ASYMPTOTIC PROPERTIES OF PENALIZED SPLINE ESTIMATORS IN CONCAVE EXTENDED LINEAR MODELS: RATES OF CONVERGENCE

BY JIANHUA Z. HUANG * AND YA SU †

Texas A&M University * and Virginia Commonwealth University †

This paper develops a general theory on rates of convergence of penalized spline estimators for function estimation when the likelihood functional is concave in candidate functions, where the likelihood is interpreted in a broad sense that includes conditional likelihood, quasi-likelihood, and pseudo-likelihood. The theory allows all feasible combinations of the spline degree, the penalty order, and the smoothness of the unknown functions. According to this theory, the asymptotic behaviors of the penalized spline estimators depends on interplay between the spline knot number and the penalty parameter. The general theory is applied to obtain results in a variety of contexts, including regression, generalized regression such as logistic regression and Poisson regression, density estimation, conditional hazard function estimation for censored data, quantile regression, diffusion function estimation for a diffusion type process, and estimation of spectral density function of a stationary time series. For multi-dimensional function estimation, the theory (presented in the Supplementary Material) covers both penalized tensor product splines and penalized bivariate splines on triangulations.

1. Introduction. Since the publication of the Statistical Science discussion paper of Eilers and Marx (1996), penalized spline estimators (or penalized splines for short) have gained much popularity and have become a standard general-purpose method for function estimation. Many applications of penalized splines are presented in the monograph Ruppert, Wand and Carroll (2003). As an indication of popularity of penalized splines, a google search on “penalized splines” yields more than 200,000 results, and the Eilers and Marx (1996) paper has more than 3000 citations. Despite the popularity of penalized splines, theoretical understanding of the method falls much behind. Existing results on asymptotic behaviors of penalized splines have focused on the nonparametric regression setting. Since application of penalized splines has gone far beyond nonparametric regression, there is a big gap between theory and practice that needs to be filled in.

Hall and Opsomer (2005) obtained the asymptotic mean squared error of penalized spline estimators under a white noise model. Li and Ruppert (2008), Wang et al. (2011) and Schwarz and Krivobokova (2016) showed
that penalized spline estimators are approximately equivalent to kernel regression estimators and used this connection to obtain asymptotic properties of penalized spline estimators. Claeskens, Krivobokova and Opsomer (2009) and Xiao (2019a) obtained asymptotic results for penalized splines under weaker conditions than previously used in the literature, and identified a breakpoint in rates of convergence to classify two asymptotic situations for penalized splines, one close to smoothing splines, and one close to polynomial splines. Results on estimation of bivariate functions have been obtained by Lai and Wang (2013) for penalized bivariate splines on triangulations, and by Xiao, Li and Ruppert (2013) and Xiao (2019b) for penalized tensor product splines with different choices of penalty functionals. Holland (2017) studied asymptotic behaviors of penalized tensor product splines for estimating multi-dimensional functions. While the above papers focused on least squares regression, Kauermann, Krivobokova and Fahrmeir (2009) obtained asymptotic behaviors of penalized spline estimators for generalized regression when the regressor is univariate.

Most of the works mentioned above have used closed-form expressions of penalized spline estimators which are only available in the regression setting. When such expressions are not available in other estimation contexts, such as estimation of density functions or conditional quantile functions, existing asymptotic approaches cannot be easily extended, imposing a challenge on studying the asymptotic behaviors of penalized splines beyond nonparametric regression.

The goal of this article is to develop a new asymptotic approach to penalized spline estimators that allows us to obtain general rates of convergence results in a broad range of contexts, called concave extended linear models (Huang, 2001). We use the term “concave extended linear models” because in all these contexts, the unknown function is searched over a linear function space using a maximum-likelihood-type criterion, while the “likelihood” is a concave functional of candidate functions. As we shall see later, the family of concave extended linear models is rich, covers many useful contexts of function estimation as special cases, including regression, generalized regression such as logistic regression and Poisson regression, density estimation, conditional hazard function estimation for censored data, diffusion function estimation for a diffusion process, quantile regression, and estimation of spectral density function of a stationary time series. For readability of the paper, we present only results for univariate function estimation in the main paper. Results for multi-dimensional function estimation are obtained in the same theoretical framework, but will be presented in the Supplementary Material since they involve more complicated notations and background
on multivariate splines.

1.1. Concave extended linear models. Suppose we are interested in estimating an unknown function \( \eta_0 \) that is associated with the distribution of a random variable or vector \( W \). This function is defined on a compact set \( U \), which for concreteness is assumed to be an interval \([a,b]\). We have available an i.i.d. sample of \( W \) of size \( n \), denoted as \( W_1, \ldots, W_n \). For a candidate function \( h \) of estimating \( \eta_0 \), the (scaled) log-likelihood is

\[
\ell(h; W_1, \ldots, W_n) = \frac{1}{n} \sum_{i=1}^{n} l(h; W_i),
\]

where \( l(h; W_i) \) is the contribution to the log-likelihood from \( W_i \), and the scaling is given by the factor \( 1/n \). The expected log-likelihood is

\[
\Lambda(h) = E\{\ell(h; W_1, \ldots, W_n)\} \quad (= E\{l(h; W_i)\} \text{ if } W_i \text{ are i.i.d.})
\]

where the expectation is taken with respect to the distribution of \( W_1, \ldots, W_n \). For the rest of the paper, when there is no confusion, we will omit \( W_1, \ldots, W_n \) in the log-likelihood expression and write \( \ell(\eta) \) to simplify notation.

Assume that the set of functions for which both the log-likelihood and the expected log-likelihood are well-defined is a convex set. We say that we have a concave extended linear model if

\begin{enumerate}
\item \( \ell(h; W_1, \ldots, W_n) \) is a concave in \( h \) for all possible values of \( W_1, \ldots, W_n \), i.e., for \( 0 \leq \alpha \leq 1 \),

\[
\ell(\alpha h_1 + (1 - \alpha) h_2; W_1, \ldots, W_n) \geq \alpha \ell(h_1; W_1, \ldots, W_n) + (1 - \alpha) \ell(h_2; W_1, \ldots, W_n);
\]

\item \( \Lambda(h) \) is strictly concave in \( h \), i.e., for \( 0 \leq \alpha \leq 1 \),

\[
\Lambda(\alpha h_1 + (1 - \alpha) h_2) \geq \alpha \Lambda(h_1) + (1 - \alpha) \Lambda(h_2),
\]

and if \( 0 < \alpha < 1 \), the strict inequality holds only when it does not hold that \( h_1 = h_2, a.e. \).
\end{enumerate}

In our framework, the functional \( \ell(h) \) can be something more general than the log-likelihood function. All we need is that the function of interest, \( \eta_0 \), maximizes \( \Lambda(h) \). For example, for the regression problem, our goal is to estimate the conditional mean \( \eta_0(x) = E(Y|X = x) \), by setting \( W_i = (X_i, Y_i) \) and

\[
\ell(h; W_1, \ldots, W_n) = -\frac{1}{n} \sum_{i=1}^{n} (Y_i - h(X_i))^2,
\]
we obtain a concave extended linear model. If the conditional distribution of $Y_i$ given $X_i$ is Gaussian, $\ell(h; \mathbf{W}_1, \ldots, \mathbf{W}_n)$ can be interpreted (up to a scale factor) as the log-conditional likelihood, but this distribution assumption is not needed when applying our results in this paper. For the problem of estimating a probability density function $\eta_0$, by setting $\mathbf{W}_i = X_i$ and

$$\ell(h; \mathbf{W}_1, \ldots, \mathbf{W}_n) = \frac{1}{n} \sum_{i=1}^{n} h(X_i) - \log \int _{U} \exp h(x) \, dx,$$

we also obtain a concave extended linear model. More detailed discussions of log-likelihood function for a variety of contexts can be found in Sections 6–7 and Sections S.2–S.6 in the Supplementary Material.

1.2. Penalized spline estimators. For sample size $n$, consider a finite-dimensional space $\mathbb{G}_n$ of spline functions with degree $m$. The penalized spline estimator $\hat{\eta}_n$ is defined as the maximizer among $g \in \mathbb{G}_n$ of the following penalized likelihood

$$(2) \quad p\ell(g; \mathbf{W}_1, \ldots, \mathbf{W}_n) = \ell(g; \mathbf{W}_1, \ldots, \mathbf{W}_n) - \lambda_n J_q(g),$$

where $\ell(g; \mathbf{W}_1, \ldots, \mathbf{W}_n)$ is the log-likelihood defined in (1), $J_q(g)$ is a penalty term, and $\lambda_n$ is a penalty parameter. The penalty term $J_q(g) = J_q(g, g)$ is chosen to be a quadratic functional that quantifies the roughness of a candidate function $g$, and we use the following specific form in this paper

$$(3) \quad J_q(g) = \int _{U} \{g^{(q)}(x)\}^2 \, dx.$$

We let $q$ be an integer and refer to it as the order of the penalty. This kind of estimator was first introduced in O’Sullivan (1986), O’Sullivan (1988), and later popularized by Eilers and Marx (1996) where a modified penalty functional is used. A multi-dimensional analog of the penalty functional (3) is given in the Supplementary Material.

If we do not restrict the maximization to a finite-dimensional space in the optimization problem (2), we perform the optimization over the set of all functions such that the penalty functional is finite, which is the usual Sobolev space of order $q$

$$W^q[a, b] = \{ h : h^{(q-1)} \text{ is absolutely continuous and } J_q(h) < \infty\},$$

where $h^{(l)}$ denotes the $l$-th derivative of $h$, then the resulting estimator is a smoothing spline (Wahba, 1990; Gu, 2013). If there is no penalty term in (2)
(or $\lambda_n = 0$), we call the resulting estimator a polynomial spline estimator. In the literature, a polynomial spline function estimator is usually called a regression spline estimator mainly because the regression problem is where such an estimator was first applied to, but we prefer the former name because its application goes far beyond the regression problem. There is an extensive literature on the asymptotic theory of the smoothing spline estimators and the polynomial spline estimators, which is reviewed in the Supplementary Material.

1.3. Overview of results in this paper. For the penalized spline estimator $\hat{\eta}_n$, we obtain a probabilistic bound on the quantity $||\hat{\eta}_n - \eta_0||^2 + \lambda_n J(\hat{\eta}_n)$, where $||\cdot||$ is a norm that is equivalent to the usual $L_2$-norm. Our result not only gives the $L_2$ rate of convergence of $\hat{\eta}_n$ to the true function $\eta_0$, but also gives a bound on $J(\hat{\eta}_n)$, which measures the roughness of the estimator.

In the framework of concave extended linear models, we establish asymptotic results for penalized spline estimators under a set of high level conditions. These high level conditions help us identify the essential factors governing the asymptotic behaviors, namely, the property of the likelihood, the approximation property of the spline space, and the eigenstructure of the penalty functional. Using high level conditions allows us to obtain results in a unified manner for a wide range of problems, including the following:

- regression (Section 6),
- generalized regression (Section S.2),
- quantile regression (Section 7),
- estimation of probability density function (Section S.3),
- hazard regression for censored data (Section S.4),
- estimation of drift coefficient of diffusion type process (Section S.5),
- spectral density estimation for a stationary time series (Section S.6).

To our knowledge, our treatment of rates of convergence for the penalized spline estimators is the most comprehensive in its ability to handle a variety of estimation contexts under weak assumptions. Using high level conditions allows us to obtain results without making the strong assumption of equally-spaced knots as used by some existing works. Our results for the later five contexts are entirely new to the literature.

Our theory shows that the asymptotic behaviors of penalized splines are governed by the spline degree $m$, penalty order $q$, degree of smoothness of the unknown function $p$ (usually denoting the number of derivatives), and the interplay between the number of knots and the penalty parameter. Our results cover all feasible combinations of $m$, $p$, and $q$, while all existing works only cover selected combinations and are obtained only in the regression or
generalized regression setting. Following our main results (Sections 3), the rates of convergence of penalized splines can be classified into seven scenarios and in six of these scenarios the optimal rate of convergence can be achieved when the spline knot number and the penalty parameter are appropriately chosen (Table 1 and its discussion, Section 3.3).

Our technical approach uses functional analysis tools and avoids the detailed calculations that involve explicit expressions of penalized spline estimators as typically used in previous works. This functional analysis approach is particularly powerful in dealing with new challenges encountered when obtaining asymptotic behaviors of penalized splines beyond the regression setting. For example, one needs to handle the integration-to-one constraints for density estimation, the non-negative constraint for hazard function estimation, and non-differentiability of the “log-likelihood” for quantile regression. Since penalized spline estimators do not have a closed-form expression in general settings, the asymptotic approaches previously used for the regression setting do not apply. The functional analysis approach also allows us to treat penalized univariate splines, penalized (multivariate) tensor product splines, and penalized bivariate splines on triangulations in a unified framework.

Our technical approach has its roots in previous works for obtaining asymptotic behaviors of (un-penalized) polynomial spline estimators, as originated by Charles J. Stone in a series of works, synthesized in Stone (1994) and Hansen (1994), and matured in Huang (2001). As such, we are able to obtain existing results for polynomial spline estimators as a special simplification of our approach. On the other hand, considering a penalized likelihood in extended linear models with a roughness penalty is a substantial advancement over existing works. We obtain a rich collection of new results that reveal interesting asymptotic behaviors of penalized spline estimators that were not anticipated by Huang (2001). We also extended previous theory to deal with some contexts that were not covered by the framework of Huang (2001), such as quantile regression and spectral density estimation.

The rest of the paper is organized as follows. (Sections labeled with S. are in the Supplementary Material.) Section 2 collects some known facts on the properties of univariate spline functions and the penalty functional to make this paper self-contained. Sections 3 and 4 present respectively two master theorems and their proofs. Section 5 gives several lemmas for assisting verification of the conditions used in the master theorems. Sections 6–7 and S.2–S.6 verify those conditions under primitive conditions in a variety of contexts. Section S.1 provides a literature review on the asymptotic theory of smoothing spline estimators and polynomial spline estimators. Sections S.7
and S.8 respectively present our theory for penalized tensor-product splines and for penalized bivariate splines on triangulations.

1.4. Notation. For two real numbers $a$ and $b$, let $a \land b$ and $a \lor b$ denote respectively the smaller and larger one of the two. Given two sequences of positive numbers $a_n$ and $b_n$, we write $a_n \lesssim b_n$ and $b_n \gtrsim a_n$ if 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 $b_n \gtrsim a_n$, we write $a_n < b_n$ and $b_n > a_n$ if $a_n/b_n \to 0$ as $n \to \infty$. Let $\|g\|_2$ denote the $L_2$-norm (relative to the Lebesgue measure) and $\|g\|_\infty$ the $L_\infty$-norm of the function $g$. Throughout the paper, we use $C$, $M$, and possibly with subscripts to denote constants whose values may vary from contexts to contexts.

2. Preliminaries: splines and penalty functionals. This section provides the necessary background about spline functions and penalty functionals, introduces notations, and presents some general assumptions. In particular, it summarizes some key results from the literature about spline functions and the penalty functionals, which are essential for our study of asymptotic properties of the penalized spline estimators.

2.1. Splines. A spline function is a numerical function that is piecewise-defined by polynomial functions, and the polynomial pieces are connected smoothly. More precisely, a function $f$ defined on a compact interval $[a, b]$ is called a spline of degree $m$ with $k$ interior knots $t_j, j = 1, \ldots, k$ (satisfying $a = t_0 < t_1 < \ldots, t_k < t_{k+1} = b$), if $f$ is a polynomial of degree $m \geq 0$ on $[t_j, t_{j+1}], j = 0, \ldots, k$, and $f$ globally has $m - 1$ continuous derivatives (no derivative if $m = 0$). Note that, for a given sequence of knots, the collection of all degree-$m$ splines on $[a, b]$ forms a linear vector space with dimension $N = m + k + 1$, denoted as $G$.

When we study the asymptotic properties of penalized spline estimators, we allow the number of knots to increase with the sample size. We write $N = N_n$ and $G = G_n$ to make this dependence explicit. We assume that the knot sequence has the bounded mesh ratio. More precisely, we assume that the ratio of the maximum and minimum distance between two neighboring knots is bounded from above and below by two positive numbers that do not depend on $n$, that is,

$$C_1 \leq \frac{\max_j (t_{j+1} - t_j)}{\min_j (t_{j+1} - t_j)} \leq C_2,$$

for some $C_1, C_2 > 0$.

Let $\delta_n$ be the largest distance between all the neighboring knots, that is,

$$\delta_n = \max_j |t_{j+1} - t_j|.$$
Under the assumption of bounded mesh ratio, we have $\delta_n \asymp 1/N_n$.

The rationale for using splines in function estimation is that splines have a good approximation property, namely, they can approximate smooth functions very well when the knot number increases to infinity, as shown in the next result (Theorem 6.25 and Corollary 6.26 of Schumaker, 1981). (Please note difference in notation. We state the result in terms of spline degree, while the result in the cited book is stated using the order of splines.)

**Proposition 2.1.** Assume $\eta_0 \in W^p[a, b]$ and $m \geq p - 1$. There exist a function $\eta_n^* \in G_n$ and constants $C_1$-$C_3$, depending on $p$ and $\eta_0$ such that

$$
\|\eta_n^* - \eta_0\|_2 \leq C_1 \delta_n^p, \quad \|\eta_n^* - \eta_0\|_\infty \leq C_2 \delta_n^{p-1/2},
$$

and moreover, if $q \leq m$, then $J_q(\eta_n^*) \leq C_3 \delta_n^{2(p-q)/2}$.

If $m < p - 1$, since $\eta_0 \in W^p[a, b]$ implies that $\eta_0 \in W^{m+1}[a, b]$, the conclusion of this theorem holds by replacing $p$ with $m+1$. This approximation rate is the best one can expect: the approximation error rate cannot be better than $\delta_n^{m+1}$ even when the function $\eta_0$ has smoothness $p > m + 1$, as shown in Theorem 6.42 of Schumaker (1981). Because of the saturation phenomenon of the spline approximation, we define $p' = p \wedge (m + 1)$ and use $p'$ to measure the rate of approximation error. Moreover, we will require later that $p > 1/2$ in order to guarantee $\|\eta_n^* - \eta_0\|_\infty = o(1)$.

Following Huang (1998a,b), we introduce a measure of the complexity of a spline space,

$$
A_n = \sup_{g \in G_n, \|g\|_2 \neq 0} \left\{ \frac{\|g\|_\infty}{\|g\|_2} \right\},
$$

This measure will play an important role in the asymptotic analysis. The next result, which follows from Theorem 5.1.2 of DeVore and Lorentz (1993), gives the rate of increase of $A_n$.

**Proposition 2.2.** Under the bounded mesh ratio condition, $A_n \asymp \delta_n^{-1/2}$.

The asymptotic analysis of spline estimators relies on an important property of spline spaces, namely, the uniformly closeness of a data-driven norm to its expectation over the entire spline space for a fixed degree and fixed knot sequence (they vary with $n$). Let $X, X_1, \ldots, X_n$ be i.i.d. random variables. Define the empirical and theoretical inner products as

$$
\langle g_1, g_2 \rangle_n = E_n[g_1(X)g_2(X)w(X)] = \frac{1}{n} \sum_{i=1}^n g_1(X_i)g_2(X_i)w(X_i),
$$

$$
\langle g_1, g_2 \rangle = E[\langle g_1(X)g_2(X)w(X) \rangle],
$$
where \( w(x) \) is a weight function bounded away from zero and infinity, that is, there exists \( C_1, C_2 > 0 \) such that

\[
C_1 \leq w(x) \leq C_2, \quad \text{for any } a \leq x \leq b.
\]

The corresponding squared empirical and theoretical norms are \( \| g \|_n^2 = \langle g, g \rangle_n \) and \( \| g \|^2 = \langle g, g \rangle \). We assume that \( X \) has a density function which is bounded away from zero and infinity, and consequently the theoretical norm \( \| \cdot \| \) is equivalent to \( \| \cdot \|_2 \), the usual \( L_2 \)-norm relative to the Lebesgue measure, i.e., there are constants \( C_3 \) and \( C_4 \) such that

\[
C_3 \| g \|_2 \leq \| g \| \leq C_4 \| g \|_2
\]

for all square-integrable function \( g \).

**Proposition 2.3.** Under the bounded mesh ratio condition, if \( \lim_n N_n \log(n)/n = 0 \), then the empirical and theoretical norms are asymptotically equivalent, that is,

\[
\sup_{g \in G_n, \| g \| \neq 0} \left| \frac{\| g \|_n}{\| g \| - 1} \right| = o_P(1).
\]

Huang (1998a) proved Proposition 2.3 for an arbitrary finite dimensional function space under the stronger condition that \( \lim_n A_n^2 N_n/n = 0 \). Huang (2003) relaxed the condition to \( \lim_n N_n \log(n)/n = 0 \) for splines. Both papers proved the results for \( w(x) = 1 \) but the same argument applies to a general weight function that is bounded away from zero and infinity.

### 2.2. The penalty functional

The asymptotic properties of the penalized spline estimator rely heavily on the eigenanalysis of the quadratic penalty functional \( J_q(h) = \int_{\mathcal{U}} \{ h^{(q)}(x) \}^2 dx \) with respect to the quadratic functional \( V(h) = \| h \|^2 = \int_{\mathcal{U}} h^2(x) \omega(x) dx \). Such eigenanalysis also plays a critical role in studying the asymptotic properties of the smoothing splines; see, i.e., Gu (2013).

A quadratic functional \( B \) is said to be completely continuous with respect to another quadratic functional \( A \), 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 \( B(h) \leq \epsilon A(h) \). See Weinberger (1974, Section 3.3).

Applying Theorem 3.1 of Weinberger (1974), it can be shown that, if \( V \) is completely continuous with respect to \( J \), then \( V \) and \( J \) can be simultaneously diagonalized in the following sense (see Section 9.1 of Gu, 2013). There exists a sequence of eigenfunctions \( \phi_\nu, \nu = 1, 2, \ldots, \) and the associated sequence of eigenvalues \( \rho_\nu \geq 0 \) of \( J \) with respect to \( V \) such that

\[
V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}, \quad J_q(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu\mu},
\]
where $\delta_{\nu\mu}$ is the Kronecker delta,

$$V(\phi_\nu, \phi_\mu) = \int_{U} \phi_\nu(x) \phi_\mu(x) \omega(x) \, dx, \quad J_q(\phi_\nu, \phi_\mu) = \int_{U} \phi_\nu^{(q)}(x) \phi_\mu^{(q)}(x) \, dx.$$  

See also Silverman (1982). Furthermore, any function $h$ satisfying $J_q(h) < \infty$ has a Fourier series expansion with the eigen basis $\{\phi_\nu\}$,

$$h = \sum_\nu h_\nu \phi_\nu, \quad h_\nu = V(h, \phi_\nu),$$

and

$$V(h) = \sum_\nu h_\nu^2, \quad J_q(h) = \sum_\nu \rho_\nu h_\nu^2.$$ 

Therefore,

$$\|h\|^2 + \lambda_n J_q(h) = (V + \lambda_n J)(h) = \sum_\nu (1 + \lambda_n \rho_\nu) h_\nu^2.$$ 

The next result (see (3.17) of Utreras, 1981) gives the rate of divergence to infinity of the eigenvalues.

**Proposition 2.4.** Assume $V(h) = \|h\|^2 = \int_{U} h^2(x) \omega(x) \, dx$ for a weight function $\omega$ that is bounded away from zero and infinity, that is, there exist constants $C_1, C_2 > 0$ such that

$$C_1 \leq \omega(x) \leq C_2, \quad \text{for any } a \leq x \leq b.$$ 

Then $V$ is completely continuous with respect to $J_q$. Moreover, we have $0 \leq \rho_\nu \uparrow \infty$, and $\rho_\nu \asymp \nu^{2q}$ for all sufficiently large $\nu$.

The following result, which is part of Lemma 9.1 of Gu (2013), will be used when studying the rate of convergence of the estimation error (see Lemma 5.2).

**Proposition 2.5.** Assume there is a constant $C > 0$ such that $\rho_\nu \geq C \nu^{2q}$ ($q > 1/2$), for all large $\nu$. If $\lambda_n \to 0$, as $n \to \infty$, then

\[
\sum_\nu \frac{1}{1 + \lambda_n \rho_\nu} = O(\lambda_n^{-1/(2q)}).
\]
3. Statement of the master theorems. The rate of convergence of a penalized spline estimator depends on three quantities:

- \( p \) —the smoothness of the unknown function (i.e., we assume \( \eta_0 \in W^p[a,b] \));
- \( m \) —the degree of the splines in \( G_n \);
- \( q \) —the order of the penalty functional \( J_q(g) = \int_{d} (g^{(q)}(x))^2 \, dx \).

Here, \( m + 1 \) is also called the order of the spline functions.

We make several (natural) restrictions on the choice of \( p, m, q \), as follows.

- \( q \leq m \). Since the \( m \)th derivative of a spline function of degree \( m \) is piecewise constant, the \((m + 1)\)th derivative of the spline function contains Dirac delta functions, therefore the \((m + 1)\)th order penalty functional is not defined, thus it is natural to require that \( q \leq m \).
- \( p > 1/2 \). This is to ensure that the spline space has desired approximation properties (see Proposition 2.1).
- \( q > 1/2 \). This is to ensure the eigenvalues of the penalty functional have desired rate of divergence (see Proposition 2.5).

In this paper, we also restrict \( p \) and \( q \) to be integer-valued, which is the most relevant in practical applications. To relax this restriction, one needs only to supply a version of Propositions 2.1 and 2.4 that allow non-integer values of \( p \) and \( q \). The rest of technical arguments is not affected.

The expected value of the penalized log-likelihood \( p\ell(\eta) \) appeared in (2) is

\[
p\Lambda(\eta) = \Lambda(\eta) - \lambda_n J_q(\eta).
\]

Denote its maximizers as

\[
\hat{\eta}_n = \arg \max_{g \in G_n} p\Lambda(g) = \arg \max_{g \in G_n} \{ \Lambda(g) - \lambda_n J_q(g) \}.
\]

We can think that \( \hat{\eta}_n \) is an approximation of \( \eta_0 \), and the penalized spline estimator \( \hat{\eta}_n \) directly estimates \( \bar{\eta}_n \). Therefore we have the decomposition

\[
\hat{\eta}_n - \eta_0 = \hat{\eta}_n - \bar{\eta}_n + (\bar{\eta}_n - \eta_0),
\]

where \( \hat{\eta}_n - \bar{\eta}_n \) and \( \bar{\eta}_n - \eta_0 \) are referred to as the estimation error and the approximation error, respectively.

3.1. Approximation Error.

**Condition 3.1.** There are constants \( B > 0 \) and constants \( M_1, M_2 > 0 \) such that

\[
-M_1 \|h\|^2 \leq \Lambda(\eta_0 + h) - \Lambda(\eta_0) \leq -M_2 \|h\|^2
\]
whenever \( \|h\|_{\infty} \leq B \).

This condition says that the expected log-likelihood behaves like a quadratic functional around its maximal point.

Recall \( p' = p \wedge (m + 1) \), as defined after Proposition 2.1.

**Theorem 3.1.** Assume Condition 3.1 holds. If \( \lim_n \delta_n \vee \lambda_n = 0 \) and

\[
\lim_n A_n^2 \{ \delta_n^2 p' \vee (\lambda_n \delta_n^2(p'-q)^\wedge 0) \} = 0,
\]

then \( \bar{\eta}_n \) exists and \( \|\bar{\eta}_n\|_{\infty} = O(1) \). Moreover,

\[
\|\bar{\eta}_n - \eta_0\|_2 + \lambda_n J_q(\bar{\eta}_n) = O\{ \delta_n^2 p' \vee (\lambda_n \delta_n^2(p'-q)^\wedge 0) \}.
\]

**3.2. Estimation Error.** To simplify notation, we shall omit \( W_1, \ldots, W_n \) in \( \ell(h; W_1, \ldots, W_n) \) when we write the log-likelihood functional.

Because \( l(\bar{\eta}_n + \alpha g) \) is a concave function of \( \alpha \), it admits left and right derivatives and is differentiable at all but countable many points. Denote the directional derivative at \( \bar{\eta}_n \) along the direction of \( g \) as

\[
\dot{l}[\bar{\eta}_n; g] = \left. \frac{d}{d\alpha}l(\bar{\eta}_n + \alpha g) \right|_{\alpha=0^+},
\]

where the dependence \( \dot{l}[\bar{\eta}_n; g] \) on \( W_i \) is suppressed in our notation for simplicity. Using the mild assumption that we can exchange differentiation and expectation, we have

\[
E\{ \dot{l}[\bar{\eta}_n; g] \} = (d/d\alpha)\Lambda(\bar{\eta}_n + \alpha g)\big|_{\alpha=0^+}.
\]

Since \( \bar{\eta}_n \) maximizes the concave functional \( p\Lambda(\cdot) \) over \( \mathcal{G}_n \), it satisfies the first order condition

\[
\left. \frac{d}{d\alpha} \Lambda(\bar{\eta}_n + \alpha g) \right|_{\alpha=0^+} - 2\lambda_n J_q(\bar{\eta}_n, g) = 0, \quad g \in \mathcal{G}_n.
\]

Thus, for any \( g \in \mathcal{G}_n \), we have that

\[
\left. \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \right|_{\alpha=0^+} - 2\lambda_n J_q(\bar{\eta}_n, g) = \left. \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \right|_{\alpha=0^+} - \left. \frac{d}{d\alpha} \Lambda(\bar{\eta}_n + \alpha g) \right|_{\alpha=0^+}.
\]

Consequently, we have

\[
(10) \quad \left. \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \right|_{\alpha=0^+} - 2\lambda_n J_q(\bar{\eta}_n, g) = (E_n - E) \dot{l}[\bar{\eta}_n; g].
\]
Condition 3.2.  

(i) 
\[ \sup_{g \in G_n} \frac{|(E_n - E) \hat{\ell}[\bar{\eta}_n; g]|^2}{\|g\|^2 + \lambda_n J_q(g)} = O_p \left( \frac{1}{n \lambda_n^{1/(2q)}} \wedge \frac{1}{n \delta_n} \right). \]

(ii) There are constants \( B > 0 \) and \( M > 0 \) such that, with probability tending to one as \( n \to \infty \), we have that for all \( g \in G_n \) with \( \|g\|_\infty \leq B \),
\[ \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \bigg|_{\alpha=1^+} - \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \bigg|_{\alpha=0^+} \leq -M \|g\|^2. \]

Theorem 3.2.  Assume Condition 3.2 holds. If \( \lim_n \delta_n \vee \lambda_n = 0 \) and
\[ \lim_n A_n^2 \left( \frac{1}{n \lambda_n^{1/(2q)}} \wedge \frac{1}{n \delta_n} \right) = 0, \]
then \( \|\hat{\eta}_n - \tilde{\eta}_n\|_\infty = o_P(1) \) and
\[ \|\hat{\eta}_n - \tilde{\eta}_n\|^2 + \lambda_n J_q(\hat{\eta}_n) = O_p \left( \frac{1}{n \lambda_n^{1/(2q)}} \wedge \frac{1}{n \delta_n} \right). \]

3.3. Summary.  Combining the results of Theorems 3.1 and 3.2, we obtain the following result that gives the rate of convergence of \( \|\hat{\eta}_n - \eta_0\|^2 \) to zero. The result also gives a bound for the size of \( J_q(\hat{\eta}_n) \).

Corollary 3.3.  Assume Conditions 3.1 and 3.2 hold. If \( \lim_n \delta_n \vee \lambda_n = 0 \) and
\[ \lim_n A_n^2 \left( \delta_n^{2p'} \vee (\lambda_n \delta_n^{2(p'-q)\wedge 0}) + \frac{1}{n \lambda_n^{1/(2q)}} \wedge \frac{1}{n \delta_n} \right) = 0, \]
then \( \|\hat{\eta}_n - \eta_0\|_\infty = o_P(1) \) and
\[ \|\hat{\eta}_n - \eta_0\|^2 + \lambda_n J_q(\hat{\eta}_n) = O_p \left( \delta_n^{2p'} \vee (\lambda_n \delta_n^{2(p'-q)\wedge 0}) + \frac{1}{n \lambda_n^{1/(2q)}} \wedge \frac{1}{n \delta_n} \right). \]

This result covers all practical combinations of \( p \), \( q \) and \( m \) with the only restriction being the necessary requirement \( q \leq m \) (otherwise the penalty functional is not defined). Following this result, the asymptotic behavior of the penalized splines can be classified into seven scenarios as shown in Table 1. Cases II.1 and II.2 contain a typical application scenario of using penalized cubic splines with a second-order penalty \( (m = 3, q = 2) \) to estimate a function with a continuous second derivative \( p = 2 \). Using Proposition 2.2, Condition (11) can be simplified in each scenario as follows:
We refer to these cases as the light penalty scenarios when using a small \( \lambda_n \) indicates light penalization, we may refer to these cases as the light penalty scenarios.

Table 1

| Rate of convergence | Parameters for achieving the best rate | Best rate |
|---------------------|---------------------------------------|-----------|
| I. \( q < p' \) (i.e., \( q < p \) and \( q < m + 1 \)) | \( \lambda_n \lesssim \delta_n^{2p'} \) \( \delta_n^{2p'} + (n\delta_n)^{-1} \delta_n \approx n^{-1/(2p'+1)} \) \( n^{-2p'/(2p'+1)} \) (*) | |
| II. \( q = p' (= p) \) (i.e., \( p = q \leq m \)) | \( \lambda_n \lesssim \delta_n^{2p} \) \( \delta_n^{2p} + (n\delta_n)^{-1} \delta_n \approx n^{-1/(2p+1)} \) \( n^{-2p/(2p+1)} \) (***) | |
| III. \( q > p' (= p) \) (i.e., \( p < q \leq m \)) | \( \lambda_n \lesssim \delta_n^{2q} \) \( \lambda_n \approx n^{1/(2q)} \) \( \lambda_n \approx n^{-2q/(2q+1)} \) \( n^{-2q/(2q+1)} \) (***) | |

(*) achieving Stone’s optimal rate when \( p' = p \), (***) achieving Stone’s optimal rate

- Cases I.1, II.1, III.1: \( p' > 1/2 \), \( n\delta_n^2 \to \infty \).
- Case I.2: \( n\delta_n^2 \to \infty \), \( \lambda_n/\delta_n \to 0 \).
- Cases I.3, II.2, III.2: \( n\delta_n \lambda_n^{1/(2q)} \to \infty \) (or its sufficient condition \( n\delta_n^2 \to \infty \), \( \lambda_n/\delta_n \to 0 \).)

An overall sufficient condition for all these conditions to hold is \( p' > 1/2 \), \( n\delta_n^2 \to \infty \), and \( \lambda_n/\delta_n^{1+2(1-p)^{1/2}} \to 0 \).

From Table 1, we observe that the asymptotic behavior of the penalized spline estimators depend on the interplays among the smoothness of unknown function, spline degree, penalty order, spline knot number, and penalty parameter.

In Cases I.1, I.2, II.1. III.1, \( \lambda_n \lesssim \delta_n^{2q} \). Since using a small \( \lambda_n \) indicates light penalization, we may refer to these cases as the light penalty scenarios.
Alternatively, since $\delta_n^{-1} \lesssim \lambda_n^{1/(2q)}$ and $\delta_n^{-1}$ essentially quantifies the number of knots, we may also refer to these cases as the small knot number scenarios. The behavior of the penalized splines in these scenarios is similar to that of an unpenalized polynomial spline estimator (e.g., Huang, 2003). In Cases I.1 (if $p \leq m + 1$), II.1, III.1, the penalized spline estimator achieves Stone’s optimal rate of convergence $n^{-2p/(2p+1)}$ (Stone, 1982), if the tuning parameter $\delta_n$ is chosen such that $\delta_n \asymp n^{-1/(2p+1)}$. In Case I.2 (if $p \leq m + 1$), Stone’s optimal rate can be achieved if we tune both parameters so that $\delta_n \asymp n^{-1/(2p+1)}$ and $\lambda_n \asymp \delta_n^{2p}$. If $p > m + 1$ (Cases I.1 and I.2), the best rate of convergence of penalized spline estimator is controlled by the spline order $m + 1$, Stone’s optimal rate of convergence cannot be achieved, as for the unpenalized polynomial spline estimators; this is due to the saturation of spline approximation (see the discussion following Proposition 2.1).

In Cases I.3, II.2, III.2, $\lambda_n \gtrsim \delta_n^{2q}$. We may refer to these cases as the heavy penalty scenarios. Alternatively, since $\delta_n^{-1} \gtrsim \lambda_n^{1/(2q)}$, we may also refer to these cases as the large knot number scenarios. The behavior shown in Case II.2 is similar to that of a smoothing spline estimator (e.g., Gu, 2013) and Stone’s optimal rate of convergence $n^{-2p/(2p+1)}$ can be achieved by choosing $\lambda_n \asymp n^{-2p/(2p+1)}$. The results for Cases I.3 and III.2 show different behaviors of the penalized spline estimators in the heavy penalty scenarios when the penalty order $q$ differs from the smoothness $p$ of the unknown function. If $q < p$ (Case I.3), the best rate of convergence of penalized spline estimator is controlled by $q$, which is $n^{-2q/(2q+1)}$ and is slower than Stone’s optimal rate $n^{-2p/(2p+1)}$. This result suggests that, in heavy penalty scenarios, using a penalty with order smaller than the true smoothness will hurt the ability of penalized splines to achieve the optimal rate of convergence. On the other hand, if $q > p$ (Case III.2), the penalized spline estimator can achieve Stone’s optimal rate if we tune both parameters so that $\lambda_n \asymp n^{-2p/(2p+1)}$ and $\delta_n \asymp \lambda_n^{1/(2q)}$.

In the context of least squares regression, rates of convergence for penalized spline estimators have been extensively studied when $q \leq p$ (corresponding to Cases I and II in Table 1); the best available results are given in Claeskens, Krivobokova and Opsomer (2009); Holland (2017); Xiao (2019a). Our results match the best available results for Cases I.1, I.3, II.2, II.3. For Case I.2, the best available result for rate of convergence is $\lambda_n^2 \delta_n^{-2q} + (n\delta_n)^{-1}$ (e.g., Theorem 1(a) of Claeskens, Krivobokova and Opsomer (2009) for $p = m + 1$, Theorem 5.1 of Xiao (2019a)), which is always no larger than the rate shown in Table I, $\lambda_n + (n\delta_n)^{-1}$. When $p \leq m + 1$ so that $p' = p$, to achieve Stone’s optimal rate, one needs to choose $\delta_n \asymp n^{-1/(2p+1)}$ in $\lambda_n^2 \delta_n^{-2q} + (n\delta_n)^{-1}$, and also require that $\lambda_n n^{-2q} \lesssim n^{-2p/(2p+1)}$, or equiva-
lently \( \lambda_n \lesssim n^{-(p+q)/(2p+1)} \). This requirement on \( \lambda_n \) is slightly looser than our requirement \( \lambda_n \lesssim n^{-2p/(2p+1)} \) in Cases I.1 and I.2 of Table 1.

It is worthwhile to point out that our result in Corollary 3.3 not only bound the squared \( L_2 \)-norm \( \| \hat{\eta}_n - \eta_0 \|^2 \) but also bound the penalty functional \( J_q(\hat{\eta}_n) \), and thus it is stronger than existing results which bound only the \( L_2 \)-norm. For this reason, we believe our rate of convergence in Case I.2, \( \lambda_n + (n\delta_n)^{-1} \), cannot be improved to match the best available result of squared \( L_2 \)-norm rate \( \lambda_n^2 \delta_n^{-2q} + (n\delta_n)^{-1} \) mentioned above. To see this, suppose otherwise, i.e.,

\[
\| \hat{\eta}_n - \eta_0 \|^2 + \lambda_n J_q(\hat{\eta}_n) = O(\lambda_n^2 \delta_n^{-2q} + (n\delta_n)^{-1}).
\]

When \( \lambda_n^2 \delta_n^{-2q} \geq (n\delta_n)^{-1} \), the first term dominates the rate of convergence, and we have \( J_q(\hat{\eta}_n) = O(\lambda_n \delta_n^{-2q}) \). If \( \lambda_n/\delta_n^{2q} \to 0 \) (which falls in Case I.2), then we obtain \( J_q(\hat{\eta}_n) \to 0 \), which is generally implausible. For instance, \( J_q(\hat{\eta}_n) = 0 \) for \( q = 2 \) means that \( \hat{\eta}_n \) is a straight line, and \( J_2(\hat{\eta}_n) \to 0 \) suggests that \( \hat{\eta}_n \) becomes closer and closer to a straight line when the sample size \( n \to \infty \).

We are not aware any existing results for Cases III.1 and III.2. Our results for these two scenarios answer the following question: When the smoothness of the unknown function is not given, if one uses a penalty that assumes more derivatives than the unknown function, how will the penalized spline estimator behave asymptotically? Our answer is that it does not hurt the ability of penalized spline estimator to achieve Stone’s optimal rate of convergence. This question is of interest because in practice prior knowledge about the degree of smoothness of the unknown function is usually unavailable.

4. Proof of the master theorems. This section gives the proof of the main theorems of convergence rates of the penalized spline estimator, that is, Theorems 3.1 and 3.2. The argument makes use of the convexity and is an extension of that in Huang (2001). We first present a lemma that will play an important role in our proof.

**Lemma 4.1 (Convexity Lemma).** Suppose \( C(\cdot) \) is a convex functional and \( L(\cdot) \) is a continuous functional defined on a convex set \( \mathcal{C} \) of functions. If there exists a function \( \eta^\dagger \in \mathcal{C} \) and a real number \( s \) with \( L(\eta^\dagger) < s \) such that for all \( \eta \in \mathcal{C} \) satisfying \( L(\eta) = s \), we have either

\[
C(\eta^\dagger) < C(\eta), \tag{12}
\]

or

\[
\frac{\partial}{\partial \beta} C(\eta^\dagger + \beta(\eta - \eta^\dagger)) \bigg|_{\beta = 1^+} > 0, \tag{13}
\]
then any minimizer \( \eta_{\text{min}} \) of \( C(\cdot) \) in \( C \) satisfies \( L(\eta_{\text{min}}) \leq s \).

**Proof.** Fix any \( \tilde{\eta} \in C \) with \( L(\tilde{\eta}) > s \). Consider the convex combination of \( \eta^\dagger \) and \( \tilde{\eta} \)
\[
\eta_\alpha = \alpha \tilde{\eta} + (1 - \alpha) \eta^\dagger, \quad 0 \leq \alpha \leq 1.
\]
Define \( f(\alpha) = L(\eta_\alpha) \). It is a continuous function of \( \alpha \). Since \( f(0) = L(\eta^\dagger) < s \) and \( f(1) = L(\tilde{\eta}) > s \), by the intermediate value theorem, there exists an \( \bar{\alpha} \in (0, 1) \) such that \( f(\bar{\alpha}) = s \). Denote \( \bar{\eta} = \bar{\alpha} \tilde{\eta} + (1 - \bar{\alpha}) \eta^\dagger \). Immediately \( L(\bar{\eta}) = f(\bar{\alpha}) = s \).

If (12) holds, from the convexity of \( C(\cdot) \), we have
\[
C(\eta^\dagger) < C(\bar{\eta}) \leq \bar{\alpha} C(\tilde{\eta}) + (1 - \bar{\alpha}) C(\eta^\dagger),
\]
which implies
\[
C(\eta^\dagger) < C(\bar{\eta}). \tag{14}
\]
On the other hand, we can write \( \tilde{\eta} = \eta^\dagger + \bar{\beta}(\bar{\eta} - \eta^\dagger) \), where \( \bar{\beta} = \bar{\alpha}^{-1} > 1 \). If (13) holds, then
\[
C(\eta^\dagger) - C(\bar{\eta}) = C(\eta^\dagger + \bar{\beta}(\bar{\eta} - \eta^\dagger)) - C(\eta^\dagger + (\bar{\eta} - \eta^\dagger)) \geq (\bar{\beta} - 1) \frac{\partial}{\partial \beta} C(\eta^\dagger + \beta(\bar{\eta} - \eta^\dagger)) \bigg|_{\beta = 1^+} > 0.
\]
Both (14) and (15) imply that \( \tilde{\eta} \) with \( L(\tilde{\eta}) > s \) cannot be the minimizer of \( C(\cdot) \). \( \square \)

**Proof of Theorem 3.1.** We assume \( p \leq m + 1 \) without loss of generality, since we can replace \( p \) by \( p' = p \wedge (m + 1) \) otherwise. For \( \eta_n^* \) as in Proposition 2.1, we have that \( \| \eta_n^* - \eta_0 \| \leq C_1 \delta_n^p \) and \( J_q(\eta_n^*) \leq C_3 \delta_n^{2(p-q)\wedge 0} \). Therefore
\[
\| \eta_n^* - \eta_0 \| + \lambda_n^{1/2} J_q^{1/2}(\eta_n^*) \leq C_1 \delta_n^p + C_3^{1/2} \lambda_n^{1/2} \delta_n^{(p-q)\wedge 0}. \tag{16}
\]
In the following we will repeatedly use the inequality
\[
\frac{1}{2}(u + v)^2 \leq u^2 + v^2 \leq (u + v)^2, \quad u, v > 0.
\]
to bound \( \delta_n^p + \lambda_n^{1/2} \delta_n^{(p-q)\wedge 0} \) and \( \delta_n^{2p} + \lambda_n \delta_n^{2(p-q)\wedge 0} \) by each other.

We apply the Convexity Lemma (Lemma 4.1) to the convex functional
\[
C(g) = -\Lambda(g) + \lambda_n J_q(g)
\]
and the continuous functional
\[ L(g) = \|g - \eta_n^*\| + \lambda_n^{1/2} J_q^{1/2}(g - \eta_n^*), \]
both defined on \( C = \mathbb{G}_n \). The continuity of \( L(g) \) follows from the fact that
\[ |L(g_1) - L(g_2)| \leq \|g_1 - g_2\| + \lambda_n^{1/2} J_q^{1/2}(g_1 - g_2). \]

When applying the lemma, take \( s = a(\delta_n^p + \lambda_n^{1/2} \delta_n^{(p-q)\wedge 0}) \), where \( a > 0 \) is a constant to be determined later.

Take \( \eta^0 = \eta_n^* \) in Lemma 4.1. We have \( L(\eta_n^*) = 0 \). We will show that
\[ C(\eta_n^*) < C(g), \quad g \in \mathbb{G}_n \text{ with } L(g) = s. \]

Then, the Convexity Lemma implies that the minimizer \( \bar{\eta}_n \) of \( C(g) \) in \( \mathbb{G}_n \) satisfies \( L(\bar{\eta}_n) < s \). Consequently, by the triangle inequality and (16),
\[ \| \bar{\eta}_n - \eta_0 \| + \lambda_n^{1/2} J_q^{1/2}(\bar{\eta}_n) \leq L(\bar{\eta}_n) + \| \eta_n^* - \eta_0 \| + \lambda_n^{1/2} J_q^{1/2}(\eta_n^*) \]
\[ \leq a(\delta_n^p + \lambda_n^{1/2} \delta_n^{(p-q)\wedge 0}) + C_1 \delta_n^p + C_3 \lambda_n^{1/2} \delta_n^{(p-q)\wedge 0}. \]

By using (17), we have that
\[ \| \bar{\eta}_n - \eta_0 \|^2 + \lambda_n J_q(\bar{\eta}_n) = O(\delta_n^{2p} \vee \lambda_n \delta_n^{(p-q)\wedge 0}), \]
which is the desired result.

It remains to show (18). By Proposition 2.1, \( \| \eta_n^* - \eta_0 \|_\infty \leq C_2 \delta_n^{p-1/2} \). For \( g \in \mathbb{G}_n \) with \( L(g) \leq s \), we have
\[ \| g - \eta_n^* \|_\infty \leq A_n \| g - \eta_n^* \| \leq A_n L(g) \leq A_n a(\delta_n^p + \lambda_n^{1/2} \delta_n^{(p-q)\wedge 0}), \]
and therefore,
\[ \| g - \eta_0 \|_\infty \leq \| g - \eta_n^* \|_\infty + \| \eta_n^* - \eta_0 \|_\infty \]
\[ \leq A_n a(\delta_n^p + \lambda_n^{1/2} \delta_n^{(p-q)\wedge 0}) + C_3 \delta_n^{p-1/2} = o(1) \]
(since \( p > 1/2 \)). Thus, \( \| g - \eta_0 \|_\infty \leq B \) when \( n \) is large, for \( B \) in Condition 3.1. Then, use Condition 3.1 to obtain
\[ C(g) + \Lambda(\eta_0) = -\Lambda(g) + \Lambda(\eta_0) + \lambda_n J_q(g) \]
\[ \geq M_1 \| g - \eta_0 \|^2 + \lambda_n J_q(g) \]
\[ \geq \frac{1}{2} (M_1 \wedge 1) \{ \| g - \eta_0 \| + \lambda_n^{1/2} J_q^{1/2}(g) \}^2, \]
and
\[
C(\eta_n^\ast) + \Lambda(\eta_0) = -\Lambda(\eta_n^\ast) + \Lambda(\eta_0) + \lambda_n J_q(\eta_n^\ast)
\leq M_2 \|\eta_n^\ast - \eta_0\|^2 + \lambda_n J_q(\eta_n^\ast)
\leq (M_2 \vee 1) \{\|\eta_n^\ast - \eta_0\| + \lambda_n^{1/2} J_q^{1/2}(\eta_n^\ast)\}^2.
\]

For \(g \in \mathcal{G}_n\) with \(L(g) = s\), by the triangle inequality and (16), we have that
\[
a(\delta_n s + \lambda_n^{1/2} \delta_n^{(p-q)/2}) = \|g - \eta_n^\ast\| + \lambda_n^{1/2} J_q^{1/2}(g - \eta_n^\ast)
\leq \|g - \eta_0\| + \lambda_n^{1/2} J_q^{1/2}(g) + \|\eta_n^\ast - \eta_0\| + \lambda_n^{1/2} J_q^{1/2}(\eta_n^\ast)
\leq \|g - \eta_0\| + \lambda_n^{1/2} J_q^{1/2}(g) + C_1 \delta_n + C_3 \lambda_n^{1/2} \delta_n^{(p-q)/2}.
\]

Using the above inequality and (16) we obtain that, by taking \(a\) large enough, the right hand side of (22) is strictly greater than the right hand side of (23). This proves (18).

It follows from (21) that, for any \(g \in \mathcal{G}_n\) with \(L(g) \leq s\), we have
\[
\|g\|_\infty \leq \|g - \eta_0\|_\infty + \|\eta_0\|_\infty < M\|\eta_0\|_\infty
\]
for large \(n\). Since \(L(\bar{\eta}_n) < s\), (24) implies that \(\|\bar{\eta}_n\|_\infty \leq M\|\eta_0\|_\infty < \infty\). It follows again from (21) that \(\|\bar{\eta}_n - \eta_0\|_\infty = o(1)\). The proof is complete.

**Proof of Theorem 3.2.** We apply the Convexity Lemma (Lemma 4.1) to the convex functional
\[
C(g) = -\ell(g) + \lambda_n J_q(g)
\]
and the continuous functional
\[
L(g) = \|g - \bar{\eta}_n\| + \lambda_n^{1/2} J_q^{1/2}(g - \bar{\eta}_n),
\]
both defined on \(\mathcal{C} = \mathcal{G}_n\). We take
\[
s^2 = a^2 \left(\frac{1}{n\lambda_n^{1/2}} \wedge \frac{1}{n\delta_n}\right)
\]
when applying this lemma, where \(a > 0\) is a constant to be determined later.

Take \(\eta^\dagger = \bar{\eta}_n\) in Lemma 4.1. We have \(L(\bar{\eta}_n) = 0 < s\). We will show that
\[
(25) \quad \frac{\partial}{\partial \alpha} C(\bar{\eta}_n + \alpha(g - \bar{\eta}_n)) \bigg|_{\alpha = 1^+} > 0, \quad g \in \mathcal{G}_n\text{ with } L(g) = s.
\]
Then, the Convexity Lemma implies that the minimizer \( \hat{\eta}_n \) of \( C(g) \) in \( \mathcal{G}_n \) satisfies \( L(\hat{\eta}_n) \leq s \). Hence, 
\[
\| \hat{\eta}_n - \bar{\eta}_n \|^2 + \lambda_n J_q(\hat{\eta}_n - \bar{\eta}_n) \leq s^2 = \alpha^2 \left( \frac{1}{n\lambda_n^{1/(2q)}} \right) \wedge \frac{1}{n\delta_n} \tag{26}
\]
which is the desired result.

It remains to show \( (25) \). Because \( J_q(\cdot) \) is a quadratic functional, we have the expansion
\[
J_q(\eta_n + \alpha(g - \bar{\eta}_n)) = J_q(\bar{\eta}_n) + 2\alpha J_q(\bar{\eta}_n, g - \bar{\eta}_n) + \alpha^2 J_q(g - \bar{\eta}_n).
\]
This together with the definition of \( C(\cdot) \) imply that
\[
\left. \frac{\partial}{\partial \alpha} C(\eta_n + \alpha(g - \bar{\eta}_n)) \right|_{\alpha=1^+} = I + II
\]
where (using \( (10) \))
\[
I = -\frac{d}{d\alpha} \ell(\eta_n + \alpha(g - \bar{\eta}_n)) \bigg|_{\alpha=0^+} + 2\lambda_n J_q(\bar{\eta}_n, g - \bar{\eta}_n) = -\left( E_n - E \right) \hat{\ell}(\bar{\eta}_n; g - \bar{\eta}_n),
\]
and
\[
II = -\frac{d}{d\alpha} \ell(\eta_n + \alpha(g - \bar{\eta}_n)) \bigg|_{\alpha=0^+} + \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha(g - \bar{\eta}_n)) \bigg|_{\alpha=0^+} + 2\lambda_n J_q(g - \bar{\eta}_n).
\]

Now consider \( g \in \mathcal{G}_n \) with \( L(g) \leq s \). By Condition 3.2 (i),
\[
|I| = \{ \| g - \bar{\eta}_n \|^2 + \lambda_n J_q(g - \bar{\eta}_n) \}^{1/2} O_P \left( \left( \frac{1}{n\lambda_n^{1/(2q)}} \right) \wedge \frac{1}{n\delta_n} \right)^{1/2} \tag{27}
\]
\[
\leq s O_P \left( \frac{s^2}{a} \right) = O_P \left( \frac{s^2}{a} \right).
\]
On the other hand, by the definition of \( A_n \),
\[
\| g - \bar{\eta}_n \| \leq A_n \| g - \bar{\eta}_n \| = A_n \alpha \left( \frac{1}{n\lambda_n^{1/(2q)}} \right) \wedge \frac{1}{n\delta_n} \right)^{1/2} = o(1).
\]
Thus, \( \| g - \bar{\eta}_n \| \leq B \) for large \( n \) where \( B \) is as in Condition 3.2 (ii). It then follows from this condition that, for \( g \in \mathcal{G}_n \) with \( L(g) = s \),
\[
II \geq M \| g - \bar{\eta}_n \|^2 + 2\lambda_n J_q(g - \bar{\eta}_n)
\]
\[
\geq \frac{1}{2} (M \wedge 2) \left( \| g - \bar{\eta}_n \|^2 + \lambda_n^{1/2} J_q^{1/2}(g - \bar{\eta}_n) \right)^2 = \frac{1}{2} (M \wedge 2) s^2.
\]
Therefore, by taking a sufficient large \( a \),
\[
I + II \geq \frac{1}{2} (M \wedge 2) s^2 - O_P \left( \frac{s^2}{a} \right) > 0.
\]
Thus we have proved \( (25) \). This completes the proof of the theorem. \( \square \)
5. Useful lemmas for verifying the conditions of the master theorems. This section develops three lemmas that provide sufficient conditions for Conditions 3.1 and 3.2(i)(ii), respectively.

**Lemma 5.1.** Suppose \(\|h_1\|_\infty \leq C\) for some constant \(C > 0\). If there are constant \(B > 0\) and constants \(M_1, M_2 > 0\) such that

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

whenever \(\|h_2\|_\infty \leq B\), then Condition 3.1 holds if \(\|\eta_0\| \leq C\).

This is Lemma A.1 of Huang (2001), which is proved easily by a Taylor expansion at the maximal point of the expected log-likelihood and noticing that the first order term is zero. As we will show later in this paper that (29) can be verified easily in various contexts.

**Lemma 5.2.** If there exists a constant \(M\) such that \(\text{Var}\{\hat{l}[\tilde{\eta}_n; h]\} \leq M\) for any \(h\) satisfying \(\|h\|^2 = 1\), then Condition 3.2(i) holds.

This lemma is a generalization of Lemma A.2 of Huang (2001), which gives a similar result for polynomial spline estimators.

**Proof of Lemma 5.2.** Consider an orthonormal basis \(\{\psi_k, k = 1, \ldots, N_n\}\) of \(G_n\). We have \(N_n \approx \delta_n^{-1}\). Any \(g \in G_n\) can be represented by this basis as \(g = \sum_k g_k \psi_k\), where \(g_k = \langle g, \psi_k \rangle\). It follows that \(\hat{l}[\tilde{\eta}_n; g] = \sum_k g_k \hat{l}[\tilde{\eta}_n; \psi_k]\).

By the Cauchy–Schwarz inequality and \(\|g\|^2 = \sum_k g_k^2\),

\[
\frac{|(E_n - E)\hat{l}[\tilde{\eta}_n; g]|^2}{\|g\|^2 + \lambda_n J_q(g)} \leq \frac{|(E_n - E)\hat{l}[\tilde{\eta}_n; g]|^2}{\|g\|^2} \leq \sum_k \{(E_n - E)\hat{l}[\tilde{\eta}_n; \psi_k]\}^2
\]

Since \(\|\psi_k\| = 1\), by the assumption of the lemma, the expectation of the right hand side of the above is bounded by \(\sum_k \{M/n\} \leq M/(n\delta_n)\). On the other hand, take the eigen decomposition \(g = \sum_\nu g_\nu \phi_\nu\). We have \(\hat{l}[\tilde{\eta}_n; g] = \sum_\nu g_\nu \hat{l}[\tilde{\eta}_n; \phi_\nu]\). By the Cauchy–Schwartz inequality and

\[
\|g\|^2 + \lambda_n J_q(g) = \sum_\nu g_\nu^2 (1 + \lambda_n \rho_\nu),
\]

we have that

\[
\frac{|(E_n - E)\hat{l}[\tilde{\eta}_n; g]|^2}{\|g\|^2 + \lambda_n J_q(g)} \leq \sum_\nu \{(E_n - E)\hat{l}[\tilde{\eta}_n; \phi_\nu]\}^2 / \left(1 + \lambda_n \rho_\nu\right).
\]
Since $\|\phi_\nu\| = 1$, by the assumption of this lemma and Proposition 2.5, the expectation of the right hand side of the above is bounded by

$$\frac{M}{n} \sum_\nu \frac{1}{1 + \lambda_n \rho_\nu} = O\left(\frac{1}{n^{1/(2q)}}\right).$$

The conclusion now follows from (30)–(31) and the Markov inequality.

**Lemma 5.3.** The following provides a sufficient condition for Condition 3.2(ii):
(i) $\|\bar{\eta}_n\|_\infty = O(1)$;
(ii) For $g \in \mathcal{G}_n$, $\ell(\bar{\eta}_n + \alpha g)$ as a function of $\alpha$ is twice continuously differentiable; moreover, there are constants $B > 0$ and $M > 0$ such that

$$\frac{d^2}{d\alpha^2} \ell(\bar{\eta}_n + \alpha g) \leq -M\|g\|^2, \quad 0 \leq \alpha \leq 1,$$

holds for $g \in \mathcal{G}_n$ with $\|g\|_\infty \leq B$, with probability tending to one as $n \to \infty$.

When using this lemma, we only need to verify Part (ii) of the condition, since Part (i) is a consequence of Theorem 3.1. Part (ii) of the condition has been used in Huang (2001) when studying rates of convergence of polynomial spline estimators.

**Proof of Lemma 5.3.** Since

$$\left. \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \right|_{\alpha=1} - \left. \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \right|_{\alpha=0} = \int_0^1 \frac{d^2}{d\alpha^2} \ell(\bar{\eta}_n + \alpha g) d\alpha,$$

the result is straightforward.

**6. Application I: regression.** Consider the problem of estimating the conditional mean function $\eta_0(x) = E(\mathcal{Y} | X = x)$ based on an i.i.d. sample of $\mathbf{W} = (X, Y)$, denoted as $\mathbf{W}_i = (X_i, Y_i), i = 1, \ldots, n$. For a candidate function $h$ of the unknown function $\eta_0$, define the “log-likelihood” functional as

$$\ell(h; \mathbf{W}_1, \ldots, \mathbf{W}_n) = -\frac{1}{n} \sum_{i=1}^n \{Y_i - h(X_i)\}^2.$$

This can be interpreted as a (conditional) log-likelihood (up to a scale factor) when the conditional distribution of $y_i$ given $x_i$ is Gaussian or a pseudo log-likelihood without the distribution assumption.

We verify conditions used in the master theorems under the following primitive assumptions.
Assumption (REG). (i) The function \( \eta_0 \) is bounded on \( U \).

(ii) There is a constant \( D > 0 \) such that \( \text{Var}(Y|X = x) \leq D \) for all \( x \).

(iii) The distribution of \( X \) is absolutely continuous and its density function is bounded away from zero and infinity on \( U \), that is, there exist constants \( C_1, C_2 > 0 \) such that

\[
C_1 \leq f_X(x) \leq C_2, \quad \text{for} \; x \in U.
\]

The expected log-likelihood is

\[
\Lambda(\eta) = -E[(Y_i - h(X_i))^2].
\]

Define the empirical and theoretical norms as in Section 2.1 with the weight function being \( w(x) \equiv 1 \). It is easy to see that

\[
\frac{d^2}{d\alpha^2} \Lambda(h_1 + \alpha h_2) = -2\|h_2\|^2,
\]

and thus (29) holds with \( M_1 = M_2 = 2 \). Condition 3.1 then follows from Lemma 5.1.

Note that

\[
\frac{d}{d\alpha} \ell(\bar{\eta}_n; h)(W_1) = \{\bar{\eta}_n(X_1) - Y_1\}h(X_1).
\]

Since we apply Theorem 3.2 after we apply Theorem 3.1, we can use the conclusion of Theorem 3.1 and assume that \( \|\bar{\eta}_n\|_\infty \leq M \) for some constant \( M > 0 \) when \( n \) is large enough. Suppose \( \|h\|^2 = 1 \). Let \( \epsilon_1 = Y_1 - \eta_0(X_1) \). We have that

\[
\text{Var}[\{\bar{\eta}_n(X_1) - Y_1\}h(X_1)] \leq E[\{\bar{\eta}_n(X_1) - Y_1\}^2 h(X_1)^2]
\]

\[
= E[\{\bar{\eta}_n(X_1) - \eta_0(X_1)\}^2 h(X_1)^2] + E[\epsilon_1^2 h(X_1)^2]
\]

\[
\leq \|\bar{\eta}_n - \eta_0\|_\infty^2 + D \leq (M + \|\eta_0\|_\infty)^2 + D,
\]

which is the condition needed for applying Lemma 5.2. Condition 3.2(i) then follows from Lemma 5.2.

Finally,

\[
\frac{d^2}{d\alpha^2} \ell(\bar{\eta}_n + \alpha g; W_1, \ldots, W_n) = -\frac{2}{n} \sum_{i=1}^{n} g^2(X_i) = -2\|g\|^2_n.
\]

Proposition 2.3 implies that Part (ii) of the condition in Lemma 5.3 holds if \( \lim_n N_n \log(n)/n = 0 \), and thus Condition 3.2(ii) holds according to this lemma.

Verification of conditions is complete.
7. Application II: quantile regression. Fixing $\tau \in (0, 1)$, let $\eta_0(x)$ be the $\tau$-th quantile of the conditional distribution of $Y|X = x$. We would like to estimate $\eta_0$ based on an i.i.d. sample of $W = (X, Y)$, denoted as $W_i = (X_i, Y_i), i = 1, \ldots, n$. For a candidate function $h$ of the unknown function $\eta_0$, define the “log-likelihood” functional as

$$
\ell(h; W_1, \ldots, W_n) = -\frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - h(X_i)),
$$

where $\rho_\tau(u) = (\tau - 1_{u<0})u$ is the check function for quantile at the level $\tau$. This can be interpreted as a pseudo log-likelihood without making a distribution assumption on the conditional distribution of $Y$ given $X$. The quantile function $\eta_0$ maximizes the expected log-likelihood functional

$$
\Lambda(h) = -E\{\rho_\tau(Y_i - h(X_i))\}.
$$

We verify conditions used in the master theorems under the following primitive assumptions.

**Assumption (QR).** (i) The function $\eta_0$ is bounded on $U$.

(ii) There are constants $B > 0$ and $M_1, M_2 > 0$ such that for any interval $A \subset [-B, B]$,

$$
M_1|A| \leq P(Y - \eta_0(x) \in A|X = x) \leq M_2|A|,
$$

where $|A|$ denotes the length of interval $A$.

(iii) The distribution of $X$ is absolutely continuous and its density function is bounded away from zero and infinity on $U$, that is, there exist constants $C_1, C_2 > 0$ such that

$$
C_1 \leq f_X(x) \leq C_2, \quad \text{for } x \in U.
$$

Similar to the regression case, define the empirical and theoretical norms as in Section 2.1 with the weight function being $w(x) \equiv 1$. Using the Knight identity (Knight, 1998),

$$
\rho_\tau(u - v) - \rho_\tau(u) = v\{1_{(u \leq 0)} - \tau\} + \int_0^v \{1_{(u \leq s)} - 1_{(u \leq 0)}\}ds,
$$

we obtain

$$
\Lambda(\eta_0 + h) - \Lambda(\eta_0) = -E\{\rho_\tau(Y - \eta_0(X) - h(X)) - \rho_\tau(Y - \eta_0(X))\}
$$

$$
= -E[h(X)\{1_{(Y - \eta_0(X) \leq 0)} - \tau\}
$$

$$
+ \int_0^{h(X)} \{1_{(Y - \eta_0(X) \leq s)} - 1_{(Y - \eta_0(X) \leq 0)}\}ds].
$$
Note the first part of the expectation is zero by the definition of \( \eta_0 \). By conditioning and then changing the order of integration, we have

\[
\Lambda(\eta_0 + h) - \Lambda(\eta_0) = -E \left[ \int_0^h E \{1_{Y - \eta_0(X) \leq s} - 1_{Y - \eta_0(X) \leq 0} \mid X \} \, ds \right] \\
= -E \left[ \int_0^h \text{sgn}(s) P\{Y - \eta_0(X)\text{ is between 0 and } s \mid X\} \, ds \right].
\]

If \( \|h\|_\infty \leq B \), by Assumption QR(ii), the above quantity is between \(-M_2\|h\|^2/2\) and \(-M_1\|h\|^2/2\). This verifies Condition 3.1.

Define \( \psi(u) = \tau - 1 \) for \( u < 0 \), and \( \psi(u) = \tau \) for \( u \geq 0 \). Then \( \psi(u) \) is the derivative of \( \rho_x(u) \) when \( u \neq 0 \) and the right derivative when \( u = 0 \). The directional derivative at \( \bar{\eta}_n \) along the direction of \( g \) is

\[
\bar{l} \{\bar{\eta}; g\}(\mathbf{W}_1) = g(X_1)\psi(Y_1 - \bar{\eta}(X_1)).
\]

Since \( |\psi(u)| \leq 1 \), \( \text{Var} \{\bar{l}[\bar{\eta}; g]\} \} \leq \|g\|^2 \). Condition 3.2(i) then follows from Lemma 5.2.

It remains to verify Condition 3.2(ii). Note that

\[
\frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \bigg|_{\alpha=1} + \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \bigg|_{\alpha=0} = \frac{1}{n} \sum_{i=1}^n g(X_i)\{\psi(Y_i - \bar{\eta}_n(X_i) - g(X_i)) - \psi(Y_i - \bar{\eta}_n(X_i))\}.
\]  

(34)

Let \( \epsilon_i = Y_i - \eta_0(X_i) \). Then \( Y_i - \bar{\eta}_n(X_i) = \epsilon_i - \{\bar{\eta}_n(X_i) - \eta_0(X_i)\} \). By the definition of \( \psi(\cdot) \), the difference \( \psi(Y_i - \bar{\eta}_n(X_i) - g(X_i)) - \psi(Y_i - \bar{\eta}_n(X_i)) \) is non-zero only when zero is between

\[
Y_i - \bar{\eta}_n(X_i) - g(X_i) = \epsilon_i - \{\bar{\eta}_n(X_i) - \eta_0(X_i)\} - g(X_i)
\]

and

\[
Y_i - \bar{\eta}_n(X_i) = \epsilon_i - \{\bar{\eta}_n(X_i) - \eta_0(X_i)\},
\]

or equivalently, when \( \epsilon_i \) is between \( \bar{\eta}_n(X_i) - \eta_0(X_i) - g(X_i) \) and \( \bar{\eta}_n(X_i) - \eta_0(X_i) \), and the value is \(-\text{sgn}\{g(X_i)\}\). Therefore,

\[
-\frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \bigg|_{\alpha=1} + \frac{d}{d\alpha} \ell(\bar{\eta}_n + \alpha g) \bigg|_{\alpha=0} = \frac{1}{n} \sum_{i=1}^n |g(X_i)| I_i
\]

(35)

where

\[
I_i = I(\epsilon_i \text{ is between } \bar{\eta}_n(X_i) - \eta_0(X_i) - g(X_i) \text{ and } \bar{\eta}_n(X_i) - \eta_0(X_i)).
\]
Applying the Hoeffding inequality, we obtain

\[
P\left(|\frac{1}{n}\sum_{i=1}^{n} |g(X_i)|\{I_i - E(I_i|X_i)\}| \geq t \, |X_i, i = 1, \ldots, n\right) \leq 2 \exp\left(-\frac{nt^2}{2\|g\|_n^2}\right).
\]

It follows that

\[
\frac{1}{n}\sum_{i=1}^{n} |g(X_i)|I_i - \frac{1}{n}\sum_{i=1}^{n} |g(X_i)|E(I_i|X_i) = \|g\|_n o\left(\sqrt{\log n}\right).
\]

We may focus on \(g \in G\) satisfying \(\|g\|_{\infty} \leq B/2\), where \(B\) is the constant in Assumption QR(ii). Since \(\|\bar{\eta}_n - \eta_0\|_{\infty} = o(1)\), both \(\bar{\eta}_n(X_i) - \eta_0(X_i) - g(X_i)\) and \(\bar{\eta}_n(X_i) - \eta_0(X_i)\) are in the interval \([-B, B]\). Using the assumption, we have that \(P(I_i|X_i) \geq M_1|g(X_i)|\). Thus,

\[
\frac{1}{n}\sum_{i=1}^{n} |g(X_i)|E(I_i|X_i) \geq M_1\|g\|_n^2.
\]

Combining (35)–(37) and using the equivalence between the empirical and theoretical norms (i.e., Proposition 2.3), we obtain the desired validity of Condition 3.2(ii).

Acknowledgement. We would like to thank the Associate Editor and reviewers for many valuable comments which helped us substantially improve upon earlier versions of the paper.

References.

Claeskens, G., Krivobokova, T. and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. Biometrika 96 529–544.

DeVore, R. A. and Lorentz, G. G. (1993). Constructive Approximation. Springer Science & Business Media.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). Statistical Science 11 89–102.

Gu, C. (2013). Smoothing Spline ANOVA Models. 2nd ed. Springer.

Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. Biometrika 92 105–118.

Hansen, M. H. (1994). Extended Linear Models, Multivariate Splines, and ANOVA. PhD Thesis, University of California, Berkeley.

Holland, A. D. (2017). Penalized Spline Estimation in the Partially Linear Model. Journal of Multivariate Analysis 153 211–235.

Huang, J. Z. (1998a). Projection estimation in multiple regression with application to functional ANOVA models. The Annals of Statistics 26 242–272.

Huang, J. Z. (1998b). Functional ANOVA models for generalized regression. Journal of Multivariate Analysis 67 49–71.
Huang, J. Z. (2001). Concave extended linear modeling: a theoretical synthesis. *Statistica Sinica* **11** 173–198.

Huang, J. Z. (2003). Asymptotics for polynomial spline regression under weak conditions. *Statistics & probability letters* **65** 207–216.

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

Knight, K. (1998). Limiting distributions for $L_1$ regression estimators under general conditions. *Annals of Statistics* **26** 755–770.

Lai, M.-J. and Wang, L. (2013). Bivariate penalized splines for regression. *Statistica Sinica* **23** 1399–1417.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika* **95** 415–436.

O’Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science* **1** 502–518.

O’Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing* **9** 363–379.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. Wiley New York.

Schwarz, K. and Krivobokova, T. (2016). A unified framework for spline estimators. *Biometrika* **103** 121–131.

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

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

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics* **22** 118–171.

Utreras, F. (1981). Optimal smoothing of noisy data using spline functions. *SIAM Journal on Scientific and Statistical Computing* **2** 349–362.

Wahba, G. (1990). *Spline Models for Observational Data*. CMBS-NSF Regional Conference Series in Applied Mathematics. SIAM.

Wang, X., Shen, J., Ruppert, D. et al. (2011). On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics* **5** 1–17.

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

Xiao, L. (2019a). Asymptotic theory of penalized splines. *Electron. J. Statist.* **13** 747–794.

Xiao, L. (2019b). Asymptotics of bivariate penalised splines. *Journal of Nonparametric Statistics* **31** 289-314.

Xiao, L., Li, Y., and Ruppert, D. (2013). Fast bivariate P-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 577-599.