_2 \in \mathbb{G}$, $l(g_1 + \alpha g_2)$ as a function of α is twice continuously differentiable. Moreover,*

(i)

$$\sup_{g \in \mathbb{G}} \frac{\left| -\frac{1}{2n}\frac{d}{d\alpha} l(\bar{\eta} + \alpha g)|_{\alpha=0} + \lambda J(\bar{\eta}, g) \right|}{(V + \lambda J)(g)^{1/2}} = O_P\left(\min\left\{ \left(\frac{1}{n\lambda^{1/2q}} \right)^{1/2}, \left(\frac{1}{n\delta} \right)^{1/2} \right\} \right).$$

Here $\bar{\eta}$ is the minimizer of $-\Lambda(\cdot) + \lambda J(\cdot)$, as defined in section 1.3.2.

(ii) For any $B > 0$, there exists a constant $M > 0$ (might depend on B) such that

$$\frac{d^2}{d\alpha^2} l(g_1 + \alpha g_2) \leq -M\|g_2\|^2, \quad 0 \leq \alpha \leq 1,$$

holds for any pair of $g_1, g_2 \in \mathbb{G}$ as long as $\|g_1\|_\infty \leq B$, $\|g_2\|_\infty \leq B$, with probability tending to one as $n \longrightarrow \infty$.

The next condition states the equivalence between $\frac{d^2}{d\alpha^2}\Lambda(h_1 + \alpha h_2)$ and $-\|h_2\|^2$.

Condition 3.1.2. *For each pair of $h_1, h_2 \in \mathbb{H}$, $\Lambda(h_1 + \alpha h_2)$ as a function of α is twice continuously differentiable. Furthermore, for any $B > 0$, there are constants $M_1 > 0$ and M_2 (might depend on B) such that*

$$-M_1\|h_2\|^2 \leq \frac{d^2}{d\alpha^2}\Lambda(h_1 + \alpha h_2) \leq -M_2\|h_2\|^2, \quad 0 \leq \alpha \leq 1,$$

holds for all $h_1, h_2 \in \mathbb{H}$ whenever $\|h_1\|_\infty \leq B$, $\|h_2\|_\infty \leq B$.

The main theorem could be established under the conditions above. For completeness, we need make it clear the circumstances under which the above conditions would hold. For this purpose, we will state several commonly used technical conditions in the smoothing spline literature (Gu (2013)), which simplifies the calculation when dealing with two quadratic functionals. Some of these conditions are satisfied for the previously defined $J(h) = \int\{h^{(q)}(x)\}^2 dx$ and $V(h) = \int h^2(x)\omega(x)dx$ (when ω meets condition 3.1.3), if this is the case, they will be phrased as propositions.

Condition 3.1.3. *The weight function $\omega(x)$ used in defining the norm $V(h)$ (equivalently $\|h\|$) is bounded away from zero and infinity, that is, there exists $c, C > 0$ such that*

$$c \leq \omega(x) \leq C, \quad \textit{for any} \quad a < x < b.$$

Definition 3.1.1. *A quardratic functional A is said to be completely continuous with respect to another quadratic functional B, if for any $\epsilon > 0$, there exists a finite number of linear functionals $L_1, \cdots, L_k$ such that $L_1(h) = \cdots = L_k(h) = 0$ implies that $A(h) \leq \epsilon B(h)$.*

Proposition 3.1.1. *Under condition 3.1.3, V is completely continuous with respect to J.*

By theorem 3.1 of Weinberger (1974), proposition 3.1.1 implies that V and J can be simultaneously diagonalized in the following sense. There exists a sequence of eigenfunctions $\phi_\nu \in \mathcal{H}$ and the associated nonnegative sequence of eigenvalues ρ_ν such that $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu\mu}$ where $\delta_{\nu\mu}$ is the Kronecker delta, $V(\phi_\nu, \phi_\mu) = \int \phi_\nu(x)\phi_\mu(x)\omega(x)dx$, $J(\phi_\nu, \phi_\mu) = \int \phi_\nu^{(q)}(x)\phi_\mu^{(q)}(x)dx$; see, e.g., Silverman (1982), Section 9.1 of Gu (2013). Furthermore, any function $h \in \mathbb{H}$ satisfying $J(h) < \infty$ can be expressed as the series expansion with basis $\{\phi_\nu\}$ defined above, $h = \sum_\nu h_\nu \phi_\nu$, where $h_\nu = V(h, \phi_\nu)$. Immediately with the Fourier series expansion, we are able to express $V(h)$ and $J(h)$ as $V(h) = \sum_\nu h_\nu^2$ and $J(h) = \sum_\nu \rho_\nu h_\nu^2$. Therefore, $(V + \lambda J)(h) = \sum_\nu (1 + \lambda\rho_\nu) h_\nu^2$.

Proposition 3.1.2. *Under condition 3.1.3, we have $\rho_\nu \asymp \nu^{2q}$ for sufficiently large ν.*

Proposition 3.1 and 3.2 are well known in the literature, readers could refer to Utreras (1981) or section 9.1 of Gu (2013).

We will close this section with a general result on the equivalence of some empirical and theoretical norm on $\mathbb{G}$ as supporting material in the process of verifying condition 3.1.1 and 3.1.2.

Let the empirical and theoretical inner product on $\mathbb{G}$ be defined by

$$< f_1, f_2 >_{n,\Psi} = E_n[\Psi(f_1, f_2, \mathbf{W})], \quad < f_1, f_2 >_{\Psi} = E[\Psi(f_1, f_2, \mathbf{W})]$$

where $\Psi(f_1, f_2; \mathbf{W})$ is a real valued functional that is symmetric and bilinear in its first two arguments: $\Psi(f_1, f_2; \mathbf{W}) = \Psi(f_2, f_1; \mathbf{W})$ and $\Psi(af_1 + bf_2, f; \mathbf{W}) = a\Psi(f_1, f; \mathbf{W}) + b\Psi(f_2, f; \mathbf{W})$. Furthermore, Ψ has the following property

$$\|\Psi(f_1, f_2, ; \mathbf{W})\|_\infty \leq M_5 \|f_1\|_\infty \|f_2\|_\infty$$

and

$$var[\Psi(f_1, f_2; \mathbf{W})] \leq M_6 \|f_1\|^2 \|f_2\|_\infty^2.$$

The norms corresponding to these inner products are denoted as $\|\cdot\|_{n,\Psi}$ and $\|\cdot\|_\Psi$. In the special case when $\mathbf{W}$ is a random variable X from a density $g_X(\cdot)$ taking value in $\mathcal{U}$ and $\Psi(f_1, f_2, ; X) = f_1(X)f_2(X)$, we get $< f_1, f_2 >_\Psi = V(f_1, f_2)$ with $\omega(\cdot) = g_X(\cdot)$. In this case, the two properties are satisfied with $W_5 = W_6 = 1$.

Condition 3.1.4. $\lim_n A_n^2 N/n = 0$.

Remark 3.1.1. *Under condition 3.1.4, the empirical and theoretical norm are asymptotically equivalent, in the sense that* $\sup_{g \in \mathbb{G}} |\|g\|_{n,\Psi}/\|g\|_\Psi - 1| = o_P(1)$. *Refer to Lemma 10 in Huang (1998b).*

3.2 Main theorem

Theorem 3.2.1. *Assuming conditions 3.1.1-3.1.2 are true. Suppose* $\lim_n A_n(\delta^m \vee \lambda^{1/2}) = 0$, *then* $\bar{\eta}$ *exists for* n *sufficiently large and* $(V + \lambda J)(\bar{\eta} - \eta_0) = O(\delta^{2m} + \lambda)$. *On the other hand, suppose* $\lim_n A_n^2(\frac{1}{n\lambda^{1/2q}} \wedge \frac{1}{n\delta}) = 0$, $(V + \lambda J)(\hat{\eta} - \bar{\eta}) = O_p\left(\min\left\{\frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta}\right\}\right)$. *Consequently,*

$$(V + \lambda J)(\hat{\eta} - \eta_0) = O_p\left(\delta^{2m} + \lambda + \min\left\{\frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta}\right\}\right).$$

Remark 3.2.1. *Consider the case when setting $q = m$, that is, $J(h) = \int\{h^{(m)}(x)\}^2 dx$. It is clearly shown in Theorem 3.2.1 that two scenarios could happen:*

(a) if $\lambda^{1/2m} < \delta$, the convergence rate, $\delta^{2m} + \frac{1}{n\delta}$, is similar to the spline estimator without penalty. The optimal rate of convergence $n^{-2m/(2m+1)}$ for $\eta_0 \in \mathbb{H}$ under L_2 norm is achieved when $\delta \asymp n^{-1/(2m+1)}$.

(b) if $\lambda^{1/2m} > \delta$, the convergence rate, $\lambda + \frac{1}{n\lambda^{1/2m}}$, is similar to the smoothing spline estimator. The optimal rate of convergence for $n^{-2m/(2m+1)}$ for $\eta_0 \in \mathbb{H}$ under the 'smoothing spline norm' is achieved when $\lambda \asymp n^{-2m/(2m+1)}$.

Remark 3.2.2. *The two scenarios in Remark 3.2.1 are consistent with the result shown in Claeskens et al. (2009) for the regression model, furthermore, it is shown here that the same convergence rate holds under the stronger norm $V + \lambda J$.*

Corollary 3.2.2. *Under the conditions given in the main theorem, it is easy to obtain that*

$$J(\hat{\eta} - \eta_0) = O_p\left(\delta^{2m}\lambda^{-1} + \min\left\{\frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta}\right\}\lambda^{-1}\right) + O_p(1).$$

3.3 Lemmas

The following lemmas will be used in the proofs. A detailed proofs can be found in section 9.2 of Gu (2013) for Lemma 3.3.1; in Theorem 6.25 of Schumaker (1981) for Lemma 3.3.2.

Lemma 3.3.1. *Under proposition 3.1.2, as $\lambda \to 0$, we have*

$$\sum_{\nu} \frac{1}{(1+\lambda\rho_\nu)^2} = O(\lambda^{-1/2q}), \tag{3.1}$$

$$\sum_{\nu} \frac{1}{1+\lambda\rho_\nu} = O(\lambda^{-1/2q}), \tag{3.2}$$

$$\sum_{\nu} \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} = O(\lambda^{-1/2q}). \tag{3.3}$$

Lemma 3.3.2. *For any function $\eta_0 \in \mathbb{H}$, there exists a function $\eta^* \in \mathbb{G}$ such that*

$$V(\eta^* - \eta_0) \le L_1^2 \delta^{2m}, \quad J(\eta^*) \le L_2^2$$

and

$$\|\eta^* - \eta_0\|_\infty \le L_3 \delta^{m-1/2}.$$

Here the constants L_1, L_2 and L_3 only depends on p and η_0.

Lemma 3.3.3. *Suppose that $\Lambda(\cdot)$ satisfies condition 3.1.2. Then for any $B > 0$, there exist positive constants M_3 and M_4 (might depend on B) such that $-M_3\|h-\eta_0\|^2 \le \Lambda(h) - \Lambda(\eta_0) \le -M_4\|h-\eta_0\|^2$ holds for all $h \in \mathbb{H}$ whenever $\|h\|_\infty \le B$.*

Proof. The proof can be easily derived by Taylor expansion (with integral remainder) of $\Lambda(h)$ at η_0 together with the fact that $\frac{d}{d\alpha}\Lambda(\eta_0 + \alpha(h-\eta_0))|_{\alpha=0} = 0$ and Condition 3.1.2. The proof can also be found in Huang (2001).

To be specific, the Taylor expansion of $\Lambda(h)$ at η_0 indicates

$$\Lambda(h) - \Lambda(\eta_0) = \frac{d}{d\alpha}\Lambda(\eta_0 + \alpha(h-\eta_0))|_{\alpha=0} + \int_0^1 (1-\alpha)\frac{d^2}{d\alpha^2}\Lambda(\eta_0 + \alpha(h-\eta_0))d\alpha.$$

The first order derivative $\frac{d}{d\alpha}\Lambda(\eta_0+\alpha(h-\eta_0))|_{\alpha=0}$ equals zero since η_0 is the minimizer of Λ in $\mathbb{H}$. The desired result follows from the above expression and condition

3.1.2. □

In the proof of the main theorem, we might as well need the following results that are of slight variations as statements in lemma 3.3.2, without loss of generality, we will not differentiate the constant in these cases, that is, there exists a L depending on p and η_0 such that

$$V(\eta^* - \eta_0) + \lambda J(\eta^*) \leq L^2(\delta^{2m} + \lambda) \tag{3.4}$$

$$V(\eta^* - \eta_0) + \lambda J(\eta^* - \eta_0) \leq L^2(\delta^{2m} + \lambda) \tag{3.5}$$

$$\|\eta^*\|_\infty \leq \|\eta_0\|_\infty + L\delta^{m-1/2}. \tag{3.6}$$

(3.4) comes directly from the first two inequalities of lemma 3.3.2; (3.5) follows from $V(\eta^* - \eta_0) + \lambda J(\eta^* - \eta_0) \leq V(\eta^* - \eta_0) + \lambda J(\eta^*) + \lambda J(\eta_0)$ and the fact that $J(\eta_0)$ is bounded; (3.3.2) is due to the triangle inequality $\|\eta^*\|_\infty \leq \|\eta_0\|_\infty + \|\eta^* - \eta_0\|_\infty$ and the third inequality of lemma 3.3.2.

Now we are ready to prove the rate of convergence of penalized spline estimator.

4. PROOF OF THE MASTER THEOREM

This chapter aims to prove the main theorem of convergence rates of the penalized spline estimator, that is, Theorem 3.2.1, in detail. Thanks to the convexity in all extended linear models and the smoothness penalty, we are able to show it in a general and uniform way.

Before preceding, we will state a lemma regarding the property of any convex functional since it appears more than once in our proof.

Lemma 4.0.4. *Suppose $L(\cdot)$ is a convex functional defined on a linear space $\mathbb{N}$. $\|\cdot\|_{psuedo}$ is a psuedo-metric associated with the linear space $\mathbb{N}$ meaning that for any $\eta_1, \eta_2 \in \mathbb{N}$, $|\|\eta_1\|_{psuedo} - \|\eta_2\|_{psuedo}| \leq \|\eta_1 - \eta_2\|_{\mathbb{N}}$, where $\|\cdot\|_{\mathbb{N}}$ is a norm defined on $\mathbb{N}$. If there exists a function η^* with $\|\eta^*\|_{psuedo} < s$ such that for all $\eta \in \mathbb{N}$ satisfying $\|\eta\|_{psuedo} = s$, we have*

$$L(\eta^*) < L(\eta). \tag{4.1}$$

Then $L(\tilde{\eta}) > L(\eta^*)$, for all $\tilde{\eta} \in \mathbb{N}$ with $\|\tilde{\eta}\|_{psuedo} > s$. Or we can say that η_m is the minimizer of $L(\cdot)$, then $\|\eta_m\|_{psuedo} < s$.

Proof. The proof is done by contradiction. Suppose the statement is incorrect, that is, there exists an $\tilde{\eta} \in \mathbb{N}$ with $\|\tilde{\eta}\|_{\text{psuedo}} > s$ such that $L(\tilde{\eta}) \leq L(\eta^*)$.

Consider the convex combination of η^* and $\tilde{\eta}$ as $\eta_\alpha \doteq \alpha\eta^* + (1-\alpha)\tilde{\eta}$, $0 \leq \alpha \leq 1$. Define $C(\alpha) \doteq \|\eta_\alpha\|_{\text{psuedo}}$. Since $C(\alpha_1) - C(\alpha_2) \leq (\alpha_1 - \alpha_2)\|(\eta^* - \tilde{\eta})\|_{\mathbb{N}}$, hence $C(\alpha)$ is continuous in α, and it is easy to check that $C(0) > s, C(1) < s$, then there exists an $\breve{\alpha} \in (0,1)$ such that $C(\breve{\alpha}) = s$. Denote $\breve{\eta} \doteq \breve{\alpha}\eta^* + (1-\breve{\alpha})\tilde{\eta}$, immediately $\|\breve{\eta}\|_{\text{psuedo}}$ = s. We conclude from the convexity of $L(\cdot)$ and the property of $\breve{\eta}$ that

$$L(\breve{\eta}) \leq \breve{\alpha}L(\eta^*) + (1-\breve{\alpha})L(\tilde{\eta}) \leq L(\eta^*).$$

This violates the condition given in (4.1). We complete the proof by contradiction.

□

Proof of Theorem 3.2.1. The proof of Theorem 3.2.1 borrows some idea from Huang (2001). In order to prove the convergence rate of $\hat{\eta} - \eta_0$, we will decompose $\hat{\eta} - \eta_0 = \hat{\eta} - \bar{\eta} + \bar{\eta} - \eta_0$, where the convergence rate of the first (second) term is the approximation (estimation) error. Here $\bar{\eta}$ is the minimizer corresponding to

$$\bar{\eta} = \text{argmin}_{\eta \in \mathbb{G}} \bigg(-\Lambda(\eta) + \lambda J(\eta) \bigg). \tag{4.2}$$

I. Approximation error:

Suppose $a > 1$ (to be determined later), take $\eta \in \mathbb{G}$ satisfying

$$V^{1/2}(\eta - \eta_0) + \lambda^{1/2} J^{1/2}(\eta) \leq a(\delta^m + \lambda^{1/2}) \tag{4.3}$$

and η^* as in Lemma 3.3.2.

Then

$$\begin{aligned} \|\eta - \eta^*\|_\infty &\leq A_n V^{1/2}(\eta - \eta^*) \\ &\leq A_n (V^{1/2}(\eta - \eta_0) + V^{1/2}(\eta^* - \eta_0)) \\ &\leq A_n C_{a,L_1}(\delta^m + \lambda^{1/2}), \end{aligned} \tag{4.4}$$

where C_{a,L_1} is a constant depending on a and L_1 only. The last inequality in (4.4) is due to (4.3) and lemma 3.3.2.

For any $\eta \in \mathbb{G}$ satisfying (4.3),

$$\begin{aligned} \|\eta\|_\infty &\le \|\eta - \eta^*\|_\infty + \|\eta^* - \eta_0\|_\infty + \|\eta_0\|_\infty \\ &\le A_n C_{a,L_1}(\delta^m + \lambda^{1/2}) + C(\delta^{m-1/2}) + \|\eta_0\|_\infty \\ &< M\|\eta_0\|_\infty. \end{aligned} \tag{4.5}$$

holds for all $\eta \in \mathbb{G}$ in (4.3). The second inequality can be obtained from (4.4) and the third inequality in lemma 3.3.2. The last inequality in (4.5) can be demonstrated under the fact that $\lim_n A_n\delta^m = 0$, $\lim_n A_n\lambda^{1/2} = 0$ and $\lim_n \delta = 0$.

Then we can apply Lemma 3.3.3 to η satisfying equality in (4.3) to obtain

$$\begin{aligned} -\Lambda(\eta) + \Lambda(\eta_0) + \lambda J(\eta) &\ge M_4 V(\eta - \eta_0) + \lambda J(\eta) \\ &\ge \min(M_4, 1)(V(\eta - \eta_0) + \lambda J(\eta)) \\ &\ge \frac{a^2}{2}\min(M_4, 1)(\delta^{2m} + \lambda). \end{aligned} \tag{4.6}$$

The last inequality can be justified using $V(\eta - \eta_0) + \lambda J(\eta) \ge \frac{1}{2}(V^{1/2}(\eta - \eta_0) + \lambda^{1/2}J^{1/2}(\eta))^2 = \frac{a^2}{2}(\delta^m + \lambda^{1/2})^2 \ge \frac{a^2}{2}(\delta^{2m} + \lambda)$.

On the other hand, it is easy to verify by (3.6) and $\lim_n \delta = 0$ that

$$\|\eta^*\|_\infty < \|\eta_0\|_\infty + O(\delta^{m-1/2}) \le M\|\eta_0\|_\infty.$$

Applying Lemma 3.3.3 again to η^* and then (3.4) shows

$$\begin{aligned} -\Lambda(\eta^*) + \Lambda(\eta_0) + \lambda J(\eta^*) &\le M_3 V(\eta^* - \eta_0) + \lambda J(\eta^*) \\ &\le (M_3 + 1)(V(\eta^* - \eta_0) + \lambda J(\eta^*)) \\ &\le (M_3 + 1)L^2(\delta^{2m} + \lambda). \end{aligned} \tag{4.7}$$

Take $a > \sqrt{\frac{2(M_3+1)}{\min(M_4,1)}}L$. The right hand side of (4.6) is greater (not equal) than that of (4.7). Thus, for all η such that $V^{1/2}(\eta-\eta_0)+\lambda^{1/2}J^{1/2}(\eta) = a(\delta^m+\lambda^{1/2})$ and this choice of a,

$$-\Lambda(\eta)+\lambda J(\eta) > -\Lambda(\eta^*)+\lambda J(\eta^*).$$

Thus we can apply lemma 4.0.4 to the convex functional $-\Lambda(\cdot)+\lambda J(\cdot)$ and the psuedo metric $V^{1/2}(\cdot-\eta_0)+\lambda^{1/2}J^{1/2}(\cdot)$ to conclude that

$$V^{1/2}(\bar{\eta}-\eta_0)+\lambda^{1/2}J^{1/2}(\bar{\eta}) < a(\delta^m+\lambda^{1/2}). \tag{4.8}$$

Furthermore, it follows immediately that

$$V(\bar{\eta}-\eta_0)+\lambda J(\bar{\eta}-\eta_0) \le V(\bar{\eta}-\eta_0)+\lambda J(\bar{\eta})+\lambda J(\eta_0) < a^2(\delta^{2m}+\lambda)+\lambda = O(\delta^{2m}+\lambda). \tag{4.9}$$

Thus the proof for the approximation error is finished.

As a by-product, we could derive the upper bound for $\|\bar{\eta}\|_\infty$. Since $\bar{\eta}$ satisfies (4.3) we would deduce from (4.5) that

$$\|\bar{\eta}\|_\infty < M\|\eta_0\|_\infty. \tag{4.10}$$

II. Estimation error:

Taking Taylor Expansion to the optimization function at $\bar{\eta}$ we will get

$$\begin{aligned} -l(\eta)+\lambda J(\eta) = -l(\bar{\eta})+\lambda J(\bar{\eta}) - \frac{d}{d\alpha}l(\bar{\eta}+\alpha(\eta-\bar{\eta}))|_{\alpha=0} + 2\lambda J(\bar{\eta},\eta-\bar{\eta}) \\ - \int_0^1 (1-\alpha)\frac{d^2}{d\alpha^2}l(\bar{\eta}+\alpha(\eta-\bar{\eta}))d\alpha + \lambda J(\eta-\bar{\eta}). \end{aligned} \tag{4.11}$$

Suppose $a > 1$(to be determined later), in the following derivation we consider

$\eta \in \mathbb{G}$ satisfying

$$(V + \lambda J)(\eta - \bar{\eta}) = a^2 \min\left\{ \frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta} \right\}. \tag{4.12}$$

Then by the definition of A_n and any η such that (4.12) holds, $\|\eta - \bar{\eta}\|_\infty \leq A_n \|\eta - \bar{\eta}\| = A_n a \min\left\{ \frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta} \right\}^{1/2} = o(1)$. Consequently, for n sufficiently large,

$$\|\eta\|_\infty \leq \|\eta - \bar{\eta}\|_\infty + \|\bar{\eta}\|_\infty \leq B \tag{4.13}$$

for all η satisfying (4.12). The second inequality also makes use of the (4.10).

We will take a closer look at (4.11). By condition 3.1.1 (ii) and (4.12), the summation of the last two terms in (4.11) has a lower bound

$$\begin{aligned} & - \int_0^1 (1 - \alpha) \frac{d^2}{d\alpha^2} l(\bar{\eta} + \alpha(\eta - \bar{\eta})) d\alpha + \lambda J(\eta - \bar{\eta}) \\ & \geq \min\{M/2, 1\}(V + \lambda J)(\eta - \bar{\eta}) \\ & = a^2 \min\{M/2, 1\} \min\left\{ \frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta} \right\}. \end{aligned} \tag{4.14}$$

On the other hand, it is also given in condition 3.1.1 (i) that the sum of the middle two terms in (4.11) has an upper bound

$$\begin{aligned} & \left| -\frac{d}{d\alpha} l(\bar{\eta} + \alpha(\eta - \bar{\eta}))|_{\alpha=0} + 2\lambda J(\bar{\eta}, \eta - \bar{\eta}) \right| \\ & = (V + \lambda J)(\eta - \bar{\eta})^{1/2} O_p\left(\min\left\{ \left(\frac{1}{n\lambda^{1/2q}} \right)^{1/2}, \left(\frac{1}{n\delta} \right)^{1/2} \right\} \right) \\ & = a O_p\left(\min\left\{ \frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta} \right\} \right). \end{aligned} \tag{4.15}$$

Thus we can choose a so large that the right hand side of (4.15) is bouned above by

the lower bound in (4.14), except on an event whose probability is less than ϵ,

$$\left|-\frac{d}{d\alpha}l(\bar{\eta}+\alpha(\eta-\bar{\eta}))|_{\alpha=0}+\lambda J(\bar{\eta},\eta-\bar{\eta})\right|<a^2\min\{M/2,1\}\left(\min\left\{\frac{1}{n\lambda^{1/2q}},\frac{1}{n\delta}\right\}\right). \tag{4.16}$$

Now with (4.11) (4.14) and (4.16) we are able to demonstrate that with probability tending to one

$$-l(\eta)+\lambda J(\eta)>-l(\bar{\eta})+\lambda J(\bar{\eta})$$

holds for all η satisfying (4.12).

Apply lemma 4.0.4 to the convex functional $-l(\cdot)+\lambda J(\cdot)$ and the psuedo metric $(V+\lambda J)^{1/2}(\cdot-\bar{\eta})$, $\hat{\eta}$ has the following property, with probability tending to one

$$(V+\lambda J)(\hat{\eta}-\bar{\eta})<a^2\min\left\{\frac{1}{n\lambda^{1/2q}},\frac{1}{n\delta}\right\}. \tag{4.17}$$

This completes the proof of rate of convergence for estimation error. □

5. APPLICATION OF THE MASTER THEOREM TO VARIOUS SETTINGS

We have established the general convergence rate theorem in chapter 3 under some regularity conditions. Before concluding on the validity of such a result for the extended linear models, we would like to show that the regularity conditions are not very stringent. In this chapter, we will verify that condition 3.1.1 and condition 3.1.2 are valid for several regular models, which garantees that our main theorem can be applied to a broad set of models.

To begin with, we state two useful lemmas, which synchronize the main techniques used in checking condition 3.1.1 (i). In the proof of lemma 5.0.6, we will make use of one simple observation that under Condition 3.1.3, the previously defined norm $\|\cdot\|$ or $V(\cdot)$ on $\mathbb{H}$ are equivalent to L_2 norm.

Lemma 5.0.5. *For any $g \in \mathbb{G}$, we have that*

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)|_{\alpha=0}+\lambda J(\bar{\eta},g)=-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)|_{\alpha=0}+\frac{1}{2}\frac{d}{d\alpha}\Lambda(\bar{\eta}+\alpha g)|_{\alpha=0}.$$

Proof. By the definition of $\bar{\eta}$, it is the minimizer of the convex functional $-\Lambda(\eta)+\lambda J(\eta)$. Therefore, $\bar{\eta}$ satisfies the first order condition

$$-\frac{d}{d\alpha}\Lambda(\bar{\eta}+\alpha g)|_{\alpha=0}+2\lambda J(\bar{\eta},g)=0.$$

The desired result is obtained simply by substituting $\lambda J(\bar{\eta},g)$ with $\frac{1}{2}\frac{d}{d\alpha}\Lambda(\bar{\eta}+\alpha g)|_{\alpha=0}$.

□

Lemma 5.0.6. *Let $\{h_n\}$ be a sequence of functions in $\mathbb{H}$ with $\|h_n\|_\infty \le M$ for some positive constant M and any $n \ge 1$. $Q_n(f,g)=\frac{1}{n}\sum_{i=1}^n q(f,g;w_i)$ is a quadratic*

functional defined on $\mathbb{H}$, $w_i's$ *are i.i.d. observation of some random variable (possibly vector valued)* w. *Also* $Q_n(f,g)$ *is linear in* g. *Denote* $Q(f,g) = E(Q_n(f,g))$. *Suppose* $E(q(h_n, \phi_\nu; w_1)^2) \leq K_1$, *for some constant* $K_1 > 0$ *that does not depend on* ϕ_ν *and* n, *then*

$$sup_{g\in\mathbb{G}}\left|\frac{Q_n(h_n,g) - Q(h_n,g)}{(V+\lambda J)(g)^{\frac{1}{2}}}\right| = O_p\left(\left(\frac{1}{n\lambda^{1/2q}}\right)^{1/2}\right). \tag{5.1}$$

Furthermore, if we have $E(q(h_n, b_{k,p}; w_1)^2) \leq K_2\delta$, *for some constant* $K_2 > 0$ *that does not depend on* $b_{k,p}$, *then*

$$sup_{g\in\mathbb{G}}\left|\frac{Q_n(h_n,g) - Q(h_n,g)}{(V+\lambda J)(g)^{\frac{1}{2}}}\right| = O_p\left(\left(\frac{1}{n\delta}\right)^{1/2}\right). \tag{5.2}$$

Proof. First let us assume $E(q(h_n, \phi_\nu; w_1)^2) \leq K_1$. Taking eigen decomposition of $g = \sum_\nu g_\nu\phi_\nu$ and by Cauchy Schwartz inequality we would obtain that

$$\begin{aligned} |Q_n(h_n,g) - Q(h_n,g)| &= |\sum_\nu g_\nu(Q_n(h_n,\phi_\nu) - Q(h_n,\phi_\nu))| \qquad (5.3)\\ &\leq \left[\sum_\nu g_\nu^2(1+\lambda\rho_\nu)\right]^{1/2}\left[\sum_\nu \frac{(Q_n(h_n,\phi_\nu) - Q(h_n,\phi_\nu))^2}{1+\lambda\rho_\nu}\right]^{1/2}. \end{aligned}$$

Since the first term on the right hand side of (5.3) is the same as the denumerator of (5.1), it remains to show the upper bound for the second term of (5.3).

It is not hard to establish the upper bound for the expectation of the random

component in each summand.

$$E(Q_n(h_n,\phi_\nu)-Q(h_n,\phi_\nu))^2 = \mathrm{Var}(Q_n(h_n,\phi_\nu)) \tag{5.4}$$
$$= \frac{1}{n}\mathrm{Var}\bigg(q(h_n,\phi_\nu,w_1)\bigg)$$
$$\leq \frac{1}{n}E\bigg(q(h_n,\phi_\nu,w_1)\bigg)^2 \leq \frac{K_1}{n}.$$

Notice that (5.4) indicates

$$E\bigg[\sum_\nu \frac{(Q_n(h_n,\phi_\nu)-Q(h_n,\phi_\nu))^2}{1+\lambda\rho_\nu}\bigg] \leq \frac{K_1}{n}\bigg(\sum_\nu \frac{1}{1+\lambda\rho_\nu}\bigg). \tag{5.5}$$

(5.5) together with (3.1) in Lemma 3.3.1 shows the upper bound for the second term of (5.3) is $O_p\bigg(\bigg(\frac{1}{n\lambda^{1/2q}}\bigg)^{1/2}\bigg)$. Notice that this upper bound does not depend on g, it would also be the upper bound when taking supreme over $g \in \mathbb{G}$, hence (5.1) holds.

On the other hand, if $E(q(h_n,b_{k,p};w_1)^2) \leq K_2\delta$ holds. Taking basis expansion of $g=\sum_k g_k b_{k,p}$ and by Cauchy Schwartz inequality we would obtain that

$$|Q_n(h_n,g)-Q(h_n,g)| = |\sum_k g_k(Q_n(h_n,b_{k,p})-Q(h_n,b_{k,p}))| \tag{5.6}$$
$$\leq \bigg[\delta\sum_k g_k^2\bigg]^{1/2}\bigg[\delta^{-1}\sum_k (Q_n(h_n,b_{k,p})-Q(h_n,b_{k,p}))^2\bigg]^{1/2}.$$

Due to lemma 1.3.1

$$\delta\sum_k g_k^2 \leq CV(g) \leq C(V+\lambda J)(g), \tag{5.7}$$

which means the first term on the right hand side of (5.6) is bounded by the denumerator of (5.2), it remains to show the upper bound for the second term of (5.6).

Repeat (5.4) and (5.5) to $Q_n(h_n,b_{k,p})-Q(h_n,b_{k,p})$ and the summation $\sum_k(Q_n(h_n,b_{k,p})-$

$Q(h_n, b_{k,p}))^2$, one would get

$$E\left[\delta^{-1}\sum_k (Q_n(h_n, b_{k,p}) - Q(h_n, b_{k,p}))^2\right] = O_p\left(\frac{1}{n\delta}\right). \tag{5.8}$$

Therefore, (5.2) follows from (5.6), (5.7) and (5.8). □

As we have said, the required conditions 3.1.1, 3.1.2 for our main theorem are not very stringent, in the following sections, we are going to devote some effort in checking them for some specific models. To do this, we need some model specific conditions together with lemma 5.0.5 and lemma 5.0.6 above. For most of the models below, it is of little necessity to verify condition 3.1.1 (ii) and condition 3.1.2 (which do not involve the penalty functional) here, since they have been demonstrated elsewhere, the reference will be pointed out later on.

5.1 Regression model

For the regression model $y_i = \eta_0(x_i) + \epsilon_i$, its conventional (negative) sum of squares criterion is defined as

$$l(\eta) = -\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \eta(x_i)\Big)^2.$$

If in addition $\epsilon_i \sim N(0, \sigma^2)$, then $l(\eta)$ coincides with the (conditional) log-likelihood, otherwise $l(\eta)$ can be treated as a psuedo log-likelihood.

Suppose that $x_i's$ are $i.i.d$ random sample from density $f_X(\cdot)$. And its expected log-likelihood would be (up to a constant)

$$\Lambda(\eta) = -\int (\eta(x) - \eta_0(x))^2 f_X(x)dx \doteq -V(\eta - \eta_0).$$

The above formulation leads to a concave extended linear model with random

variable $\mathbf{W}$ containing a random pair, that is, $\mathbf{W} = (X, Y)$.

By lemma 5.0.5,

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)|_{\alpha=0} + \lambda J(\bar{\eta}, g) = \frac{1}{n}\sum_{i=1}^{n}\Big(\bar{\eta}(x_i) - y_i\Big)g(x_i) - V(\bar{\eta}-\eta_0, g). \quad (5.9)$$

Immediately we have condition 3.1.3 is equivalent to

Condition 5.1.1.

$$c \le f_X(x) \le C, \quad \textit{for any} \quad a < x < b.$$

Define $Q_n(\eta, g) = \frac{1}{n}\sum_{i=1}^{n} q(\eta, g; w_i) = \frac{1}{n}\sum_{i=1}^{n}(\eta(x_i) - y_i)g(x_i)$, then $Q(\eta, g) = V(\eta - \eta_0, g)$. With these definitions, (5.9) can be rewritten as

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)|_{\alpha=0} + \lambda J(\bar{\eta}, g) = Q_n(\bar{\eta}, g) - Q(\bar{\eta}, g). \quad (5.10)$$

Because of (5.10), condition 3.1.1 (i) is just a direct application of lemma 5.0.6. The rest is to show the conditions in lemma 5.0.6 holds, namely,

$$E(q(\bar{\eta}, \phi_\nu; w_1)^2) < K_1 \quad \text{and} \quad E(q(\bar{\eta}, b_{k,p}; w_1)^2) < K_2\delta. \quad (5.11)$$

$$\begin{aligned} E(q(\bar{\eta}, \phi_\nu; w_1)^2) &= E\left[\Big(\bar{\eta}(x_1) - y_1\Big)^2 \phi_\nu(x_1)^2\right] \\ &= E\left[\Big(\bar{\eta}(x_1) - \eta_0(x_1)\Big)^2 \phi_\nu(x_1)^2\right] + E\left[\epsilon_1^2 \phi_\nu(x_1)^2\right] \\ &\le (M+1)^2\|\eta_0\|_\infty^2 + \sigma^2 \doteq K_1. \end{aligned}$$

The last inequality comes from (4.10) and $V(\phi_\nu^2) = E(\phi_\nu^2) = 1$.

Similarly, pointed out by lemma 1.3.1 $E(b_{k,p}^2) \leq C\delta$, the above derivation regarding $q(\bar{\eta}, \phi_\nu; \omega_1)$ can be applied to $q(\bar{\eta}, b_{k,p}; \omega_1)$, hence

$$E(q(\bar{\eta}, b_{k,p}; w_1)^2) \leq C\delta[(M+1)^2\|\eta_0\|_\infty^2 + \sigma^2] \cdot K_2.$$

Thus (5.11) holds.

Since

$$\frac{d^2}{d\alpha^2} l(g_1 + \alpha(g_2 - g_1)) = -\frac{2}{n}\sum_{i=1}^{n}(g_2(x_i) - g_1(x_i))^2.$$

The empirical norm on the right hand side of the above equation is equivalent to the theoretical norm in the sense that $\frac{1}{n}\sum_{i=1}^{n} g(x_i)^2 = (1 + o_p)V(g)$ as long as condition 3.1.4 is met. Condition (3.1.1) (ii) follows.

Similarly

$$\frac{d^2}{d\alpha^2}\Lambda(h_1 + \alpha(h_2 - h_1)) = -2V(h_2 - h_1).$$

Therefore, it is obvious that condition 3.1.2 is satisfied.

5.2 Generalized regression model

Like the nonparametric regression model in the previous section, the generalized regression model involves a random pair $W = (X, Y)$, where X is a $\mathcal{U}$-valued covariate and Y is a real-valued response variable. In this context, the conditional distribution of Y given X is characterized.

$$P(Y \in dy | X = x) = \exp[B(\eta_0(x))y - C(\eta_0(x))]\Phi(dy), \tag{5.12}$$

where $C(\cdot)$ is a normalizing coefficient function, the dependence of Y on X is through the function of interest η_0. The generalized regression model is a natural extension

of generalized linear model, where η_0 is supposed to be linear. The generalized regression model itself contains a variety of popular models (see below), $B(\cdot)$ and $C(\cdot)$ are known functions that take certain form for any specific model.

We would need some conditions on $B(\cdot)$ and $C(\cdot)$ such that the regularity conditions for the main theorem is valid.

Condition 5.2.1. *$B(\cdot)$ is twice continuously differentiable and its first derivative is strictly positive. Φ is a nonzero reference measure on $\mathbb{R}$ not degenerated at a single point.*

Condition 5.2.2. *S_1 is a subintervals of $\mathbb{R}$ such that Φ is concentrated on S_1 the following inequality holds*

$$B^{''}(\zeta)y - C^{''}(\zeta) < 0,$$

for all $y \in S_1$ and $\zeta \in \mathbb{R}$.

Remark 5.2.1. *Under condition 5.2.1, $C(\cdot)$ is also twice continuously differentiable. Moreover, assuming condition 5.2.2, we could deduce that $B^{'}(\cdot)$, $B^{''}(\cdot)$, $C^{'}(\cdot)$ and $C^{''}(\cdot)$ are bounded functions on any compact subinterval S_2.*

The generalized regression model (5.12) belongs to a concave extended linear model with $\mathbf{W} = (X, Y)$. As special cases, we get logistic regression, Poisson regression models, which are introduced in the following examples. We will also pay attention to whether conditions 5.2.1 and 5.2.2 hold for these specific examples.

Example 5.2.1 (Logistic Regression)**.** *Y is a binary $\{0, 1\}$-valued response variable. The probability of $Y = 1$ usually depends on the covariate X, thus $P(Y = 1|X = x)$ shall be characterized. The logistic regression takes the logistic transform of $P(Y =$*

$1|X = x)$ *as a function of* x,

$$\log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = \eta_0(x).$$

Solving the above equality for $P(Y=1|X=x)$ *we could get*

$$P(Y=1|X=x) = \frac{\exp\eta_0(x)}{1+\exp\eta_0(x)}.$$

Equivalently,

$$P(Y=y|X=x) = \exp\{\eta_0(x)y - \log(1+\exp\eta_0(x))\}, y = 0, 1$$

This is a special case of generalized regression model (5.12) with $B(s) = s$ *and* $C(s) = \log(1+\exp(s))$. *Condition 5.2.1 and 5.2.2 are satisfied. For this special case, the boundedness of* $B^{'}(\cdot)$, $B^{''}(\cdot)$, $C^{'}(\cdot)$ *and* $C^{''}(\cdot)$ *are always guaranteed so there is no need to assume that the range of* η_0 *is a subinterval of* $\mathbb{R}$.

Example 5.2.2 (Poisson Regression)**.** *Poisson regression assumes that* Y *representing count data has a Poisson distribution whose intensity parameter (expected value) depends on* X. *Therefore,*

$$P(Y=y|X=x) = \frac{\lambda_0(x)^y}{y!}\exp(-\lambda_0(x)), y = 0, 1, 2, \ldots.$$

To get around the positive constraint on $\lambda_0(\cdot)$, *it is convenient to reparameterize* $\lambda_0(\cdot) = \exp(\eta_0(\cdot))$. *The conditional probability can be easily reformulated as*

$$P(Y=y|X=x) = \exp[y\eta_0(x) - \log(y!) - \exp(\eta_0(x))].$$

Clearly, Poisson regression belongs to the generalized regression model with $B(s) = s$ and $C(s) = \log(y!) + \exp(s)$. It is very easy to verify that the conditions 5.2.1 and 5.2.2 hold.

The conditional mean and variance of Y given $X = x$ have closed form expressions

$$\mu(x) = E(Y|X = x) = \frac{C^{'}(\eta_0(x))}{B^{'}(\eta_0(x))}, \tag{5.13}$$

$$V(x) = \mathrm{Var}(Y|X = x) = \frac{C^{''}(\eta_0(x))}{B^{'}(\eta_0(x))^2} - C^{'}(\eta_0(x))\frac{B^{''}(\eta_0(x))}{B^{'}(\eta_0(x))^3}. \tag{5.14}$$

Given $i.i.d$ samples $w_1 = (x_1, y_1), \ldots, w_n = (x_n, y_n)$, the scaled (conditional) log-likelihood for generalized regression at a candidate function η is given by

$$l(\eta) = \frac{1}{n}\sum_{i=1}^{n}\Big[B(\eta(x_i))y_i - C(\eta(x_i))\Big].$$

And its expectation

$$\Lambda(\eta) = E\Big[B(\eta(x_1))\frac{C^{'}(\eta_0(x_1))}{B^{'}(\eta_0(x_1))} - C(\eta(x_1))\Big].$$

By lemma 5.0.5,

$$\begin{aligned} -\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta} + \alpha g)|_{\alpha=0} + \lambda J(\bar{\eta}, g) = \frac{1}{2n}\sum_{i=1}^{n}[C^{'}(\bar{\eta}(x_i)) - B^{'}(\bar{\eta}(x_i))y_i]g(x_i) \\ + \frac{1}{2}E\Big\{\Big[B^{'}(\bar{\eta}(x_1))\frac{C^{'}(\eta_0(x_1))}{B^{'}(\eta_0(x_1))} - C^{'}(\bar{\eta}(x_1))\Big]g(x_1)\Big\}. \end{aligned} \tag{5.15}$$

Define $Q_n(\eta, g) = \frac{1}{2n}\sum_{i=1}^{n} q(\eta, g; w_i) = \frac{1}{2n}\sum_{i=1}^{n}[B^{'}(\eta(x_i))y_i - C^{'}(\eta(x_i))]g(x_i)$, then $Q(\eta, g) = \frac{1}{2}E\Big\{\Big[B^{'}(\eta(x_1))\frac{C^{'}(\eta_0(x_1))}{B^{'}(\eta_0(x_1))} - C^{'}(\eta(x_1))\Big]g(x_1)\Big\}$.

With these definitions, (5.15) can be rewritten as

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)|_{\alpha=0}+\lambda J(\bar{\eta},g)=Q_n(\bar{\eta},g)-Q(\bar{\eta},g). \tag{5.16}$$

Because of (5.16), condition 3.1.1 (i) is just a direct application of lemma 5.0.6. The rest is to show that

$$E(q(\bar{\eta},\phi_\nu;w_1)^2)<K_1 \quad \text{and} \quad E(q(\bar{\eta},b_{k,p};w_1)^2)<K_2\delta. \tag{5.17}$$

$$\begin{aligned}
&E(q(\bar{\eta},\phi_\nu;w_1)^2)\\
&=E([B^{'}(\bar{\eta}(x_1))y_1-C^{'}(\bar{\eta}(x_1))]^2\phi_\nu(x_1)^2)\\
&=E([B^{'}(\bar{\eta}(x_1))(y_1-\mu(x_1))+B^{'}(\bar{\eta}(x_1))\mu(x_1)-C^{'}(\bar{\eta}(x_1))]^2\phi_\nu(x_1)^2)\\
&=E[B^{'}(\bar{\eta}(x_1))^2V(x_1)\phi_\nu(x_1)^2]+E([B^{'}(\bar{\eta}(x_1))\mu(x_1)-C^{'}(\bar{\eta}(x_1))]^2\phi_\nu(x_1)^2)\\
&\leq K_1E(\phi_\nu(x_1)^2)=K_1.
\end{aligned}$$

The second but last inequality makes use of the boundedness of $B^{'}(\cdot), C^{'}(\cdot), B^{''}(\cdot)$ and $C^{''}(\cdot)$, under which $\mu(\cdot)$ and $V(\cdot)$ defined in (5.13) and (5.14) are also bounded. Therefore, the first claim in (5.17) follows.

Similarly, as pointed out by lemma 1.3.1 $E(b_{k,p}^2)\leq C\delta$, replacing ϕ_ν in the above derivation with $b_{k,p}$ we would get

$$E(q(\bar{\eta},b_{k,p};w_1)^2)\leq K_2\delta.$$

Thus the second claim in (5.17) arises.

For part (ii) of condition 3.1.1 and condition 3.1.2, details are provided in lemma

4.1 and 4.3 of Huang (1998a).

5.3 Counting process regression

The counting process regression is so broad that it serves as a general framework for survival analysis, e.g., hazard regression is a special case of counting process regression, and even more complicated event history analysis.

Let $\mathcal{T} = [0, \tau]$ for some positive τ be the time interval for the counting process $N(t)$. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space. There is an associated filtration $\{\mathcal{F}_t : t \in \mathcal{T}\}$ satisfying $\mathcal{F}_t$ is a family of right-continuous, increasing σ-algebras and $\mathcal{F}_0$ contains the P-null sets of $\mathcal{F}$. The expected count within a very short period of time dt at time t, given all the past information $\mathcal{F}_{t-} = \cup_{s<t}\mathcal{F}_s$, is characterized as

$$E[N(dt)|\mathcal{F}_{t-}] = Y(t) \exp \eta_0(X(t))dt, \tag{5.18}$$

the intensity is proportional to the product of $Y(t)$ and the hazard function at $X(t)$ (to get rid of the positivity constraint on the hazard function, the log-hazard function η_0 is modeled), where $Y(t)$ is a $\{0, 1\}$-valued, predictable process, equals 1 when $N(t)$ is observed and 0 otherwise, and $X(t)$ is an $\mathcal{U}$-valued, predicatable stochastic process. As a special case, in the context of hazard regression, τ corresponds the so-called censoring time, $Y(t)$ is constant process that equals one or zero corresponding to $T \leq \tau$ (uncensored) or $T > \tau$ (censored), where T denotes the survival time. $N(t) = I(T \leq \tau \wedge t)$ is the counting process with a single jump at the survival time T if uncensored. Let η_0 is again the log-hazard function. Then (5.18) becomes the hazard regression model for the right censored data. Rigorously speaking, we are considering a "marker dependent hazard model", see Nielsen and Linton (1995), where the log-hazard function $\eta_0(t, X(t))$ depends only on the marker process, that is, $\eta_0(t, X(t)) = \eta_0(X(t))$.

To estimate the log-hazard function η_0, the random realizations $\mathbf{W}_i = (N_i(t), Y_i(t), X_i(t)), 1 \leq i \leq n$ are collected. We write the scaled log-likelihood for a candidate function h for η_0 as

$$l(h) = \frac{1}{n} \sum_i \left(\int_{\mathcal{T}} h(X_i(t)) N_i(dt) - \int_{\mathcal{T}} Y_i(t) \exp h(X_i(t)) dt \right).$$

The expected log-likelihood is thus given by

$$\Lambda(h) = E\left(\int_{\mathcal{T}} h(X(t)) N(dt) - \int_{\mathcal{T}} Y(t) \exp h(X(t)) dt \right).$$

Again for the marker dependent hazard model, the definition of the above log-likelihood agrees with the traditional log-likelihood in hazard regression based on the random sample $(T_i, X_i(t))$

$$l(h) = \frac{1}{n} \sum_i \left(h(X(T_i)) I\{T_i \leq \tau\} - \int_0^{T_i \wedge \tau} \exp h(X_i(t)) dt \right).$$

In this context, we could define the empirical inner product and squared norm for square-integrable functions h_1, h_2 by $\langle h_1, h_2 \rangle_n = E_n \int_{\mathcal{T}} Y(t) h_1(X(t)) h_2(X(t)) dt$ and $\|h_1\|_n^2 = \langle h_1, h_1 \rangle_n$. The corresponding theoretical quantities are denoted as $\langle \cdot, \cdot \rangle$ and $\| \cdot \|^2$, where $\langle h_1, h_2 \rangle = E \int_{\mathcal{T}} Y(t) h_1(X(t)) h_2(X(t))) dt$ and $\|h_1\|^2 = \langle h_1, h_1 \rangle$.

The counting process regression falls into the extended linear modeling framework with $W = (N(t), Y(t), X(t))$. The main theorem applies when some additional model specific conditions are satisfied.

Condition 5.3.1. *The function η_0 is bounded on $\mathcal{U}$.*

Condition 5.3.2. *For $t \in \mathcal{T}$ fixed, the Radon-Nikodym derivative of the measure $P(Y(t) = 1, X(t) \in \cdot)$ is $f_{Y(t)=1,X(t)}(t, x)$ w.r.t. the Lebesgue measure on $\mathcal{U}$. It is*

required that this RN derivative $f_{Y(t)=1,X(t)}(t,x)$ *as a function of* (t,x) *is bounded from below and above uniformly in* $t \in \mathcal{T}$ *and* $x \in \mathcal{U}$.

Remark 5.3.1. *Under condition 5.3.2, the inner product defined in this section can be rewritten as the* $\langle h_1, h_2 \rangle = \int_x h_1(x)h_2(x) \int_0^\tau f_{Y(t)=1,X(t)}(t,x)dtdx$. *Therefore, it agrees with our general notation* $V(h) = \|h\|^2 = \int h^2(x)\omega_{cp}(x)dx$, *for* $\omega_{cp}(x) = \int_0^\tau f_{Y(t)=1,X(t)}(t,x)dt$.

The rest is to show that condition 3.1.1 and 3.1.2 hold.

By lemma 5.0.5,

$$\begin{aligned} &\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)|_{\alpha=0} + \lambda J(\bar{\eta}, g) \\ &= \frac{1}{2}(E_n - E)\left(\int_{\mathcal{T}} g(X(t))dN(t)\right) - [\langle \exp(\bar{\eta}), g\rangle_n - \langle \exp(\bar{\eta}), g\rangle]. \end{aligned} \tag{5.19}$$

We are about to show that

$$\sup_{g\in\mathbb{G}}\left|\frac{\langle \exp(\bar{\eta}), g\rangle_n - \langle \exp(\bar{\eta}), g\rangle}{(V+\lambda J)(g)^{\frac{1}{2}}}\right| = O_p\left(\min\left\{\left(\frac{1}{n\lambda^{1/2q}}\right)^{1/2}, \left(\frac{1}{n\delta}\right)^{1/2}\right\}\right). \tag{5.20}$$

and

$$\sup_{g\in\mathbb{G}}\left|\frac{(E_n - E)\int_{\mathcal{T}} g(X(t))dN(t)}{(V+\lambda J)g^{1/2}}\right| = O_p\left(\min\left\{\left(\frac{1}{n\lambda^{1/2q}}\right)^{1/2}, \left(\frac{1}{n\delta}\right)^{1/2}\right\}\right). \tag{5.21}$$

Both (5.20) and (5.21) are direct applications of lemma 5.0.6. It remains to demonstrate the conditions in lemma 5.0.6 hold.

Noticing that

$$
\begin{aligned}
& E\left(\int_{\mathcal{T}} Y(t)\exp(\bar{\eta}(X(t)))\phi_\nu(X(t))dt\right)^2 \qquad (5.22)\\
& \leq E\left(\int_{\mathcal{T}} Y(t)\exp(\bar{\eta}(X(t)))^2 dt\right) E\left(\int_{\mathcal{T}} Y(t)\phi_\nu(X(t))^2 dt\right)\\
& = \int \exp(\bar{\eta}(x))^2 \omega_{cp}(x)dx \int \phi_\nu(x)^2 \omega_{cp}(x)dx\\
& \leq K_1.
\end{aligned}
$$

The last inequality follows from (4.10) and condition 5.3.2.

The same argument can be used to see that

$$
E\left(\int_{\mathcal{T}} Y(t)\exp(\bar{\eta}(X(t)))b_{k,p}(X(t))dt\right)^2 \leq K_2\delta. \qquad (5.23)
$$

Therefore, we are able to apply lemma 5.0.6 to obtain that (5.20) holds.

On the other hand, we could construct a square integrable martingale from the counting process $N(\cdot)$ as follows

$$
M(\cdot) = N(\cdot) - \int_0^{\cdot} E(N(dt|\mathcal{F}_{t-})),
$$

with its predictable variation process $\langle M\rangle = \int_0^{\cdot} E(N(dt|\mathcal{F}_{t-})$. From stochastic integral theory, the process

$$
\left(\int_0^{\cdot} g(X(t))dM(t)\right)^2 - \int_0^{\cdot} g^2(X(t))N(dt|\mathcal{F}_{t-})
$$

is also a martingale.

Now we are ready to show (5.21) with the help of lemma 5.0.6. Thus it boils

down to show

$$E\left(\int_{\mathcal{T}} \phi_\nu(X(t))dN(t)\right)^2 \leq K_1, \text{ for some } K_1 > 0 \tag{5.24}$$

and

$$E\left(\int_{\mathcal{T}} b_{k,p}(X(t))(X(t))dN(t)\right)^2 \leq K_2\delta, \text{ for some } K_2 > 0. \tag{5.25}$$

$$\begin{aligned} &E\left(\int_{\mathcal{T}} \phi_\nu(X(t))dN(t)\right)^2 \\ &\leq E\left(\int_{\mathcal{T}} \phi_\nu(X(t))dM(t)\right)^2 + E\left(\int_{\mathcal{T}} \phi_\nu(X(t))E(N(dt)|\mathcal{F}_{t-})\right)^2. \end{aligned} \tag{5.26}$$

From the theory of martingale and Markov inequality, the first and second term on the right hand side of (5.26) is equal and bounded by the following quantity

$$E\int_{\mathcal{T}} \phi_\nu^2(X(t))E(N(dt)|\mathcal{F}_{t-}) = E\int_{\mathcal{T}} \phi_\nu^2(X(t))Y(t)\exp\eta_0(X(t))dt. \tag{5.27}$$

Under Condition 5.3.1 and 5.3.2

$$\begin{aligned} &E\int_{\mathcal{T}} \phi_\nu^2(X(t))Y(t)\exp\eta_0(X(t))dt \\ &\leq CE\int_{\mathcal{T}} \phi_\nu^2(X(t))Y(t)dt \\ &= CV(\phi_\nu) \doteq K_1. \end{aligned} \tag{5.28}$$

Combining (5.26), (5.27) and (5.28), we have proved (5.24). The same argument can be used to see that (5.25) is true.

With (5.20) and (5.21) verified, we conclude that condition 3.1.1 (i) holds for

counting process.

Proof of condition 3.1.1 (ii) and condition 3.1.2 can be found in Huang (2001).

5.4 Probability density estimation

Let $X_i, i = 1, \ldots, n$ be *i.i.d.* random samples being drawn from a probability density $f_0(x)$ on a bounded interval $\mathcal{U}$. There are two intrinsic constraints that a f_0 must satisfy, the positivity constraint that $f_0 \geq 0$ and the unity constraint that $\int_{\mathcal{U}} f_0 dx = 1$. Our interest lies in estimating f_0 based on x_i's. Assuming $f_0 > 0$ on $\mathcal{U}$, make a logistic density transform to f_0 in which $f_0(\cdot) = \exp \eta_0(\cdot) / \int_{\mathcal{U}} \exp \eta_0(x) dx$ and the quantity to estimate is η_0 instead, which itself is free of the two constraints on f_0. Due to the identifiability issue, to make the estimator unique, it is convenient to consider a subspace of the spline space $\mathbb{G}$ under the orthogonality relative to the L_2 inner product.

Specifically, let this subspace of the spline space be $\mathbb{G}_1$, to make a one to one map $f_0(\cdot) = \exp \eta_0(\cdot) / \int_{\mathcal{U}} \exp \eta_0(x) dx$ between $\mathbb{G}_1$ and the set of all density functions, take $\mathbb{G}_1$ to be those splines whose integral equals zero, $\mathbb{G}_1 = \{g \in \mathbb{G} : \int_{\mathcal{U}} g(x) dx = 0\}$. Conditioning on $\mathbb{G}_1$ the identifiability of η is ensured. Denote the L_2 inner product and norm by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, $\mathbb{G}_0$ be the space consisting of all constant functions on $\mathcal{U}$, it follows that $\mathbb{G}_1 = \{g \in \mathbb{G} : \langle g, g_c \rangle = 0, \text{ for all } g_c \in \mathbb{G}_0\}$.

The scaled log-likelihood at a candidate function h based on x_i's is written as

$$l(h) = \frac{1}{n} \sum_i \left(h(x_i) - \log \int_{\mathcal{U}} \exp h(x) dx \right).$$

The expected log-likelihood therefore is given by

$$\Lambda(h) = E(h(X)) - \log \int_{\mathcal{U}} \exp h(x) dx.$$

The density model just formulated is a special case of extended linear model with $\mathbf{W} = X$. It turns out that we can apply the main theorem 3.2.1 to the context of probability density estimation without any higher level conditions. As a note, replacing $\mathbb{G}$ by the subspace $\mathbb{G}_1$, the lemmas in chapter 3 are valid, thus the proof of the main theorem can be extended to the penalized spline estimator in $\mathbb{G}_1$. On the other hand, it is sufficient to impose the regularity conditions on $\mathbb{G}_1$. In the following, we will still check the regularity condition 3.1.1 (i) for $\mathbb{G}$ (thus $\mathbb{G}_1$) while verify that condition 3.1.1 (ii) and condition 3.1.2 holds for $\mathbb{G}_1$.

By lemma 5.0.5,

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)\bigg|_{\alpha=0} + \lambda J(\bar{\eta}, g) = -\frac{1}{2}\sum_{i=1}^{n} g(x_i) + \frac{1}{2}E(g(X)). \tag{5.29}$$

Define $Q_n(f,h) = -\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} q(f,h;w_i) = -\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} f(x_i)h(x_i)$, then $Q(f,h) = -\frac{1}{2}E\bigg(f(X)h(X)\bigg)$.

With these definitions, (5.29) can be rewritten as

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)|_{\alpha=0} + \lambda J(\bar{\eta}, g) = Q_n(1,g) - Q(1,g). \tag{5.30}$$

Because of (5.30), condition 3.1.1 (i) is just a direct application of lemma 5.0.6. The rest is to show that

$$E(q(1,\phi_\nu;w_1)^2) < K_1 \quad \text{and} \quad E(q(1,b_{k,p};w_1)^2) < K_2\delta. \tag{5.31}$$

Meanwhile (5.31) is equivalent to

$$E(\phi_\nu(X)^2) < K_1 \quad \text{and} \quad E(b_{k,p}(X)^2) < K_2\delta, \tag{5.32}$$

which are the properties of the basis functions ϕ_ν and $b_{k,p}$ that have been introduced before. Till now we complete the proof for condition 3.1.1 (i).

Next, we claim that

$$\frac{d^2}{d\alpha^2} l(g_1 + \alpha(g_2 - g_1)) = \frac{d^2}{d\alpha^2} \Lambda(g_1 + \alpha(g_2 - g_1)) = -\mathrm{Var}(g_2(X_\alpha) - g_1(X_\alpha)), \quad (5.33)$$

where X_α has the density $f_{X_\alpha}(x) = \exp g_\alpha(x) / \int_{\mathcal{U}} \exp g_\alpha(x) dx$ and $g_\alpha = g_1 + \alpha(g_2 - g_1)$.

Under the assumption $\|g_1\|_\infty \leq K$ and $\|g_2\|_\infty \leq K$, we are able to show that

$$\mathrm{Var}(g_2(X_\alpha) - g_1(X_\alpha)) \asymp \int_{\mathcal{X}} (g_2(x) - g_1(x))^2 dx. \quad (5.34)$$

Condition 3.1.1 (ii) and condition 3.1.2 now follow from (5.33) and (5.34).

5.5 Spectral density estimation

In this section, we are going to verify condition 3.1.1 and 3.1.2 holds in the context of spectral density estimation. We want to point it out that the formulation of the maximum likelihood here has also been done elsewhere, such as Kooperberg et al. (1995). The conditions appeared here is very similar to theirs.

We will first impose an additional condition on the eigenfunctions $\{\phi_j\}_{j=1}^\infty$, which are defined under Proposition 3.1.1.

Condition 5.5.1. *The eigenfunctions are uniformly bounded, to be specific, that is* $\max_j \|\phi_j(\lambda)\|_\infty < \infty$.

We consider a stationary linear time series $\{X_t\}$ taking the form

$$X_t = \sum_{j=-\infty}^{\infty} a_j Z_{t-j}$$

where $\{Z_j\}_{j=-\infty}^\infty$ are white noise with mean zero and variance σ^2.

The theoretical autocovariance function $\gamma(\cdot)$ for this linear process $\{X_t\}$ has the expression

$$\gamma(u) = \mathrm{cov}(X_t, X_{t+u}) = \sigma^2 \sum_j a_{j-u} a_j,$$

while its spectral density function $f(\cdot)$ is given by

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left| \sum_{j=-\infty}^{\infty} a_j \exp(-ij\lambda) \right|^2 \quad -\pi \leq \lambda \leq \pi.$$

To fit into the general extended linear model framework, we need to impose the smoothness condition and positive condition on the spectral density f. In the context of linear time series $\{X_t\}$, it is more common to put the condition on the original series if possible.

Definition 5.5.1. *For any positive real number s, let l be the floor of s and α be the remaining decimal. We call a function g on $[0, \pi]$ s-smooth if g is l-times differentiable on $[0, \pi]$ and $g^{(l)}$ satisfies a Hölder condition with exponent α, that is, $|g^{(l)}(\lambda) - g^{(l)}(\lambda_0)| \leq c|\lambda - \lambda_0|^\alpha$ for $\lambda, \lambda_0 \in [0, \pi]$.*

Condition 5.5.2. *$\{X_t\}$ is a stationary linear process with $\sum_j |a_j| j^s < \infty$ for some $s > 1/2$. In addition, the white noise process has normal distribution, $Z_j \sim_{i.i.d.} N(0, \sigma^2)$.*

Condition 5.5.3. *The spectral density function f is bounded away from zero on $[0, \pi]$.*

Remark 5.5.1. *Under condition 5.5.2, the spectral density function is s-smooth, so is the logarithm $\eta = \log f$. Condition 5.5.3 indicates that η is bounded. Meanwhile, the time series $\{X_t\}$ is Gaussian given condition 5.5.2, the distribution assumption on X_t will play a role in determining the asymptotic distribution of the periodogram.*

Suppose $X_0, \ldots, X_{T-1}$ be a realization of length T of the series, it is known that the periodogram

$$I^{(T)}(\lambda) = (2\pi T)^{-1}\left|\sum_{t=0}^{T-1}\exp(-i\lambda t)X_t\right|^2 \quad -\pi \leq \lambda \leq \pi$$

has the following asymptotic properties:

$$I^{(T)}(\lambda_k) = f(\lambda_k)W_k \quad \lambda_k = \frac{2\pi k}{T} \quad \text{for} \quad k = 0, \ldots, [T/2] \tag{5.35}$$

where $W_k,\ k = 0, \ldots, [T/2]$, are the ratios of the periodogram and the spectral density function evaluated at the grid points λ_k between $[0, \pi]$. W_k, $k = 0, \ldots, T/2$ are asymptotically independent; The asymptotic distribution of W_k is free of $f(\cdot)$, which is exponential distribution with mean one when λ_k is not on the boundary of $[0, \pi]$. W_0 and $W_{[T/2]}$ (if T is even) have approximately the χ^2 distribution with degree of freedom one, and $W_0, W_1, \ldots, W_{[T/2]}$; see Brockwell and Davis (1991) and references therein.

Since the spectral density function is symmetric about zero on $[-\pi, \pi]$ and is periodic (with period 2π), it is sufficient to model the segment on $[0, \pi]$ with additional constraints that $f^{'}(0) = f^{'''}(0) = f^{'}(\pi) = f^{'''}(\pi) = 0$ and $\eta^{'}(0) = \eta^{'''}(0) = \eta^{'}(\pi) = \eta^{'''}(\pi) = 0$. Let $\mathbb{G}_1$ denote the spline space and we use the following subspace of spline (denoted as $\mathbb{G}$) as the estimation space.

$$\mathbb{G} = \{g \in \mathbb{G}_1 : g^{'}(0) = g^{'''}(0) = g^{'}(\pi) = g^{'''}(\pi) = 0\}.$$

Define

$$\psi(y, \lambda; g) = \left\{\frac{\delta_\pi(\lambda)}{2} - 1\right\}[g(\lambda) + y\exp(-g(\lambda))]$$

for $0 < \lambda \leq \pi$ and $y \geq 0$, where $\delta_\pi(\lambda) = 1$ if $\lambda = \pi$ and $\delta_\pi(\lambda) = 0$ otherwise.

Set $I_k = I^{(T)}(\lambda_k) \quad k = 1, 2, \ldots, [T/2]$, we can write the (approximate) log-likelihood function of the periodogram for a candidate function $h \in \mathbb{G}$, according to (5.35), as

$$l(h) = \frac{1}{[T/2]} \sum_{k=1}^{[T/2]} \psi(I_k, \lambda_k, h) = \frac{1}{[T/2]} \sum_{k=1}^{[T/2]} \left\{ \frac{\delta_\pi(\lambda_k)}{2} - 1 \right\} [h(\lambda_k) + I_k \exp(-h(\lambda_k))].$$

Define the (approximate) expected log-likelihood function

$$\Lambda(h) = \frac{1}{[T/2]} \sum_{k=1}^{[T/2]} \left\{ \frac{\delta_\pi(\lambda_k)}{2} - 1 \right\} [h(\lambda_k) + E(I_k) \exp(-h(\lambda_k))].$$

The context of nonparametric spectral density belongs to the extended linear model with $\mathbf{W} = (\lambda, I^{[T]}(\lambda))$. One may notice that instead of *i.i.d.* random sample W_i's, the observations W_i are independent but not identically distributed. The trouble it causes when checking condition 3.1.1 can be overcome later on in this section. Also, it can be seen in Theorem 8.12 of Schumaker (1981) that Lemma 3.3.2 holds when the space is changed from $\mathbb{G}_1$ to $\mathbb{G}$. Therefore, the proof of convergence rate of $\|\hat{\eta} - \eta_0\|$ for spectral density model is exactly the same as that for extended linear model.

The rest is to show that Condition 3.1.1 and 3.1.2 holds.

We first check that part (i) of Condition 3.1.1 holds.

By lemma 5.0.5,

$$-\frac{1}{2} \frac{d}{d\alpha} l(\bar{\eta} + \alpha g) \Big|_{\alpha=0} + \lambda J(\bar{\eta}, g) = \frac{1}{2[T/2]} \sum_k \left\{ \frac{\delta_\pi(\lambda_k)}{2} - 1 \right\} g(\lambda_k) [I_k - E(I_k)] \exp(-\bar{\eta}(\lambda_k)). \tag{5.36}$$

Define $Q_n(f, h) = \frac{1}{2[T/2]} \sum_k \left\{ \frac{\delta_\pi(\lambda_k)}{2} - 1 \right\} h(\lambda_k) I_k \exp(-f(\lambda_k))$, then $Q(f, h) =$

$\frac{1}{2[T/2]}\sum_k\left\{\frac{\delta_\pi(\lambda_k)}{2}-1\right\}h(\lambda_k)E(I_k)\exp(-f(\lambda_k))$.

With these definitions, (5.36) can be rewritten as

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)|_{\alpha=0}+\lambda J(\bar{\eta},g)=Q_n(\bar{\eta},g)-Q(\bar{\eta},g). \tag{5.37}$$

Lemma 5.0.6 can not be directly applied here since the summands in $Q_n(\bar{\eta},g)$ are not identically distributed. However, we can modify the proof of lemma 5.0.6 to accommodate the situation here. Overthere instead of bounding $E(Q_n(h_n,\phi_\nu)-Q(h_n,\phi_\nu))^2$ by $\frac{1}{n}E(q(h_n,\phi_\nu,w_1)^2)$, we can directly find a bound for this term.

$$\begin{aligned}&E(Q_n(\bar{\eta},\phi_\nu)-Q(\bar{\eta},\phi_\nu))^2=\mathrm{Var}(Q_n(\bar{\eta},\phi_\nu)) \\ &=\mathrm{Var}\left(\frac{1}{2[T/2]}\sum_k\left\{\frac{\delta_\pi(\lambda_k)}{2}-1\right\}\phi_\nu(\lambda_k)[I_k-E(I_k)]\exp(-\bar{\eta}(\lambda_k))\right).\end{aligned} \tag{5.38}$$

By applying Theorem 10.3.2 (ii) of Brockwell and Davis (1991) together with Condition 5.5.1 and the boundedness of $\|\bar{\eta}\|_\infty$ given in (4.5),

$$E(Q_n(\bar{\eta},\phi_\nu)-Q(\bar{\eta},\phi_\nu))^2=O\left(\frac{1}{T}\right). \tag{5.39}$$

Plug in (5.39) to the proof of Lemma 5.0.6, we could obtain the first result in lemma 5.0.6

$$\sup_{g\in\mathbb{G}}\left|\frac{Q_n(\bar{\eta},g)-Q(\bar{\eta},g)}{(V+\lambda J)(g)^{\frac{1}{2}}}\right|=O_p\left(\left(\frac{1}{T\lambda^{1/2q}}\right)^{1/2}\right). \tag{5.40}$$

Similarly, we are able to derive the second result in lemma 5.0.6

$$\sup_{g\in\mathbb{G}}\left|\frac{Q_n(\bar{\eta},g)-Q(\bar{\eta},g)}{(V+\lambda J)(g)^{\frac{1}{2}}}\right|=O_p\left(\left(\frac{1}{T\delta}\right)^{1/2}\right). \tag{5.41}$$

Therefore, Condition 3.1.1 (i) follows from (5.38), (5.40) and (5.41).

Next, we want to check that condition 3.1.1(ii) holds.

Define $g_\alpha = g_1 + \alpha(g_2 - g_1)$, taking second derivative of $l(g_\alpha)$ with respect to α yields

$$\frac{d^2}{d\alpha^2} l(g_\alpha) = \frac{1}{[T/2]} \sum_{k=1}^{[T/2]} \left\{ \frac{\delta_\pi(\lambda_k)}{2} - 1 \right\} [I_k(g_2(\lambda_k) - g_1(\lambda_k))^2 \exp(-g_\alpha(\lambda_k))]. \quad (5.42)$$

To see that Condition 3.1.1 (ii) holds. It is sufficient to check that except on an event whose probability tends to zero when $T \longrightarrow \infty$,

$$\frac{1}{[T/2]} \sum_{k=1}^{[T/2]} I_k(g_2(\lambda_k) - g_1(\lambda_k))^2 \geq M \|g_2 - g_1\|^2. \quad (5.43)$$

Similarly, Condition 3.1.2 is equivalent to

$$M_2 \|g_2 - g_1\|^2 \leq \frac{1}{[T/2]} \sum_{k=1}^{[T/2]} E(I_k)(g_2(\lambda_k) - g_1(\lambda_k))^2 \leq M_1 \|g_2 - g_1\|^2. \quad (5.44)$$

It turns out that we do not bother to check (5.43) once (5.44) is satisfied. The reason is due to the next lemma.

Lemma 5.5.1.

$$\sup_{g \in \mathbb{G}} \frac{|\frac{1}{[T/2]} \sum_{k=1}^{[T/2]} (I_k - E(I_k)) g(\lambda_k)^2|}{\|g\|^2} = O_p\left(\frac{1}{\sqrt{T}}\right).$$

Proof. We first introduced some notations. Let $A_i = [x_i, x_{i+1}], 0 \leq i \leq K-1$. Denote $\mathbb{G}_i = \{g|_{A_i}, g \in \mathbb{G}\}$ and $g_i = g|_{A_i}$.

If $\frac{|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))g_i(\lambda_k)^2|}{\|g_i\|^2} \le t$ for all i and $g\in\mathbb{G}$, then

$$\begin{aligned}\left|\frac{1}{[T/2]}\sum_{k=1}^{[T/2]}(I_k-E(I_k))g(\lambda_k)^2\right| &\le \sum_{i=0}^{K-1}\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))g_i(\lambda_k)^2\right| \\ &\le \sum_{i=0}^{K-1} t\|g_i\|^2 = t\|g\|^2.\end{aligned}$$

Thus

$$\sup_{g\in\mathbb{G}}\frac{\left|\frac{1}{[T/2]}\sum_{k=1}^{[T/2]}(I_k-E(I_k))g(\lambda_k)^2\right|}{\|g\|^2} \le \sup_i\sup_{g\in\mathbb{G}}\frac{\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))g_i(\lambda_k)^2\right|}{\|g_i\|^2}.$$

It follows that for any $t>0$

$$\begin{aligned}&P\left(\sup_{g\in\mathbb{G}}\left|\frac{1}{[T/2]}\sum_{k=1}^{[T/2]}(I_k-E(I_k))g(\lambda_k)^2\right| > t\|g\|^2\right) \\ &\le P\left(\sup_i\sup_{g\in\mathbb{G}}\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))g_i(\lambda_k)^2\right| > t\|g_i\|^2\right) \\ &\le \sum_i P\left(\sup_{g\in\mathbb{G}}\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))g_i(\lambda_k)^2\right| > t\|g_i\|^2\right).\end{aligned}$$

Let $\phi_j, j=1,\dots,p+1$ be an orthonormal basis of $\mathbb{G}_i$ relative to $\langle\cdot,\cdot\rangle$, $g_i = \sum_j \beta_j\phi_j$.

If

$$\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))\phi_j(\lambda_k)\phi_{j'}(\lambda_k)\right| \le \frac{t}{p+1}, \quad \text{for} \quad j,j'=1,\dots,p+1,$$

then

$$\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))g_i(\lambda_k)^2\right| \le \sum_{j,j'}|\beta_j||\beta_{j'}|\frac{t}{p+1} \le t\|g_i\|^2.$$

Consequently,

$$P\left(\sup_{g\in\mathbb{G}}\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))g_i(\lambda_k)^2\right|>t\|g_i\|^2\right)$$
$$\leq\sum_i(p+1)^2P\left(\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))\phi_j(\lambda_k)\phi_{j'}(\lambda_k)\right|>\frac{t}{p+1}\right).$$

Due to condition 5.5.1, $M_\phi=sup_j\|\phi_j\|_\infty<\infty$, hence

$$P\left(\sup_{g\in\mathbb{G}}\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))\phi_j(\lambda_k)\phi_{j'}(\lambda_k)\right|>\frac{t}{p+1}\right)$$
$$\leq P\left(\sup_{g\in\mathbb{G}}\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))\right|>\frac{t}{(p+1)M_\phi}\right).$$

Applying Cauchy-Schwarz inequality to the right hand side of the last inequality yields

$$P\left(\sup_{g\in\mathbb{G}}\left|\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))\right|>\frac{t}{p+1}\right)\leq\frac{(p+1)^2M_\phi^2E[(\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k)))^2]}{t^2}.$$

By Theorem 10.3.2 (ii) of Brockwell and Davis (1991),

$$E\left[\left(\frac{1}{[T/2]}\sum_{\lambda_k\in A_i}(I_k-E(I_k))\right)^2\right]=O\left(\frac{N_i}{T^2}\right),$$

where $N_i=\#\{\lambda_k,\lambda_k\in A_i\}$.

Combining all the above results, it can be obtained that for any t

$$P\left(\sup_{g\in\mathbb{G}}\left|\frac{1}{[T/2]}\sum_{k=1}^{[T/2]}(I_k-E(I_k))g(\lambda_k)^2\right|>t\|g\|^2\right)\leq O\left(\frac{1}{Tt^2}\right).$$

We finish the proof of the lemma. □

From Lemma 5.5.1, both Condition 3.1.1 (ii) and 3.1.2 reduces to show (5.44), which can be simplified to the argument that

$$M_1\|g\|^2 \leq \frac{1}{[T/2]}\sum_{k=1}^{[T/2]} g(\lambda_k)^2 \leq M_2\|g\|^2 \tag{5.45}$$

holds except on an event whose probability tends to zero as $T \longrightarrow 0$ for some $M_1 > 0$ and $M_2 > 0$.

Finally, (5.45) can be easily justified using the definition of integral and Riemann sum.

5.6 Diffusion process

Diffusion type processes are widely used to describe continuous time stochastic processes with application to physical, biological, medical, economic, and social sciences. See Rao (1999) for a study on the development of statistical estimation and inference in the field of diffusion processes.

Here, as in Stone and Huang (2003), the focus is on the nonparametric estimation of the drift coefficient as a function of some time dependent covariates while assuming the diffusion coefficient as a function of time is known. The estimator is constructed given continuous realizations of the relevant processes.

To be specific, we will define a one-dimensional diffusion type process $\{Y(t), 0 \leq t \leq \tau\}$ accompanied by a covariate process $\{X(t), 0 \leq t \leq \tau\}$

$$dY(t) = \eta_0(X(t))dt + \sigma(t)dW(t), \quad 0 \leq t \leq \tau,$$

where $0 < \tau < \infty$ and $W(t)$ is a Wiener process. The diffusion coefficient $\sigma(t)$ is a known function of time, while the drift coefficient $\eta_0(X(t))$ is an unknown function of a covariate process $X(t)$. Moreover, let $Z(t)$ be a $\{0, 1\}$ valued process as a censoring

indicator, $Z(t) = 1$ indicating the processes $X(t)$ and $Y(t)$ are observed , $Z(t) = 0$ otherwise.

The estimation of η_0 will be based on a random sample of n realizations of $\{(X_i(t), Y_i(t), Z_i(t)) : 0 \leq t \leq \tau\}, 1 \leq i \leq n$. The scaled (partial) log-likelihood at a candidate function h can be expressed as

$$l(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\int Z_i(t)\frac{h(X_i(t))}{\sigma^2(t)}dY_i(t) - \frac{1}{2}\int Z_i(t)\frac{h^2(X_i(t))}{\sigma^2(t)}dt\right).$$

The expected (partial) log-likelihood is given by conditioning $Y(t)$ on $X(t)$, then take expectation on $X(t)$,

$$\begin{aligned}\Lambda(h) &= E\left(\int Z(t)\frac{h(X(t))}{\sigma^2(t)}dY(t) - \frac{1}{2}\int Z(t)\frac{h^2(X(t))}{\sigma^2(t)}dt\right)\\ &= E\left(\int Z(t)\frac{h(X(t))\eta_0(X(t))}{\sigma^2(t)}dt - \frac{1}{2}\int Z(t)\frac{h^2(X(t))}{\sigma^2(t)}dt\right)\end{aligned}$$

With $l(h)$ and $\Lambda(h)$ being defined above, the diffusion type of process fits the extended linear model with $\mathbf{W} = (X(t), Y(t), Z(t))$. Moreover, theorem 3.2.1 holds for the above diffusion process under two conditions.

Condition 5.6.1. *There are two positive constants $M_2 \geq M_1$ such that $M_1 \leq \sigma^{-2}(t) \leq M_2$ whenever $Z(t) = 1$.*

Condition 5.6.2. *There are constants $M_4 \geq M_3 > 0$ such that*

$$M_3\psi(A) \leq E\left(\int Z(t)ind(X(t) \in A)\right) \leq M_4\psi(A)$$

for all Borel subset A of $\mathcal{X}$. Moreover, $\psi(A)$ denotes the Lebesgue measure of A.

All we need is to verify condition 3.1.1 and condition 3.1.2 hold.

By lemma 5.0.5

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)\Big|_{\alpha=0}+\lambda J(\bar{\eta},g)=-\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\int Z_i(t)\frac{g(X_i(t))}{\sigma^2(t)}[dY_i(t)-\bar{\eta}(X_i(t))dt] \tag{5.46}$$
$$+\frac{1}{2}E\bigg(\int Z(t)\frac{g(X(t))}{\sigma^2(t)}[\eta_0(X(t))-\bar{\eta}(X(t))dt]\bigg)$$

Define $Q_n(f,h)=-\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\int Z_i(t)\frac{h(X_i(t))}{\sigma^2(t)}[dY_i(t)-f(X_i(t))dt]$, then it immediately follows that $Q(f,h)=-\frac{1}{2}E\bigg(\int Z_i(t)\frac{h(X_i(t))}{\sigma^2(t)}[dY_i(t)-f(X_i(t))dt]\bigg)$.

With these definitions, (5.46) can be rewritten as

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)|_{\alpha=0}+\lambda J(\bar{\eta},g)=Q_n(\bar{\eta},g)-Q(\bar{\eta},g). \tag{5.47}$$

Because of (5.47), condition 3.1.1 (i) is just a direct application of lemma 5.0.6. The rest is to show that

$$E\bigg(\bigg[\int Z(t)\frac{\phi_\nu(X(t))}{\sigma^2(t)}[dY(t)-\bar{\eta}(X(t))dt]\bigg]^2\bigg)<K_1 \tag{5.48}$$

and

$$E\bigg(\bigg[\int Z(t)\frac{b_{k,p}(X(t))}{\sigma^2(t)}[dY(t)-\bar{\eta}(X(t))dt]\bigg]^2\bigg)<K_2\delta. \tag{5.49}$$

For (5.48), it suffices to verify that

$$E\bigg(\bigg[\int Z(t)\frac{\phi_\nu(X(t))}{\sigma^2(t)}[dY(t)-\eta_0(X(t))dt]\bigg]^2\bigg)\leq M \tag{5.50}$$

and

$$E\bigg(\bigg[\int Z(t)\frac{\phi_\nu(X(t))}{\sigma^2(t)}[\eta_0(X(t))-\bar{\eta}(X(t))]dt\bigg]^2\bigg)\leq M. \tag{5.51}$$

Using the fact that $E(\int f(t,X(t))dW(t))^2=E(\int f^2(t,X(t))dt)$, we can express

(5.50) as

$$E\left(\left[\int Z(t)\frac{\phi_\nu(X(t))}{\sigma^2(t)}[dY(t)-\eta_0(X(t))dt]\right]^2\right)=E\left(\int Z(t)\frac{\phi_\nu^2(X(t))}{\sigma^2(t)}dt\right). \quad (5.52)$$

Under Condition 5.6.1 and 5.6.2, we can bound (5.52) by

$$M_1\cdot M_3\le E\left(\int Z(t)\frac{\phi_\nu^2(X(t))}{\sigma^2(t)}dt\right)\le M_2\cdot M_4. \quad (5.53)$$

Thus, (5.50) is proved to be true.

On the other hand, by applying the fact that $(\int g(t)dt)^2\le\int g^2(t)dt$ and $\|\eta_0\|_\infty\le\infty$ and boundedness of $\|\bar{\eta}\|_\infty$ (4.5), (5.51) is upper bounded by

$$E\left(\left[\int Z(t)\frac{\phi_\nu(X(t))}{\sigma^2(t)}[\eta_0(X(t))-\bar{\eta}(X(t))]dt\right]^2\right)\le\tau E\left(\int Z(t)\frac{\phi_\nu^2(X(t))}{\sigma^4(t)}dt\right). \quad (5.54)$$

Similarly as in (5.53), the right hand side of (5.54) is bounded by

$$E\left(\int Z(t)\frac{\phi_\nu^2(X(t))}{\sigma^4(t)}dt\right)\le M_2^2*M_4. \quad (5.55)$$

Combining (5.54) and (5.55), we finished the proof of (5.51).

In summary, (5.48) is true. The same argument remains true for (5.49). Hence we complete the proof of condition 3.1.1 (i).

We can refer to Stone and Huang (2003) for the verification of condition 3.1.1 (ii) and condition 3.1.2.

5.7 Nonparametric M-regression

The theoretical results in our extended linear modeling can be modified to handle contexts in which the log-likelihood function may not be twice differentiable (the regularity conditions are no longer feasible), e.g., nonparametric M-regression, that

includes but may not be limited to, least absolute deviations (LAD) regression or more generally, quantile regression and Huber's robust regression. The current set up finds its root in Stone (2005) and the references therein.

The idea of Nonparametric M-regression is to find some other loss functions than the squared loss used in the ordinary regression. The (negative) loss function could be specified through a concave function Ψ on $\mathbb{R}$. One feature of Ψ is that there exists some $a_1 \leq a_2$, Ψ is linear with positive slope when its argument is smaller than a_1 while takes the form of another linear function with negative slope when its argument is to the right of a_2. Furthermore, let ψ be the average of the left- and right-hand derivatives of Ψ. By the above property of Ψ, $\psi(x)$ is nonincreasing and that it equals $\psi(-\infty) > 0$ when $x \leq a_1$ and equals $\psi(\infty) < 0$ when $x \geq a_2$. Consequently, $\psi(-\infty) \geq \psi(y) \geq \psi(\infty)$ for $y \in \mathbb{R}$ and $\|\psi\|_\infty = \max(\psi(-\infty), -\psi(\infty))$.

Though ψ can only be thought of as the psuedo derivative of Ψ, the Newton-Leibniz formula still holds, meaning

$$\Psi(y_2) - \Psi(y_1) = \int_{y_1}^{y_2} \psi(y)dy, \quad y_1, y_2 \in \mathbb{R}$$

and hence that

$$|\Psi(y - \theta) - \Psi(y)| \leq |\theta| \|\psi\|_\infty \quad \theta, y \in \mathbb{R}.$$

As in the regression setup, consider a pair of random variables (X, Y) with X being an $\mathcal{U}$-valued and Y being real-valued. Suppose the joint density function of (X, Y) is positive on $\mathcal{U} \times \mathbb{R}$. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be independent observations from the same distribution as (X, Y). Then the scaled (normalized) log-likelihood function is formulated by replacing the negative squared loss with Ψ

(the shift term does not affect the estimation of η_0, is added just for later derivation),

$$l(h) = \frac{1}{n}\sum_{i=1}^{n}[\Psi(Y_i - h(X_i)) - \Psi(Y_i)].$$

Hence the expected log-likelihood is simply written as

$$\Lambda(h) = E[\Psi(Y - h(X)) - \Psi(Y)].$$

For better understanding, we will pause for some concrete examples in the context of nonparametric M-estimation. Consider the choice of $\Psi(y) = -(1-p)y_- - py_+$, where $y_- = \max(-y, 0)$ while $y_+ = \max(y, 0)$, this type of Ψ will lead to a conditional p-th quantile function estimator. The corresponding ψ-function is $\psi(y) = 1 - p$ for $|y| < 0$, $\psi(0) = 1/2 - p$ and $\psi(y) = -p$ for $|y| > 0$. Another famous example was introduced by Huber et al. (1964) with the purpose of robust estimation, there they proposed $\Psi(y) = -y^2/2$ for $|y| \leq c$ and $\Psi(y) = -c|y| + c^2/2$ for $|y| > c$, where c is a positive constant. The corresponding psuedo derivative function ψ can be easily obtained also, $\psi(y) = -y$ for $|y| \leq c$ and $\psi(y) = -c\text{sign}(y)$ for $|y| > c$.

As pointed out in Stone (2005), the function $-\psi$ could define a finite measure on $\mathbb{R}$ which assigns measure in the following way: For $a < b$, the interval $(a, b]$ has measure $\psi(a+)-\psi(b+)$. In particular, the whole real line has a finite positive measure $\psi(-\infty) - \psi(\infty)$. As the general measure theory shows, the newly defined measure can be decomposed to be the summation of an absolutely continuous component, a discrete component and a singular component.

It is denoted later in this section that (X, Y) has a joint density $f(x, y)$ and a marginal density $f_X(x)$ for X, the former is bounded on $\mathcal{U} \times \mathbb{R}$ while the latter is bounded on $\mathcal{U}$.

Using the joint density $f(x,y)$, the expected log-likelihood has an expression

$$\Lambda(h) = \int_{\mathcal{X}} \left(\int_{\mathbb{R}} [\Psi(y - h(x)) - \Psi(y)] f(x,y) dy \right) dx.$$

Let g_1 and g_2 be bounded functions on $\mathcal{X}$. Then

$$\frac{d}{d\alpha}\Lambda(g_1 + \alpha g_2) = \int_{\mathcal{X}} g_2(x) \left(\int_{\mathbb{R}} \psi(y - g_1(x) - \alpha g_2(x)) f(x,y) dy \right) dx.$$

Moreover, by Fubini's theorem,

$$\frac{d^2}{d\alpha^2}\Lambda(g_1 + \alpha g_2) = -\int_{\mathcal{X}} g_2^2(x) \left(\int_{\mathbb{R}} f(x,y) d\psi(y - g_1(x) - \alpha g_2(x)) \right) dx \tag{5.56}$$

in the sense that

$$\int_{\alpha_1}^{\alpha_2} \frac{d^2}{d\alpha^2}\Lambda(g_1 + \alpha g_2) d\alpha = \frac{d}{d\alpha}\Lambda(g_1 + \alpha g_2)\Big|_{\alpha_2} - \frac{d}{d\alpha}\Lambda(g_1 + \alpha g_2)\Big|_{\alpha_1}.$$

It follows from (5.56) that there are positive numbers M_3 and M_4 such that

$$-M_3 \|g_2\|^2 \leq \frac{d^2}{d\alpha^2}\Lambda(g_1 + \alpha g_2) \leq -M_4 \|g_2\|^2.$$

Thus condition 3.1.2 holds for the nonparametric M-regression model.

Now let's verify that condition 3.1.1 (i) holds for this model as well.

According to lemma 5.0.5 and the definition of $l(h)$ and $\Lambda(h)$

$$-\frac{1}{2}\frac{d}{d\alpha} l(\bar{\eta} + \alpha g)\Big|_{\alpha=0} + \lambda J(\bar{\eta}, g) = -\frac{1}{2}(E_n - E)[g(X)\psi(Y - \bar{\eta}(X))]. \tag{5.57}$$

Define $Q_n(f,g) = -\frac{1}{2}E_n[g(X)\psi(Y - \bar{\eta}(X))]$, then $Q(f,g) = -\frac{1}{2}E[g(X)\psi(Y - \bar{\eta}(X))]$.

With these definitions, (5.57) can be rewritten as

$$-\frac{1}{2}\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)\Big|_{\alpha=0}+\lambda J(\bar{\eta},g)=Q_n(\bar{\eta},g)-Q(\bar{\eta},g). \tag{5.58}$$

The rest is to show that

$$E[\phi_\nu^2(X)\psi^2(Y-\bar{\eta}(X))]<K_1 \tag{5.59}$$

and

$$E[b_{k,p}^2(X)\psi^2(Y-\bar{\eta}(X))]<K_2\delta. \tag{5.60}$$

(5.59) ((5.60)) is easily obtained by the boundedness of ψ and the fact that $E[\phi_\nu^2(X)]=1$ ($E[b_{k,p}^2(X)]\leq C\delta$ for some C not depending on k).

Since $l(h)$ is not twice differentiable, in the light of Stone (2005), we need the following condition instead of 3.1.1 (ii), which only involves the first derivative

Condition 5.7.1. *There is a constant $M>0$ such that, with probability tending to one as $n\longrightarrow\infty$, we have*

$$\sup_{g\in\mathbb{G}}\left(\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)\Big|_{\alpha=1}-\frac{d}{d\alpha}l(\bar{\eta}+\alpha g)\Big|_{\alpha=0}\right)\leq -M\|g\|^2.$$

With condition 3.1.1 (ii) replaced by condition 5.7.1 and other conditions remained, we are still able to prove that theorem 3.2.1 holds for the nonparametric M-regression model.

The approximation error part remains the same as its correspondence in chapter 4. Some adjustments are needed to obtain the estimation error.

Suppose condition 5.7.1 is true, it is easy to derive that, for n sufficiently large

and any $1 > \epsilon > 0$,

$$\sup_{g\in\mathbb{G}}\left(\frac{d}{d\alpha}[-l(\bar{\eta}+\alpha g)+\lambda J(\bar{\eta}+\alpha g)]\Big|_{\alpha=1} - \frac{d}{d\alpha}[-l(\bar{\eta}+\alpha g)+\lambda J(\bar{\eta}+\alpha g)]\Big|_{\alpha=0}\right) \tag{5.61}$$

$$\geq \min\{M,2\}(V+\lambda J)(g)$$

holds except on an event whose probability is less than ϵ.

On the other hand, it is follows from condition 3.1.1 (i) that for all $g \in \mathbb{G}$ satisfying $(V+\lambda J)(g) = a^2 \min\left\{\frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta}\right\}$

$$\begin{aligned} &\left|\frac{d}{d\alpha}\Big[-l(\bar{\eta}+\alpha g)+\lambda J(\bar{\eta}+\alpha g)\Big]\Big|_{\alpha=0}\right| \\ &\leq 2aO_p\left(\min\left\{\frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta}\right\}\right). \end{aligned} \tag{5.62}$$

Thus for g satisfying $(V+\lambda J)(g) = a^2 \min\left\{\frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta}\right\}$ we can choose a sufficiently large such that for n sufficiently large, except on an event whose probability is less than ϵ,

$$\frac{d}{d\alpha}\Big[-l(\bar{\eta}+\alpha g)+\lambda J(\bar{\eta}+\alpha g)\Big]\Big|_{\alpha=1} > 0. \tag{5.63}$$

Hence

$$-l(\bar{\eta}+\alpha g)+\lambda J(\bar{\eta}+\alpha g) > -l(\bar{\eta}+g)+\lambda J(\bar{\eta}+g) \quad \alpha > 1$$

for all g with $(V+\lambda J)(g) = a^2 \min\left\{\frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta}\right\}$.

Consequently, for n sufficiently large, except on an event having probability less than 2ϵ, $(V+\lambda J)(\hat{\eta}-\bar{\eta}) \leq a^2 \min\left\{\frac{1}{n\lambda^{1/2q}}, \frac{1}{n\delta}\right\}$.

Finally for a detailed proof of validity of condition 5.7.1, refer to section 5 of

Stone (2005).

6. CONCLUSION AND FUTURE WORK

In this dissertation, the convergence rate of penalized spline is brought into attention. Compared to the existing work, which deals with one specific model or some selected asymptotic scenarios, the main theorem here is very general, it holds for the set of extended linear models under various asymptotic scenarios. The next step might be to analyze the asymptotic distribution of the penalized spline estimator. There are some related results, see chapter 2. However, it is worthy to explore a general approach.

REFERENCES

DL Barrow and PW Smith. Asymptotic properties of best l2 [0, 1] approximation by splines with variable knots. *Quart. Appl. Math*, 36(3):293–304, 1978.

Peter J. Brockwell and Richard Davis. Time series: Theory and methods. *New York-Berlin*, 1991.

Gerda Claeskens, Tatyana Krivobokova, and Jean D Opsomer. Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544, 2009.

Ronald A DeVore and George G Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer, 2013.

Peter Hall and Jean D Opsomer. Theory for penalised spline regression. *Biometrika*, 92(1):105–118, 2005.

Mark Henry Hansen. *Extended linear models, multivariate splines, and ANOVA*. University of California, Berkeley, 1994.

Ashley D Holland. *Penalized Spline Estimation in the Partially Linear Model*. PhD thesis, The University of Michigan, 2012.

Jianhua Z Huang. Functional anova models for generalized regression. *Journal of Multivariate Analysis*, 67(1):49–71, 1998a.

Jianhua Z Huang. Projection estimation in multiple regression with application to functional anova models. *The Annals of Statistics*, 26(1):242–272, 1998b.

Jianhua Z Huang. Concave extended linear modeling: a theoretical synthesis. *Statistica Sinica*, 11(1):173–198, 2001.

Jianhua Z Huang. Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635, 2003.

Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Göran Kauermann, Tatyana Krivobokova, and Ludwig Fahrmeir. Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):487–503, 2009.

Charles Kooperberg, Charles J Stone, and Young K Truong. Rate of convergence for logspline spectral density estimation. *Journal of Time Series Analysis*, 16(4): 389–401, 1995.

Yingxing Li and David Ruppert. On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436, 2008.

Jens P Nielsen and Oliver B Linton. Kernel estimation in a nonparametric marker dependent hazard model. *The Annals of Statistics*, pages 1735–1748, 1995.

Finbarr O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical science*, pages 502–518, 1986.

Finbarr O'Sullivan. Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on scientific and statistical computing*, 9(2):363–379, 1988.

BLS Prakasa Rao. *Statistical inference for diffusion type processes*. Arnold, 1999.

David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003.

Larry L Schumaker. *Spline functions: basic theory*, volume 1981. Wiley New York, 1981.

Bernard W Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810,

1982.

Bernard W Silverman. Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics*, pages 898–916, 1984.

Paul Speckman. Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics*, pages 970–983, 1985.

Charles J Stone. Nonparametric m-regression with free knot splines. *Journal of statistical planning and inference*, 130(1):183–206, 2005.

Charles J Stone and Jianhua Z Huang. Statistical modeling of diffusion processes with free knot splines. *Journal of statistical planning and inference*, 116(2):451–474, 2003.

Charles J Stone, Mark H Hansen, Charles Kooperberg, Young K Truong, et al. Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. *The Annals of Statistics*, 25(4):1371–1470, 1997.

Florencio Utreras. Optimal smoothing of noisy data using spline functions. *SIAM Journal on Scientific and Statistical Computing*, 2(3), 1981.

Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

Matt P Wand. Smoothing and mixed models. *Computational statistics*, 18(2):223–249, 2003.

MP Wand and MC Jones. Kernel smoothing, vol. 60 of monographs on statistics and applied probability, 1995.

Xiao Wang, Jinglai Shen, David Ruppert, et al. On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics*, 5:1–17, 2011.

Hans F Weinberger. *Variational methods for eigenvalue approximation*, volume 15. SIAM, 1974.

APPENDIX A

PROOFS ON DETAILS

A.1 Proof of equation (5.34) in probability density estimation

In section 5.4, we have outlined the lines in checking condition 3.1.1 (ii) and condition 3.1.2 with (5.34) left to be proved. We display (5.34) again here.

$$\mathrm{Var}(g_2(X_\alpha) - g_1(X_\alpha)) \asymp \int_{\mathcal{X}} (g_2(x) - g_1(x))^2 dx. \tag{A.1}$$

Let's denote the density of X_α by f_α, refering back to section 5.4, $f_\alpha(y) = \frac{\exp(g_1(y)+\alpha(g_2(y)-g_1(y)))}{\int \exp(g_1(x)+\alpha(g_2(x)-g_1(x)))dx}$.

By the definition of variance, the left hand side of (A.1) is equal to

$$\int (g_2(y) - g_1(y))^2 f_\alpha(y) dy - E(g_2(X_\alpha) - g_1(X_\alpha))^2, \tag{A.2}$$

where $E(g_2(X_\alpha) - g_1(X_\alpha)) = \int (g_2 - g_1)(x) f_\alpha(x) dx$.

It is very easy to see that the left hand side of (A.1) is bounded by the right hand side. The rest is to show that the right hand side is upper bounded by the left hand side up to multiplication of a constant. Since we have the equivalent form (A.2), it is enough to verify that

$$E(g_2(X_\alpha) - g_1(X_\alpha)) \leq A \int (g_2(y) - g_1(y))^2 f_\alpha(y) dy \tag{A.3}$$

for some $0 < A < 1$.

Since g_1, g_2 satisfy that $\int g_i(x)dx = 0, i = 1, 2$, therefore

$$E(g_2(X_\alpha)-g_1(X_\alpha)) = \int (g_2-g_1)(x)f_\alpha(x)dx = \int (g_2-g_1)(x)(f_\alpha(x)-t)dx, \text{ for any } t \in \mathbb{R}. \tag{A.4}$$

Applying the Cauchy Schwartz inequality to the right hand side of (A.4) leads to

$$E(g_2(X_\alpha) - g_1(X_\alpha)) \leq \int (g_2(x) - g_1(x))^2 f_\alpha(x)dx \int \frac{(f_\alpha(x) - t)^2}{f_\alpha(x)}dx. \tag{A.5}$$

Hence it remains to show that

$$\int \frac{(f_\alpha(x) - t)^2}{f_\alpha(x)}dx \leq A, \quad \text{for some} \quad 0 < A < 1. \tag{A.6}$$

Or equivalently,

$$t \int 2dx - t^2 \int \frac{1}{f_\alpha(x)}dx \geq \delta, \quad \text{for some} \quad \delta > 0. \tag{A.7}$$

The existence of some t such that (A.7) holds will follow from the basic knowledge of quadratic functions. Hence the proof is finished.