Biometrika (2023), **xx**, x, *p.* 1
Printed in Great Britain

NAIVE PENALIZED SPLINE DERIVATIVE ESTIMATORS ACHIEVE OPTIMAL L_2 RATES OF CONVERGENCE

BY B. A. BOASIAKO
Department of Mathematics and Statistics, University of Massachusetts Amherst,
Amherst Massachusetts 01003, U.S.A.
bantwiboasia@umass.edu

J. W. STAUDENMAYER
Department of Mathematics and Statistics, University of Massachusetts Amherst,
Amherst Massachusetts 01003, U.S.A.
jstauden@umass.edu

SUMMARY

This paper studies the asymptotic behavior of penalized spline estimators of derivatives. In particular, we show that simply fitting a penalized spline with a smoothing parameter that was chosen as if the function were of interest and differentiating the estimated spline with the estimated parameters held fixed results in an estimate of the derivative that achieves the optimal L_2 rate of convergence.

Some key words: Derivative Estimation, Penalized Spline, L_2 Convergence, Nonparametric Smoothing

1. INTRODUCTION

We consider the situation where data $\{x_i, y_i\}_{i=1}^n$ are sampled from the model:

$$y_i = f(x_i) + \varepsilon_i, \ \ \forall i = 1, \ldots, n \tag{1}$$

with $f \in \mathcal{C}^p(\mathcal{K})$, the space of functions with p continuous derivatives over $\mathcal{K} = [0, 1]$, the x_i's, for $x_i \in \mathcal{K}$, are either random or deterministic, and ε_i's are independent and identically distributed random error terms with $E(\varepsilon_i) = 0$ and $var(\varepsilon_i) = \sigma^2$. There are many cases when it is of interest to estimate some derivative of the mean regression function f, with nonparametric assumptions on the functional form of f. For example, in human growth studies, the first derivative of the function relating height and age indicates the speed of growth (Müller (1988); Ramsay & Silverman (2002)). Additionally, Chaudhuri & Marron (1999) apply derivative estimation in the development of a visual mechanism for studying curve structures, and Park & Kang (2008) compare regression curves using those structures. In economics, derivatives are used to calculate the marginal propensity to consume, which measures the effect of changes in disposable income on personal consumption (Fisher et al. (2020)). In addition, average derivatives of mean regression functions are used to empirically validate the so-called "law of demand" (Härdle et al. (1992)). It is sufficient for a random matrix composed of average derivatives to be positive definite for the law of demand to hold (Hildenbrand (1989)). In nonparametric regression itself, estimates of derivatives of the true function, f, are used in plug-in bandwidth selection techniques such as in local polynomial regression (Ruppert et al. (1995)) and to construct confidence bands for nonparametric estimators (Eubank & Speckman (1993)).

Previous work has taken three major approaches to estimate derivatives of functions nonparametrically: local polynomial regression, empirical derivatives, and spline-based methods. In local polynomial regression, a derivative of $f(x)$ can be estimated using a coefficient of the fitted local polynomial at x (Fan & Gijbels, 1996, p 22). Empirical methods generally transform the data and smooth difference-based estimates of derivatives. Earlier works include Müller et al. (1987), which used Kernel-based approaches to estimate the derivatives of the mean regression function while utilizing difference quotients to identify the best kernel bandwidth via cross-validation. More recently, De Brabanter et al. (2013) used symmetric difference quotients to estimate derivatives of mean regression functions and showed that their approach improved the asymptotic order of the variance. Spline-based methods use the fact that splines are piecewise polynomials. As a result, differentiating the basis function with respect to the covariate gives a basis function for the derivative, and an estimate of that function then can be obtained using a subset of the estimated coefficients (de Boor (1978), Eilers & Marx (1996, 2010)).

While it is straightforward to compute nonparametric derivative estimates, the challenge is that those estimates also require some sort of regularization to balance estimation bias and overfitting; and methods to choose that regularization are usually designed for estimating the function itself, not derivatives (Ruppert et al. (2003); Eilers & Marx (1996)). Several authors have designed methodologies to address that problem (e.g. Charnigo et al. (2011); Simpkin & Newell (2013)), but it has also been suggested that methods that choose the amount of smoothing for the derivatives as if the function were of interest often work well in practice (Ruppert et al. (2003) section 6.8.2, Craven & Wahba (1978)). We call such methods naive. In this paper, we add evidence to that debate by exploring the asymptotic behavior of naive nonparametric derivative estimators, focusing on penalized splines.

The past few decades have seen some progress in related areas. In local polynomial regression, results from Ruppert & Wand (1994) can be used to show that when a bandwidth is chosen to minimize the integrated mean squared error (IMSE) of a p^{th} degree polynomial estimate of f, and that bandwidth is used to estimate $f^{(r)}$ ($p \geq r$, i.e. a naive estimator), then the derivative estimate's IMSE converges at an optimal rate if r is even. Otherwise, the naive bandwidth over- or under-smooths. In particular, a naive estimator of the first derivative using cubic local polynomials under-smooths. A derivation of this is given in the appendix. We have not found this result in the literature.

De Brabanter et al. (2013) showed that the asymptotic order of the bias of empirical estimators does not depend on the order of the derivative being estimated, and they employ a method by De Brabanter et al. (2011) to address correlations that result from creating the empirical dataset for the derivative. Dai et al. (2016) generalized those results by considering linear combinations of observations to better fit both interior and boundary points. They demonstrated that their method achieves optimal rates of convergence (Stone (1982)).

The asymptotic properties of two types of spline-based derivative estimators have been considered too. Wahba & Wang (1990) studied smoothing splines and found that the optimal smoothing parameter depends on the order of the derivative being estimated. In contrast, Zhou & Wolfe (2000) analyzed the asymptotics of regression spline-based derivative estimators where the number of knots increases with the sample size. They showed that the MSE goes to zero at the optimal rate (Stone (1982)), and the required rate of increase in the number of knots does not depend on the order of the derivative.

Somewhat surprisingly though, comparable results about penalized spline estimators of derivatives do not seem to exist. Building on work on the asymptotics of penalized spline estimators of functions (Xiao (2019)), we derive the apparently new result that naive methods to estimate derivatives with penalized splines achieve optimal global L_2 rates of convergence (Stone (1982)).

2. Penalized Splines & the Naive Derivative Estimator

2.1. Splines

Splines provide a flexible mechanism to estimate derivatives of the mean regression function f, and in the case of estimating the function itself, they can achieve the best possible rates of convergence (Xiao (2019); Zhou et al. (1998); Stone (1982)). A spline is a piece-wise polynomial with continuity conditions at the points where the pieces join together (called knots).
More specifically, for $q \geq 2$, we let

$$\boldsymbol{\mathcal{S}}(q, \underline{\mathrm{t}}) = \left\{ s \in \mathcal{C}^{q-2}(\mathcal{K}) : s \text{ is a } q\text{-order polynomial on each } [t_i, t_{i+1}] \right\}$$

be a space of $q-$order splines over $\mathcal{K} = [0, 1]$ with knot locations $\underline{\mathrm{t}} = (t_0, \dots, t_{K+1})$ where $t_0 = 0$, $t_{K+1} = 1$ and $t_i < t_j \ \forall_{i<j}$. For $q = 1$, $\boldsymbol{\mathcal{S}}(q, \underline{\mathrm{t}})$ consists of step functions with jumps at the knots.

This space has a number of equivalent bases and one notable for having stable numerical properties is the *B-Spline* basis (de Boor (1978); Ruppert et al. (2003); Schumaker (2007)).

de Boor (1978) defines the q^{th} order B-spline basis function $B_{j,q}(x)$ over the knot locations $\underline{\mathrm{t}}$ through a recurrence relation. Eilers & Marx (2010) show that when the distance between the knots is constant, $B_{j,q}(x)$ reduces to

$$B_{j,q}(x) = \frac{(-1)^q \Delta^q (x - t_j)_+^{q-1}}{(q-1)! h^{q-1}}$$

where Δ is the backward difference operator ($\Delta t_j = t_j - t_{j-1}$), and h is the common distance between the knots. Observe that $B_{j,q}(x)$, in this case, is a rescaled q-order difference of truncated polynomials. To get a complete set of B-Spline basis, we need $2q$ extra knots with q knots on each side of $[0, 1]$. This is referred to as the expanded basis (Eilers & Marx (2010)).

Without losing generality, we will assume a B-Spline basis for $\boldsymbol{\mathcal{S}}(q, \underline{\mathrm{t}})$ for the rest of this paper. We refer the reader to de Boor (1978); Schumaker (2007); Eilers & Marx (1996) for an introduction to B-Splines, and Eilers & Marx (2010) for how the B-Spline basis compares to the Truncated Polynomial Functions (TPF) on metrics including fit quality, numerical stability, and multidimensional smoothing.

2.2. Penalized Splines & the Naive Estimator

Penalized splines are often viewed as a compromise between regression and smoothing splines because they combine penalization and low rank bases to achieve computational efficiency. They vary slightly based on the basis functions used and the object of penalization. For example, P-Splines (Eilers & Marx (1996)) use B-Spline basis functions and penalize differences of the coefficients to a specific order. In this section, we will focus on P-Splines. Our later results hold for the general penalized spline estimator defined by Xiao (2019).

A P-Spline estimator of f in (1) based on an iid sample of size n finds a *spline* function $g(x) = B(x)\underline{\boldsymbol{\alpha}}$, that minimizes:

$$Q(\underline{\boldsymbol{\alpha}}, \lambda_n) = \frac{1}{n} \sum_{i=1}^{n} (y_i - B(x_i)\underline{\boldsymbol{\alpha}})^2 + \lambda_n \underline{\boldsymbol{\alpha}}^T \boldsymbol{P_m} \underline{\boldsymbol{\alpha}} \tag{2}$$

where, $\underline{\boldsymbol{\alpha}} = (\alpha_1, \ldots, \alpha_{K+q})$ is a vector of coefficients, and $B(x_i) = [B_{1,q}(x_i), \ldots, B_{K+q,q}(x_i)] \in \mathbb{R}^{K+q}$ is a vector of basis functions at x_i, for $i = 1, \ldots, n$. The penalty matrix $\boldsymbol{P_m} = \boldsymbol{D_m^T D_m} \in \mathbb{R}^{(K+q)\times(K+q)}$ where $\boldsymbol{D_m}\underline{\boldsymbol{\alpha}} = \Delta^m \underline{\boldsymbol{\alpha}}$ is a vector of m^{th} order differences of $\underline{\boldsymbol{\alpha}}$. Finally, $\lambda_n \geq 0$ is the smoothing parameter and needs to be chosen. Three prevalent methods for choosing λ_n are generalized cross-validation (GCV), maximum likelihood (ML), and restricted (or residual) maximum likelihood (REML); we refer the reader to Wood (2017) chapter 4 and Ruppert et al. (2003) Chapters 4 and 5 for details.

Minimizing (2) with respect to $\underline{\boldsymbol{\alpha}}$ gives $\hat{\underline{\boldsymbol{\alpha}}} = \left(\boldsymbol{B}^T\boldsymbol{B}/n + \lambda_n \boldsymbol{P_m}\right)^{-1} \boldsymbol{B}^T y/n$ which results in $\hat{f}(x) = B(x)\hat{\underline{\boldsymbol{\alpha}}}$. Here, $\boldsymbol{B} = [B(x_1), B(x_2), \ldots, B(x_n)]^T \in \mathbb{R}^{n \times K+q}$. From $\hat{f}$, we can derive the naive estimator of the r^{th} derivative of f as follows:

$$\begin{aligned}\hat{f}^{(r)}(x) &= \frac{d^{(r)}}{dx} B(x)\hat{\underline{\boldsymbol{\alpha}}} \\ &= \frac{d^{(r)}}{dx}\left(\sum_{j=1}^{K+q} \hat{\alpha}_j B_{j,q}(x)\right) \\ &= \sum_{j=1}^{K+q-r} \hat{\alpha}_j^{(r)} B_{j,q-r}(x) \end{aligned} \tag{3}$$

where $\hat{\alpha}_j^{(r)} = (q-r)\frac{\left(\hat{\alpha}_{j+1}^{(r-1)} - \hat{\alpha}_j^{(r-1)}\right)}{t_j - t_{j-q+r}}$, with $\hat{\alpha}_j^{(0)} = \hat{\alpha}_j$ for $1 \leq j \leq K+q-r$, and $r = 1, \ldots, q-2$. (de Boor (1978); Zhou & Wolfe (2000)).

Xiao (2019) showed that under some conditions on the distribution of the knots and λ_n, $\hat{f}$ achieves the optimal L_2 rate of convergence (Stone (1982)) to the true f but they do not discuss derivative estimators. We extend this result to show that under same conditions that do not depend of the order of the derivative, the naive derivative estimator $\hat{f}^{(r)}$ of $f^{(r)}$ achieves optimal L_2 rates of convergence.

3. MAIN RESULTS

In this section, we provide our main result in Theorem 1 and remark on how this result relates to regression and smoothing splines. The findings in this section apply to the general penalized spline estimator as defined by Xiao (2019). This general estimator is based on the realization that the various types of penalized splines differ mainly by their penalty matrices. However, the eigenvalues of the penalty matrices decay at similar rates, making their unified asymptotic study tractable. We refer the interested reader to a derivation of the decay rates of various penalty matrices in Xiao (2019).

3.1. Notation

We start by defining the following notations relating to norms and limits. For a real matrix A, $||A||_\infty = \max_i \sum_j |a_{ij}|$ is the largest row absolute sum. $||A||_2$ is the operator norm of A induced by the vector norm $||.||_2$. $||A||_F = \left\{tr(A^TA)\right\}^{1/2}$ is the Frobenius norm. For a real vector, $||\underline{\boldsymbol{a}}|| = \max_i |a_i|$. For a real-valued function $g(x)$ defined on $\mathcal{K} \subset \mathbb{R}$, $||g|| = \sup_{x\in\mathcal{K}} |g(x)|$

and $||g||^2_{L2} = \int_{x\in\mathcal{K}} \{g(x)\}^2\, dx$ is the squared L_2-norm of g. For two real sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n \sim b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$.

3.2. Assumptions

Next, we state assumptions on the knot placement and penalty matrix. We note that these assumptions are the same as those made in Xiao (2019) for the asymptotic analysis of estimates of functions rather than derivatives.

Assumption 1. $K = o(n)$

Assumption 2. $\max_{1\leq i\leq K} |h_{i+1} - h_i| = o(K^{-1})$, where $h_i = t_i - t_{i-1}$.

Assumption 3. $\frac{h}{\min_{1\leq i\leq K} h_i} \leq M$, where $h = \max_{1\leq i\leq K} h_i$ and $M > 0$ is some predetermined constant.

Assumption 4. For a deterministic design,

$$\sup_{x\in[0,1]} |Q_n(x) - Q(x)| = o(K^{-1})$$

where $Q_n(x)$ is the empirical CDF of x and $Q(x)$ is a distribution with continuously differentiable positive density $q(x)$.

Assumption 5. The penalty matrix $\boldsymbol{P_m}$ is a banded symmetric positive semi-definite square matrix with a finite bandwidth and $||\boldsymbol{P_m}||_2 = O(h^{1-2m})$. This assumption is similar to Assumption 3 of Xiao (2019) where it is stated in terms of the eigenvalues of $\boldsymbol{P_m}$. This assumption is verifiable for P-Splines, O-Splines, and T-Splines. See Propositions 4.1 and 4.2 of Xiao (2019). Also, we assume $\underline{\boldsymbol{\beta}}^T \boldsymbol{P_m} \underline{\boldsymbol{\beta}} = O(1)$ where $\underline{\boldsymbol{\beta}}$ is the coefficient vector for approximating f in 1 with the best approximating spline function $\boldsymbol{s_f}$ in $\boldsymbol{\mathcal{S}}(q, \underline{\mathrm{t}})$ (see Lemma A4).

Assumption 6. $\lambda_n = o(1)$.

Assumptions (2) and (3) are necessary conditions on the placements of the knots and also imply that $h \sim K^{-1}$. This ensures that $M^{-1} < Kh < M$ and is necessary for numerical computations (Zhou et al., 1998).

THEOREM 1. *Let the mean regression function in* (1) *be such that* $f \in \mathcal{C}^p(\mathcal{K})$. *Under Assumptions* (1) - (6) *above, and for* $m \leq \min(p, q)$*:*

$$\mathbb{E}\left(||\hat{f}^{(r)} - f^{(r)}||^2_{L_2}\right) = O\left(\frac{K_e}{n}\right) + O\left(K^{-2(q-r)}\right) + o(K^{-2(p-r)})$$
$$+O\{\min(\lambda_n^2 K^{2m+2r}, \lambda_n K^{2r})\}$$

where $K_e = \min\left\{K^{2r+1}, K^{2r}\lambda_n^{-1/2m}\right\}$ *and* $r = 1, \ldots, q-2$.

The proof of the theorem is given in the appendix.

3.3. ***Remarks***

Remark 1. The asymptotics of penalized splines are either similar to those of regression splines or smoothing splines depending on how fast the number of knots increases as the sample size increases (Claeskens et al., 2009; Xiao, 2019). This creates two scenarios: the small number of knots scenario with asymptotics similar to regression splines and the large number of knots scenario with asymptotics similar to smoothing splines. We explore the rates of convergence of the naive estimator under each of these scenarios in Remarks 1a and 1b below.

Remark 1a (Small number of knots scenario): Suppose the mean regression function is q-times continuously differentiable, where q is the order of the spline used to estimate f. Thus, $f \in$

$\mathcal{C}^q(\mathcal{K})$. Also suppose $\lambda_n K^{2m} = O(1)$, then

$$\begin{aligned}\mathbb{E}\left(||\hat{f}^{(r)} - f^{(r)}||^2_{L_2}\right) &= O\left(\frac{K_e}{n}\right) + O\left(K^{-2(q-r)}\right) + o(K^{-2(p-r)}) \\ &\quad + O\{\min(\lambda_n^2 K^{2m+2r}, \lambda_n K^{2r})\} \\ &= O\left(\frac{K^{2r+1}}{n}\right) + O\left(K^{-2(q-r)}\right) + O(\lambda_n^2 K^{2m+2r}).\end{aligned}$$

Choosing K such that $K \sim n^{1/(2q+1)}$ and $\lambda_n = O(n^{-(q+m)/(2q+1)})$, the estimator $\hat{f}^{(r)}$ of $f^{(r)}$ converges at the optimal L_2 rate of $n^{-(q-r)/(2q+1)}$. In the above, we have used the fact that $p = q$ and that $\min\left\{\lambda_n^2 K^{2m+2r}, \lambda_n K^{2r}\right\} = \lambda_n^2 K^{2m+2r}$, $K_e = K^{2r+1}$ for $\lambda_n K^{2m} = O(1)$. We note that the λ_n's rate of decrease does not depend on (r), the order of the derivative.

Remark 1b (Large number of knots scenario): Suppose $f \in \mathcal{C}^m(\mathcal{K})$, and there exists a sufficiently large constant, C, independent of K such that for $K \geq C^{1/2m}\lambda_n^{-1/2m} = C^{1/2m}n^{1/(2m+1)}$, with $m \leq q$, we have

$$\begin{aligned}\mathbb{E}\left(||\hat{f}^{(r)} - f^{(r)}||^2_{L_2}\right) &= O\left(\frac{K_e}{n}\right) + O\left(K^{-2(q-r)}\right) + o(K^{-2(p-r)}) \\ &\quad + O\{\min(\lambda_n^2 K^{2m+2r}, \lambda_n K^{2r})\} \\ &= O\left(\frac{K^{2r}\lambda_n^{-1/2m}}{n}\right) + O\left(K^{-2(q-r)}\right) + o\left(K^{-2(m-r)}\right) \\ &\quad + O(\lambda_n K^{2r}).\end{aligned}$$

Choosing λ_n such that $\lambda_n \sim n^{-2m/(2m+1)}$, the estimator $\hat{f}^{(r)}$ of $f^{(r)}$ converges at the optimal L_2 rate of $n^{-(m-r)/(2m+1)}$. Again, we note that the λ_n's rate of decrease does not depend on (r).

Remark 2. While the naive estimator of the derivative achieves an optimal rate of convergence, that does not mean that the naive approach is optimal in a finite sample. We compare the performance of the naive estimator to an "oracle estimator" that minimizes mean integrated squared error in Section 4.1.4.

Remark 3. The theorem is derived under conditions on the growth in the number of knots, the spacings between them, and the smoothing parameter (λ_n). Specific rates of growth for K and for λ_n in Remarks 1a and 1b led to optimal rates of convergence. That said, it is not clear whether standard ways of choosing smoothing parameters would lead to optimal rates of converged. This too is explored in Section 4.

4. SIMULATIONS

4.1. Overview

In this section, we present a simulation to assess the naive estimator's rate of convergence and its finite-sample performance. The simulation is divided into three parts. The first part examines the L_2 rates of convergence of the naive estimator when GCV and REML are used to choose the smoothing parameter. The second part of this section focuses on the finite sample performance of the naive estimator. We compared it to an "oracle" method that uses knowledge of the true function (or derivatives) to choose the optimal smoothing parameter. That "oracle" method is not a practical estimator, but it provides an upper bound benchmark for P-spline performance. Finally, the third part of this section compares the naive method to other derivative estimation methods in the literature.

Except where noted, we use the same mean regression function f as De Brabanter et al. (2013). We simulated data $\{x_i, y_i\}_{i=1}^n$ from the model:

$$Y_i = f(x_i) + \varepsilon_i, \ \forall\, i = 1, \ldots, n$$

where x_i's are a grid over $\mathcal{K} = [0, 1]$, ε_i's are iid with $\varepsilon_i \sim N(0, \sigma^2 = 0.1^2)$ and

$$f(x) = 32e^{-8(1-2x)^2}\,(1 - 2x) \tag{4}$$

Figure 1 shows the mean regression function in (4) and its first two derivatives. We use a range of sample sizes as shown in the results.

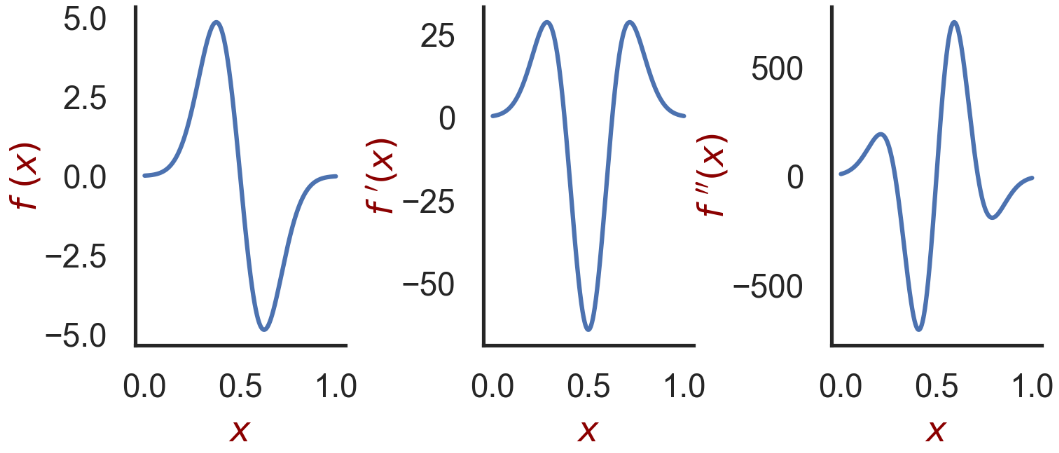


Fig. 1: Mean regression function with its first two derivatives.

As discussed in Xiao (2019), Claeskens et al. (2009), and our Remark (1), the asymptotics of the penalized spline estimator are similar to those of Regression Splines (small K scenario) or Smoothing Splines (large K scenario), depending on the rate at which the number of knots, K, increases with the sample size, n. In our simulation, we considered these two scenarios: when K increases slowly with n, and when K increases at a faster rate with n. For the slow K scenario we use $K \sim n^{1/(2p+1)}$, and K in the fast scenario is chosen such that $K \geq C^{1/p}\lambda_n^{-1/2p}$ for some large constant, C.

We investigated the L_2 rate of convergence for the first two derivatives of the mean regression function in (4) using a P-Spline with 2^{nd} ($m = 2$) order penalty (Eilers & Marx, 1996). Note that with $m = 2$, the equivalent kernel methodology (Silverman, 1984, Lemma 9.13 of Xiao et al., 2012) implies that the assumed differentiability of f is $p = 2m = 4$.

Stone (1982) provided optimal rates of convergence for non-parametric regression estimators. The optimal rate of convergence for a non-parametric estimator of the r^{th} derivative of $g : \mathbb{R}^d \to \mathbb{R}$ where $g \in \mathcal{C}^p$ is $n^{-(p-r)/(2p+d)}$, in our simulations, we have the optimal L_2 rate of convergence for estimating the r^{th} derivative of f as $n^{-(p-r)/(2p+d)} = n^{-(4-r)/(2\times4+1)} = n^{-(4-r)/9}$

4.2. *L_2 Convergence of the Naive Estimator*

Figure 2 illustrates the L_2 rate of the naive estimator when the smoothing parameter λ_n is chosen by the GCV approach. The naive estimator achieves the optimal L_2 rates of convergence

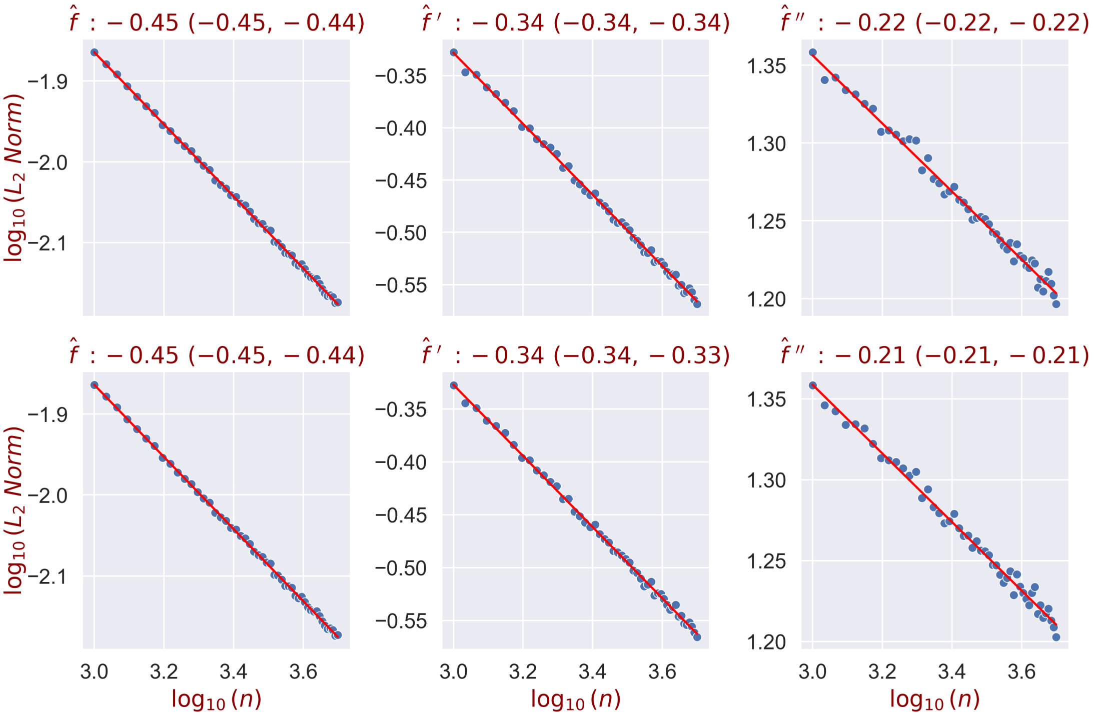


Fig. 2: L_2 convergence rates for f and its first two derivatives under two scenarios for increasing K with n. The top three figures show results for the slowly increasing K scenario while the bottom three show results for the fast increasing K scenario. The smoothing parameter λ_n is chosen by the GCV method.

for the mean regression function and its first two derivatives when GCV is used to choose the smoothing parameter, but it is slightly slower for REML. This deviation from the optimal rate using REML appears to worsen for higher derivatives. Also, we observed that the fast K scenario was overall slightly slower than the slow K scenario for REML. These results agree with known results in the literature for smoothing splines when estimating the mean regression function. For instance, Craven & Wahba (1978) showed that GCV achieves the optimal rate of convergence when used to choose the smoothing parameter in smoothing splines. However, Wahba (1985) found that Maximum Likelihood (ML) based methods may be slower than GCV for sufficiently smooth functions.

Table 1 below summarizes the rates of convergence of the naive estimator for estimating derivatives of the mean regression function in (4) under the various scenarios of the number of knots K as n increases.

4.3. Finite sample performance of naive estimator.

In this section, we compare the naive estimator to an “oracle” method that uses knowledge of the true form of the target (mean regression function or its derivatives) to choose the optimal amount of smoothing, which we did with a grid search. While this “oracle” is not an estimator, it shows the minimum loss when estimating the function in question with a penalized spline. GCV was used to choose the appropriate smoothing parameter for the various spline-based estimators in what follows.

Table 1: Summary of L_2 rates of convergence for estimating the mean regression function in (4) and its first two derivatives

λ_n **Method**	**Target**	**Optimal** L_2 **Rate**	**Slow** K	**Fast** K
	f	-0.44	$-0.45(-0.45, -0.44)$	$-0.45(-0.45, -0.44)$
GCV	f'	-0.33	$-0.34(-0.34, -0.34)$	$-0.34(-0.34, -0.33)$
	f''	-0.22	$-0.22(-0.22, -0.22)$	$-0.21(-0.21, -0.21)$
	f	-0.44	$-0.44(-0.44, -0.43)$	$-0.43(-0.44, -0.43)$
REML	f'	-0.33	$-0.32(-0.32, -0.31)$	$-0.31(-0.32, -0.31)$
	f''	-0.22	$-0.19(-0.19, -0.18)$	$-0.18(-0.18, -0.17)$

In Fig. 3 below, we show that the naive estimator corresponds to the median MSE in the Monte Carlo experiment. To summarize, we see that the naive estimator appears to accurately estimate both the true mean regression function (f) and its first derivative (f'). However, we observe some lack of fit around the boundaries of the second derivative, (f'').

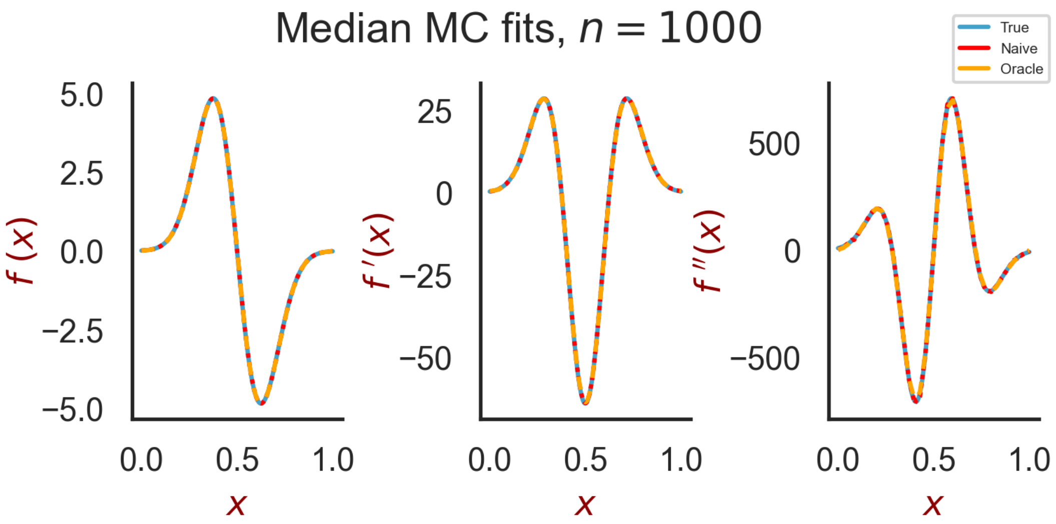


Fig. 3: Median Monte-Carlo fits of the mean regression function in (4) with its first two derivatives using the naive and oracle estimators.

Next, Fig. 4 compares the naive and oracle methods for the mean regression function in (4) and its first two derivatives across the two increasing K scenarios. Overall, in comparison to the oracle method, the naive estimator's finite sample performance degrades with increasing derivatives, with an average error difference (logarithmic scale) of about 0.5% for the mean regression function, 17% for its first derivative, and 29% for its second derivative. While the naive penalized spline derivative estimator is shown to converge at the optimal L_2 rate of convergence (Theorem 1), it may also have a higher mean squared error in finite samples, especially for higher derivatives. This is largely driven by a higher variance of the naive estimator, compared with the oracle method. We note that the results summarized in Fig. 4 are similar for the two increasing K scenarios.

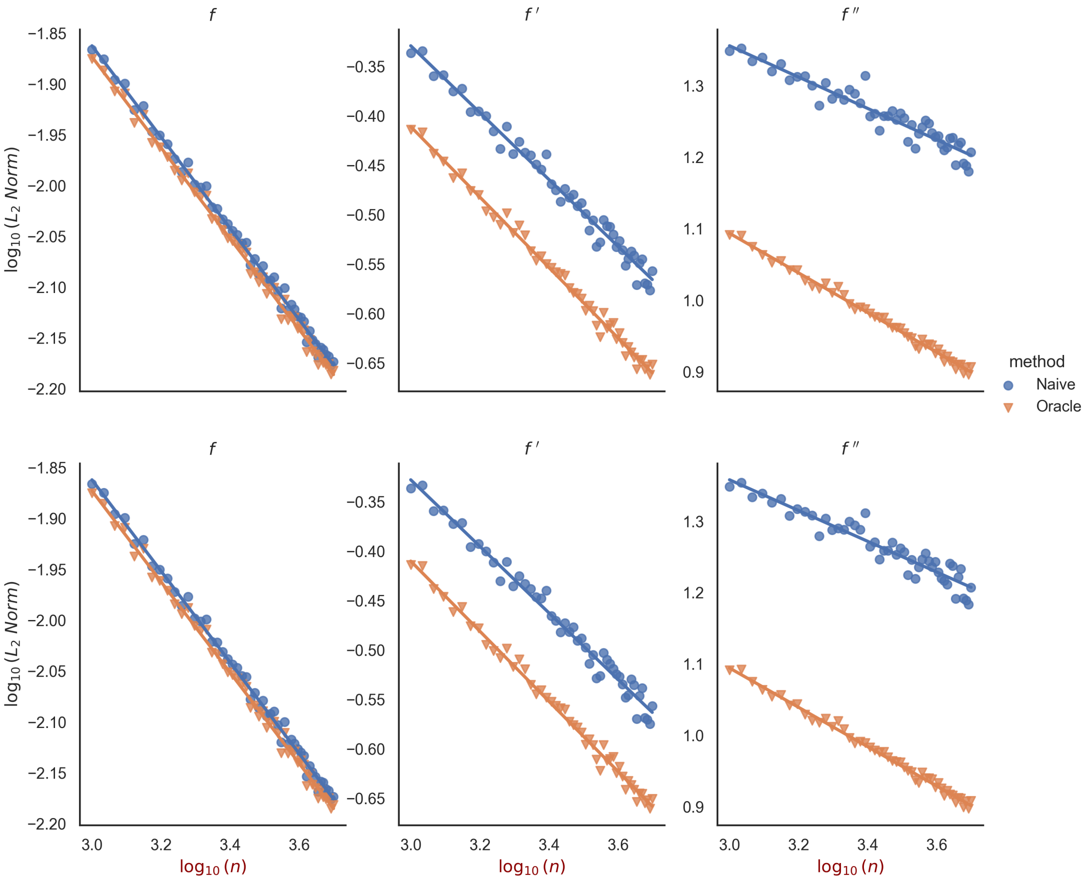


Fig. 4: L_2 convergence rates for f and its first two derivatives with two scenarios for increasing K with n and how they compare with their corresponding oracle estimators. Figures in the top row show results for slowly increasing K scenario while figures in the bottom row show results for the fast increasing K scenario. The smoothing parameter λ_n is chosen by the GCV method.

260 *4.4. Comparison with other methods*

In this section, we compare the finite sample MSE of the naive estimator to other derivative estimation methods in the literature. We considered the adaptive penalty penalized spline estimator by Simpkin & Newell (2013). We also used the linear combination method of Dai et al. (2016), but it consistently had higher MSE values and results are not shown.

265 We evaluated the methods using three mean regression functions from the literature (De Brabanter et al., 2013; Dai et al., 2016). As proxies for low, medium, and high noise scenarios, we considered noise levels that were 10 percent, 30 percent, and 60 percent of the range of each function. This was to understand how the methods compare at different levels of noise. The following are the three functions considered:

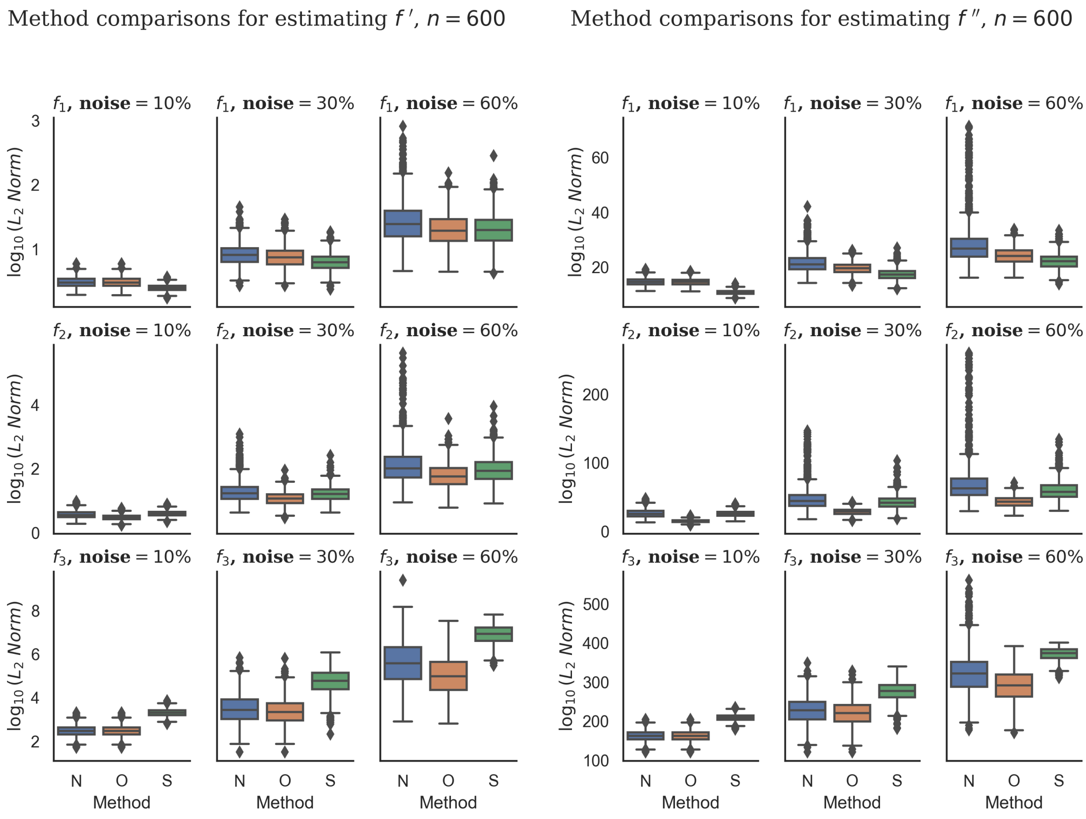
$$f_1(x) = \sin^2(2\pi x) + \log(4/3 + x) \quad \text{for} \quad x \in [-1, 1],$$

$$f_2(x) = 32e^{-8(1-2x)^2}(1-2x) \quad \text{for} \quad x \in [0,1],$$

and the doppler function

$$f_3(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x+0.05}\right) \quad \text{for} \quad x \in [0.25,1].$$

Figure 5 below shows the results for estimating the first (panel a) and second (panel b) derivatives of the three mean regression functions across the three noise levels. These results indicate that the adaptive penalty methods and the naive method often perform similarly, depending on the form of the function, the noise level, and the order of the derivative. We also note that the adaptive penalty method sometimes performs better than the oracle method. This is possible since the oracle method only finds the best P-splines estimate with the form of the penalty held constant.



(a)

(b)

Fig. 5: Comparing derivative estimation methods in reference to the oracle estimator across different functions and noise levels. Panel (a) shows results for estimating first derivatives of the mean regression functions f_1, f_2 and f_3 while Panel (b) shows results for estimating second derivatives. The first row is for f_1, second row is for f_2 and the last row is for f_3.

5. CONCLUSION

In this paper, we have shown that the naive penalized spline estimator of the r^{th} derivative of the mean regression function achieves the optimal L_2 rate of convergence (Stone, 1982) under standard assumptions on knot placement and the penalty matrix. This builds on the work by Xiao (2019) which derived the L_2 rate of convergence for estimating the mean regression function. As stated in Remark 1 and noted by others (Claeskens et al., 2009; Xiao, 2019), the rate at which the number of knots, K, increases with n gives rise to two scenarios: the fast K scenario, which is similar to smoothing spline asymptotics, and the slow K scenario, which is similar to regression spline asymptotics.

Using simulations, we investigated how two prevalent methods for choosing the smoothing parameter (GCV and REML) affect the L_2 convergence of the naive estimator. We found that, for both slow and fast K scenarios, the naive estimator achieves the optimal L_2 rate of convergence when GCV is used. For REML, the estimator did not quite achieve the optimal rate.

To access the finite sample performance of the naive estimator, we compared the MSEs of the estimator with an "oracle" method that uses information about the true function to be estimated to choose the P-spline's smoothing parameter. We found that, in finite samples, the naive estimator may have noticeably larger mean squared errors, especially for higher derivatives, but the estimates can still be quite visually similar. We found that the adaptive penalty penalized spline estimator by Simpkin & Newell (2013) performed similarly to the naive estimator.

APPENDIX

A.1. Local Polynomial Result

Suppose a local polynomial estimator of degree p is used to estimate $f^{(r)}(\cdot), r = 0, 1, 2, \ldots$. Let the naive bandwidth be the one that is optimal for estimating $f(\cdot) = f^{(0)}(\cdot)$. Asymptotic mean integrated squared error expressions can be found in Theorems 4.1 $(r = 0)$ and 4.2 $(r = 1, 2, \ldots)$ in Ruppert & Wand (1994). Minimizing those expressions with respect the bandwidths shows that when $(p - r)$ is odd the optimal bandwidth for estimating $f^{(r)}(\cdot)$ is $O\left(n^{\frac{-1}{2p+3}}\right)$, and when $(p - r)$ is even the optimal bandwidth is $O\left(n^{\frac{-1}{2p+5}}\right)$. As a result, we can conclude that the naive bandwidth results in an estimator that achieves the optimal rate when r is even.

A.2. Proof of Theorem

The proof proceeds in two steps. We first derive the L_2 rate of convergence for the bias of the naive estimator and then we derive that of the variance. The approach of the proof closely follows the proof for the L_2 rate of convergence of the mean regression function itself found in Xiao (2019) with a bit more clarity. We start by defining some terms to simplify the notation.

Let $G_{n,q} = \boldsymbol{B}^T\boldsymbol{B}/n$ and $H_n = G_{n,q} + \lambda_n \boldsymbol{P_m}$. To ease exposition, we follow Zhou & Wolfe (2000) and write $\hat{f}^{(r)}(x)$ as

$$\hat{f}^{(r)}(x) = B_{q-r}(x)D^{(r)}(G_{n,q} + \lambda_n \boldsymbol{P_m})^{-1}\boldsymbol{B}^T\boldsymbol{y}/n$$

where $B_{q-r}(x) \in \mathbb{R}^{K+q-r}$ is a vector of B-Spline basis functions of order $q - r$ and $D^{(r)}$ is defined as $D^{(r)} = M_r^T \times M_{r-1}^T \times \cdots \times M_1^T$ with

$$M_l = (q-1)\begin{bmatrix} \frac{-1}{t_1 - t_{1-q+l}} & 0 & 0 & \ldots & 0 \\ \frac{1}{t_1 - t_{1-q+l}} & \frac{-1}{t_2 - t_{2-q+l}} & 0 & \ldots & 0 \\ 0 & \frac{1}{t_2 - t_{2-q+l}} & \frac{-1}{t_3 - t_{3-q+l}} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \frac{1}{t_{K+q-l} - t_K} \end{bmatrix}$$

for $1 \leq l \leq r \leq q-2$.

Let $B_q^{(r)}(x) = B_{q-r}(x)D^{(r)}$, implying

$$\hat{f}^{(r)}(x) = B_q^{(r)}(x)\left(G_{n,q} + \lambda_n \boldsymbol{P_m}\right)^{-1} \boldsymbol{B}^T \boldsymbol{y}/n$$

We use the identity $(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1}$ to expand the inverse term in the estimator. This later allows us to split the bias term into the part due to approximating $f^{(r)}(x)$ with a spline (approximation bias) and the other part due to penalization (shrinkage bias).

$$\begin{aligned}
(G_{n,q} + \lambda_n \boldsymbol{P_m})^{-1} &= G_{n,q}^{-1} - G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})H_n^{-1} \\
&= G_{n,q}^{-1} - H_n^{-1}(\lambda_n \boldsymbol{P_m})G_{n,q}^{-1}
\end{aligned}$$

where the last equality is by symmetry.

Substituting into $\hat{f}^{(r)}(x)$, we have:

$$\begin{aligned}
\implies \hat{f}^{(r)}(x) &= B_q^{(r)}(x)\left(G_{n,q}^{-1} - H_n^{-1}(\lambda_n \boldsymbol{P_m})G_{n,q}^{-1}\right)\boldsymbol{B}^T\boldsymbol{y}/n \\
&= B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{y}/n - B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{y}/n
\end{aligned}$$

We now focus on the bias of the naive estimator, $E\left[\hat{f}^{(r)}(x)\right] - f^{(r)}(x)$.

From Lemma A1, $\exists s_f \in \boldsymbol{S}(q, \underline{\mathrm{t}})$, the space of spline functions of order q defined on knots $\underline{\mathrm{t}}$ such that $||f^{(r)} - s_f^{(r)}|| = O(h^{q-r}) + o(h^{p-r})$. The bias of $\hat{f}^{(r)}(x)$ can be written as:

$$E\left[\hat{f}^{(r)}(x)\right] - f^{(r)}(x) = \left[E\left(\hat{f}^{(r)}(x)\right) - s_f^{(r)}(x)\right] + \left[s_f^{(r)}(x) - f^{(r)}(x)\right] \tag{A1}$$

Equation (A1) above allows us to separately evaluate the approximation bias and shrinkage bias for estimating $f^{(r)}(x)$. Notice that Lemma A1 provides information on the rate of convergence on the second term in (A1), we will next focus expressing the first term in a form that isolates the effect of penalization on the bias. Substituting the previously derived expression for $\hat{f}^{(r)}(x)$ into the first term, we have

$$\begin{aligned}
E\hat{f}^{(r)}(x) - s_f^{(r)}(x) &= B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{f}/n - s_f^{(r)}(x) \\
&\quad -B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{f}/n \\
&= B_q^{(r)}(x)\underline{\gamma} - s_f^{(r)}(x) - B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\gamma}
\end{aligned}$$

where $\underline{\gamma} = G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{f}/n$ and $\boldsymbol{f} = E\left[\boldsymbol{y}\right]$.

But $s_f^{(r)}(x) = B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{s_f}/n$, where $\boldsymbol{s_f} = \{s_f(x_1), s_f(x_2), \ldots, s_f(x_n)\}$

$$\begin{aligned}
\implies E\hat{f}^{(r)}(x) - s_f^{(r)}(x) &= B_q^{(r)}(x)\underline{\gamma} - B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\gamma} \\
&\quad -B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{s_f}/n \\
&= B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{f}/n - B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{s_f}/n \\
&\quad -B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\gamma} \\
&= B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T(\boldsymbol{f} - \boldsymbol{s_f})/n - B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\gamma} \\
&= B_q^{(r)}(x)G_{n,q}^{-1}\underline{\boldsymbol{\alpha}} - B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\gamma} \qquad \text{(A2)}
\end{aligned}$$

where $\underline{\boldsymbol{\alpha}} = \boldsymbol{B}^T(\boldsymbol{f} - \boldsymbol{s_f})/n$

Let $Q(x)$ be a distribution of x on $[0,1]$ with positive continuous density $q(x)$. Then substituting (A2) into (A1) and using the triangle inequality, we can evaluate the squared bias of $\hat{f}^{(r)}(x)$ as:

$$\begin{aligned}\int_0^1 \left(E\left[\hat{f}^{(r)}(x)\right] - f^{(r)}(x)\right)^2 q(x)dx \le & \int_0^1 \left(s_f^{(r)}(x) - f^{(r)}(x)\right)^2 q(x)dx \\ & + \underline{\boldsymbol{\alpha}}^T G_{n,q}^{-1} G_q^{(r)} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} \\ & + \underline{\boldsymbol{\gamma}}^T (\lambda_n \boldsymbol{P_m}) H_n^{-1} G_q^{(r)} H_n^{-1} (\lambda_n \boldsymbol{P_m}) \underline{\boldsymbol{\gamma}}\end{aligned} \tag{A3}$$

where $G_q^{(r)} = \int_0^1 B_q^{(r)T}(x) B_q^{(r)}(x) q(x)dx$. The first and second terms in (A3) represent the part of the bias due to using spline functions to estimate $f^{(r)}(x)$, and the last term represents the part of the bias due to penalization.

Observe that, by Lemma A1,

$$\begin{aligned}\int_0^1 \left(s_f^{(r)}(x) - f^{(r)}(x)\right)^2 q(x)dx \le q_{\max} & \int_0^1 \left(s_f^{(r)}(x) - f^{(r)}(x)\right)^2 dx \\ & = O\left(h^{2(q-r)}\right) + o\left(h^{2(p-r)}\right)\end{aligned}$$

where, $q_{\max} = \max\limits_{0 \le x \le 1} q(x) < \infty$.

For the second term in (A3), we use the result $||G_q^{(r)}||_\infty = O(h^{-2r})$, from Lemma A2. We also use $||G_{n,q}^{-1}||_\infty = O(h^{-1})$ from Lemma A3 and Lemma 6.10 of Agarwal & Studden (1980) that $||\underline{\boldsymbol{\alpha}}||_{\max} = o(h^{p+1})$.

Let $G_q^{(r)\frac{1}{2}}$ be a square and symmetric matrix such that $G_q^{(r)} = G_q^{(r)\frac{1}{2}} G_q^{(r)\frac{1}{2}}$.

We write

$$\begin{aligned}\underline{\boldsymbol{\alpha}}^T G_{n,q}^{-1} G_q^{(r)} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} &= \left(G_q^{(r)\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}}\right)^T \left(G_q^{(r)\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}}\right) \\ &= ||G_q^{(r)\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}}||_2^2 \\ &= ||G_q^{(r)\frac{1}{2}} G_{n,q}^{-1}||_2^2 ||\underline{\boldsymbol{\alpha}}||_2^2\end{aligned}$$

Using the fact that for a real matrix A, $||A||_2^2 = \rho(A^T A) \le ||A^T A||_\infty$, here, $\rho(A^T A)$ is the largest eigen value of $A^T A$, we write:

$$\begin{aligned}||G_q^{(r)\frac{1}{2}} G_{n,q}^{-1}||_2^2 &\le ||G_{n,q}^{-1} G_q^{(r)\frac{1}{2}} G_q^{(r)\frac{1}{2}} G_{n,q}^{-1}||_\infty \\ &= ||G_{n,q}^{-1} G_q^{(r)} G_{n,q}^{-1}||_\infty \\ &\le ||G_{n,q}^{-1}||_\infty ||G_q^{(r)}||_\infty ||G_{n,q}^{-1}||_\infty \\ &= O(h^{-1}) O(h^{-2r}) O(h^{-1})\end{aligned}$$

Also, from $||\underline{\boldsymbol{\alpha}}||_{\max} = o(h^{p+1})$, we have:

$$\begin{aligned}\underline{\boldsymbol{\alpha}}^T G_{n,q}^{-1} G_q^{(r)} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} &= O(h^{-1}) O(h^{-2r}) O(h^{-1}) o(h^{2(p+1)}) \\ &= o(h^{2p-2r})\end{aligned}$$

Next, we focus on the part of the bias due to penalization as given by the third term in (A3). First, note that from de Boor (1978) and Lemma 5.2 of Zhou & Wolfe (2000), $D^{(r)}$ in $B_q^{(r)}(x) = B_{q-r}(x) D^{(r)}$, is such that

$$||D^{(r)}||_\infty = O(h^{-r})$$

This can be easily seen by inspecting the elements of $D^{(r)}$.

$$\therefore B_q^{(r)T}(x) B_q^{(r)}(x) = D^{(r)T} B_{q-r}^T(x) B_{q-r}(x) D^{(r)} = O(h^{-2r}) B_{q-r}^T(x) B_{q-r}(x)$$

Thus, we can write

$$\begin{aligned} G_q^{(r)} &= \int_0^1 B_q^{(r)T}(x) B_q^{(r)}(x) q(x) dx \\ &= O(h^{-2r}) \int_0^1 B_{q-r}^T(x) B_{q-r}(x) q(x) dx \\ &= O(h^{-2r}) G_{q-r} \end{aligned}$$

where $G_{q-r} = \int_0^1 B_{q-r}^T(x) B_{q-r}(x) q(x) dx.$

Also, by the WLLN, $G_{n,q-r} = G_{q-r} + o(1)$.

Therefore:

$$\begin{aligned} \underline{\gamma}^T (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) H_n^{-1} G_q^{(r)} H_n^{-1} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) \underline{\gamma} &= O(h^{-2r}) \underline{\gamma}^T (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) H_n^{-1} \\ &\quad \times G_{q-r} H_n^{-1} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) \underline{\gamma} \\ &= O(h^{-2r}) \underline{\gamma}^T (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) H_n^{-1} \\ &\quad \times G_{n,q-r} H_n^{-1} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) \underline{\gamma} \end{aligned}$$

where $G_{n,q-r} = B_{q-r}^T B_{q-r}/n$, the version of $G_{n,q}$ based on B-splines of order $q-r$. Note that the decay of the eigenvalues of $G_{n,q}$ does not depend on q (see Lemma A3). Therefore, we will use $G_{n,q}$ instead of $G_{n,q-r}$ in the derivations that follow for asymptotic order. This simplifies the calculations since H_n^{-1} depends on $G_{n,q}$.

From

$$\begin{aligned} H_n^{-1} &= [G_{n,q} + (\lambda_n \boldsymbol{P}_{\boldsymbol{m}})]^{-1} \\ &= \left[G_{n,q}^{\frac{1}{2}} \left(G_{n,q}^{\frac{1}{2}} + \lambda_n G_{n,q}^{-\frac{1}{2}} \boldsymbol{P}_{\boldsymbol{m}} \right) \right]^{-1} \\ &= \left(G_{n,q}^{\frac{1}{2}} + \lambda_n G_{n,q}^{-\frac{1}{2}} \boldsymbol{P}_{\boldsymbol{m}} \right)^{-1} G_{n,q}^{-\frac{1}{2}} \end{aligned}$$

we can write

$$\begin{aligned} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) H_n^{-1} G_{n,q} H_n^{-1} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) &= (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) \left(G_{n,q}^{\frac{1}{2}} + \lambda_n G_{n,q}^{-\frac{1}{2}} \boldsymbol{P}_{\boldsymbol{m}} \right)^{-1} G_{n,q}^{-\frac{1}{2}} G_{n,q} \\ &\quad \times \left(G_{n,q}^{\frac{1}{2}} + \lambda_n G_{n,q}^{-\frac{1}{2}} \boldsymbol{P}_{\boldsymbol{m}} \right)^{-1} G_{n,q}^{-\frac{1}{2}} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) \end{aligned}$$

Let $\tilde{P} = G_{n,q}^{-\frac{1}{2}} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) G_{n,q}^{-\frac{1}{2}} \implies \tilde{P} G_{n,q}^{\frac{1}{2}} = G_{n,q}^{-\frac{1}{2}} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}})$

Substituting into (A4), we have

$$\begin{aligned} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) H_n^{-1} G_{n,q} H_n^{-1} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) &= (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) \left(G_{n,q}^{\frac{1}{2}} + \tilde{P} G_{n,q}^{\frac{1}{2}} \right)^{-1} G_{n,q}^{\frac{1}{2}} \\ &\quad \times \left(G_{n,q}^{\frac{1}{2}} + \tilde{P} G_{n,q}^{\frac{1}{2}} \right)^{-1} \tilde{P} G_{n,q}^{\frac{1}{2}} \\ &= (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) G_{n,q}^{-\frac{1}{2}} \left(I + \tilde{P} \right)^{-1} G_{n,q}^{\frac{1}{2}} G_{n,q}^{-\frac{1}{2}} \\ &\quad \times \left(I + \tilde{P} \right)^{-1} \tilde{P} G_{n,q}^{\frac{1}{2}} \\ &= G_{n,q}^{\frac{1}{2}} \tilde{P} (I + \tilde{P})^{-2} \tilde{P} G_{n,q}^{\frac{1}{2}} \end{aligned}$$

where in the second equality, we have used the fact that $G_{n,q}^{\frac{1}{2}} + \tilde{P} G_{n,q}^{\frac{1}{2}} = (I + \tilde{P}) G_{n,q}^{\frac{1}{2}}$ and that $(\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) G_{n,q}^{-\frac{1}{2}} = G_{n,q}^{\frac{1}{2}} \tilde{P}$ in the last equality.

Using the above, we can then write:

$$\underline{\gamma}^T (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) H_n^{-1} G_{n,q} H_n^{-1} (\lambda_n \boldsymbol{P}_{\boldsymbol{m}}) \underline{\gamma} = \underline{\gamma}^T G_{n,q}^{\frac{1}{2}} \tilde{P} \left(I + \tilde{P} \right)^{-2} \tilde{P} G_{n,q}^{\frac{1}{2}} \underline{\gamma}$$

It follows from the symmetry of $\tilde{P}$ that $||\tilde{P}||_2\tilde{P} - \tilde{P}\left(I+\tilde{P}\right)^{-2}\tilde{P}$ and $\tilde{P} - \tilde{P}(I+\tilde{P})^{-2}\tilde{P}$ are positive semidefinite.

First, for $||\tilde{P}||_2\tilde{P} - \tilde{P}\left(I+\tilde{P}\right)^{-2}\tilde{P}$ positive semidefinite, we have

$$\begin{aligned}
\underline{\gamma}^T(\lambda_n\boldsymbol{P_m})H_n^{-1}G_q^{(r)}H_n^{-1}(\lambda_n\boldsymbol{P_m})\underline{\gamma} &= O(h^{-2r})||\tilde{P}||_2\underline{\gamma}^T G_{n,q}^{\frac{1}{2}}\tilde{P}G_{n,q}^{\frac{1}{2}}\underline{\gamma} \\
&= O(h^{-2r})||\tilde{P}||_2\underline{\gamma}^T(\lambda_n\boldsymbol{P_m})\underline{\gamma} \\
&= O(h^{-2r})||G_{n,q}^{-\frac{1}{2}}(\lambda_n\boldsymbol{P_m})G_{n,q}^{-\frac{1}{2}}||_2\underline{\gamma}^T(\lambda_n\boldsymbol{P_m})\underline{\gamma} \\
&= O(h^{-2r})||G_{n,q}^{-1}||_2||(\lambda_n\boldsymbol{P_m})||_2\underline{\gamma}^T(\lambda_n\boldsymbol{P_m})\underline{\gamma}
\end{aligned}$$

where we have used $G_{n,q}^{\frac{1}{2}}\tilde{P}G_{n,q}^{\frac{1}{2}} = (\lambda_n\boldsymbol{P_m})$ in the second equality and substituted $\tilde{P}$ in the third.

By Assumption 5, $||\boldsymbol{P_m}||_2 = O(h^{1-2m})$ and from Lemma A4, $\underline{\gamma}^T\boldsymbol{P_m}\underline{\gamma} = O(1)$.
Therefore:

$$\begin{aligned}
\underline{\gamma}^T(\lambda_n\boldsymbol{P_m})H_n^{-1}G_q^{(r)}H_n^{-1}(\lambda_n\boldsymbol{P_m})\underline{\gamma} &= O(h^{-2r})O(h^{-1})O(\lambda_n h^{1-2m})O(\lambda_n) \\
&= O(\lambda_n^2 h^{-2m-2r}).
\end{aligned}$$

Also, $\tilde{P} - \tilde{P}(I+\tilde{P})^{-2}\tilde{P}$ positive semidefinite, we have

$$\begin{aligned}
\underline{\gamma}^T(\lambda_n\boldsymbol{P_m})H_n^{-1}G_q^{(r)}H_n^{-1}(\lambda_n\boldsymbol{P_m})\underline{\gamma} &= O(h^{-2r})\underline{\gamma}^T G_{n,q}^{\frac{1}{2}}\tilde{P}(I+\tilde{P})^{-2}\tilde{P}G_{n,q}^{\frac{1}{2}}\underline{\gamma} \\
&= O(h^{-2r})\underline{\gamma}^T G_{n,q}^{\frac{1}{2}}\tilde{P}G_{n,q}^{\frac{1}{2}}\underline{\gamma} \\
&= O(h^{-2r})\underline{\gamma}^T(\lambda_n\boldsymbol{P_m})\underline{\gamma} \\
&= O(\lambda_n h^{-2r})
\end{aligned}$$

$\therefore \underline{\gamma}^T(\lambda_n\boldsymbol{P_m})H_n^{-1}G_q^{(r)}H_n^{-1}(\lambda_n\boldsymbol{P_m})\underline{\gamma} = O\left\{\min\left(\lambda_n^2 h^{-2m-2r}, \lambda_n h^{-2r}\right)\right\}$
This concludes the proof for bias in (A3).

Next, we look at the variance part:

$$\begin{aligned}
Var(\hat{f}^{(r)}(x)) &= Var\left(B_q^{(r)}(x)H_n^{-1}\boldsymbol{B}^T\boldsymbol{y}/n\right) \\
&= B_q^{(r)}(x)H_n^{-1}\boldsymbol{B}^T Var(\boldsymbol{y}/n)\boldsymbol{B}H_n^{-1}B_q^{(r)T}(x) \\
&= \frac{\sigma^2}{n}tr\left\{B_q^{(r)}(x)H_n^{-1}(\boldsymbol{B}^T\boldsymbol{B}/n)H_n^{-1}B_q^{(r)T}(x)\right\} \\
&= \frac{\sigma^2}{n}tr\left\{B_q^{(r)}(x)H_n^{-1}G_{n,q}H_n^{-1}B_q^{(r)T}(x)\right\} \\
&= \frac{\sigma^2}{n}tr\left\{H_n^{-1}G_{n,q}H_n^{-1}B_q^{(r)T}(x)B_q^{(r)}(x)\right\}
\end{aligned}$$

Note that we have used the rotation property of the trace in the last equality.

Therefore,

$$\begin{aligned}
\int_0^1 Var(\hat{f}^{(r)}(x))q(x)dx &= \frac{\sigma^2}{n}tr\left\{H_n^{-1}G_{n,q}H_n^{-1}G_q^{(r)}\right\} \\
&= O(h^{-2r})\frac{\sigma^2}{n}tr\left\{H_n^{-1}G_{n,q}H_n^{-1}G_{n,q}\right\}
\end{aligned}$$

where in the last equality, we have used the fact that $G_q^{(r)} = O(h^{-2r})G_{q-r}$ and that the decay rates of $G_{n,q}$ do not depend on q.

From

$$\begin{aligned} H_n^{-1} &= (G_{n,q} + (\lambda_n \boldsymbol{P_m}))^{-1} \\ &= \left[G_{n,q}\left(I + G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})\right)\right]^{-1} \\ &= \left[I + G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})\right]^{-1} G_{n,q}^{-1} \end{aligned}$$

$$\implies H_n^{-1} G_{n,q} = \left[I + G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})\right]^{-1}.$$

Note that $G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m}) = G_{n,q}^{-\frac{1}{2}} G_{n,q}^{-\frac{1}{2}}(\lambda_n \boldsymbol{P_m})$ and by the rotation property of the trace,

$$\begin{aligned} tr\left[G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})\right] &= tr\left[G_{n,q}^{-\frac{1}{2}} G_{n,q}^{-\frac{1}{2}}(\lambda_n \boldsymbol{P_m})\right] \\ &= tr\left[G_{n,q}^{-\frac{1}{2}}(\lambda_n \boldsymbol{P_m}) G_{n,q}^{-\frac{1}{2}}\right] \\ &= tr\left[\tilde{P}\right] \end{aligned}$$

$$\begin{aligned} \therefore \int_0^1 Var(\hat{f}^{(r)}(x)) q(x) dx &= O(h^{-2r}) \frac{\sigma^2}{n} tr\left[(I + \tilde{P})^{-2}\right] \\ &= O(h^{-2r}) \frac{\sigma^2}{n} \left|\left|(I + \tilde{P})^{-2}\right|\right|_F^2 \\ &= O(h^{-2r}) \frac{\sigma^2}{n} O\left\{\frac{1}{\max(h, \lambda_n^{1/2m})}\right\} \\ &= O(h^{-2r}) \frac{\sigma^2}{n} O\left\{\min(h^{-1}, \lambda_n^{-1/2m})\right\} \\ &= O(K^{2r}) \frac{\sigma^2}{n} O\left\{\min(K, \lambda_n^{-1/2m})\right\} \\ &= O\left(\frac{K_e}{n}\right) \end{aligned}$$

Where in the above, we have used $\left|\left|(I + \tilde{P})^{-2}\right|\right|_F^2 = O\left\{\frac{1}{\max(h, \lambda^{1/2m})}\right\}$ from Lemma 5.2 of Xiao (2019), $K \sim h^{-1}$, and $K_e = \min\left\{K^{2r+1}, K^{2r}\lambda_n^{-1/2m}\right\}$

This completes the proof of the theorem.

*A.3. **Technical Lemmas***

LEMMA A1. *Let $f \in \mathcal{C}^p$, then there exists $s_f \in \boldsymbol{\mathcal{S}}(q, \underline{\mathrm{t}})$ such that*

$$||f^{(r)} - s_f^{(r)}|| = O(h^{q-r}) + o(h^{p-r})$$

for all $r = 0, 1, \ldots, q-2$ and $p \leq q$.

Here, $b(x) = -\frac{f^{(q)}(x)h_i^q}{q!} B_q\left(\frac{x-t_i}{h_i}\right)$ for $t_i \leq x < t_{i+1}$ where $B_q(.)$ is the q^{th} Bernoulli polynomial defined as $B_0(x) = 1$, and $B_k(x) = \displaystyle\int_0^x k B_{k-1}(x) dx + B_k$

and B_k is chosen such that $\int_0^1 B_k(x) dx = 0$.

B_k is known as the k^{th} Bernoulli number (Barrow & Smith, 1979). This Lemma also appears in Barrow & Smith (1979) and Xiao (2019) adopts the general result in Barrow & Smith (1979) to prove the case where $p < q$.

Proof of Lemma A1

We provide a proof for the case where $p = q$ and refer to Remark 3.1 of Xiao (2019) for the case where $p < q$. Xiao (2019) showed that when $p < q$, $||f^{(r)} - s_f^{(r)}|| = o(h^{p-r})$.
For $p = q$, first note that under Assumption 3, Barrow & Smith (1979) showed that

$$\inf_{s(x)\in\boldsymbol{\mathcal{S}}(q,\underline{\mathbf{t}})} ||f^{(r)}(x) - s^{(r)}(x) + b^{*(r)}(x)||_{L_\infty} = o(h^{q-r})$$

This means, there exists an $s_f(x) \in \boldsymbol{\mathcal{S}}(q, \underline{\mathbf{t}})$ such that

$$||f^{(r)}(x) - s_f^{(r)}(x) + b^{*(r)}(x)|| = o(h^{q-r})$$

where $b^*(x) = -\frac{f^{(q)}(t_i)h_i^q}{q!} B_q\left(\frac{x-t_i}{h_i}\right)$, for $t_i \le x < t_{i+1}$ and $b^{*(r)}$ is the r^{th} derivative of b^*. Note that $f^{(q)}(x)$ in $b(x)$ is replaced with $f^{(q)}(t_i)$ in $b^*(x)$.

With $p = q$, we have that $f \in \mathcal{C}^q[0, 1]$. Therefore, from Taylor's theorem, $f^{(q)}(x) = f^{(q)}(t_i) + o(1)$.

$$\implies b(x) = b^*(x) + o(h^q)$$

The derivative of the Bernoulli polynomial of order k is given by $\boldsymbol{B}'_k(x) = \boldsymbol{B}_{k-1}(x)$ (Barrow & Smith, 1979), it therefore follows that

$$b^{(r)}(x) = b^{*(r)}(x) + o(h^{q-r})$$

for $r = 0, 1, 2, \dots, q - 2$. But $||b^*|| = O(h^q)$ by definition, giving $||b^{(r)}|| = O(h^{q-r})$.

Combining this with the case where $p < q$, we have that $||f^{(r)} - s_f^{(r)}|| = O(h^{q-r}) + o(h^{p-r})$ for all $p \le q$.

LEMMA A2. *Given* $G_q^{(r)} = \int_0^1 B^{(r)}(x)B_q^{(r)}(x)q(x)dx$,

$$||G_q^{(r)}||_\infty = O(h^{-2r})$$

A.4. Proof of Lemma A2

Note that $B_q^{(r)}(x) = B_{q-r}(x)D^{(r)}$

$$\begin{aligned}\therefore G_q^{(r)} &= \int_0^1 B_{q-r}(x)D^{(r)}D^{T(r)}B_{q-r}^T(x)q(x)dx \\ &= O(h^{-2r})\int_0^1 B_{q-r}(x)B_{q-r}^T(x)q(x)dx \\ &= O(h^{-2r}) \times q_{\max} \\ &= O(h^{-2r})\end{aligned}$$

Where $q_{\max} = \max\limits_{x\in[0,1]} q(x) < \infty$. Also, note that B-spline bases are bounded by 1 $\forall x \in [0, 1]$.

LEMMA A3. *Let* $G_{n,q} = \boldsymbol{B}^T\boldsymbol{B}/n$ *where* $\boldsymbol{B} = [B(x_1), B(x_2), \dots, B(x_n)]^T \in \mathbb{R}^{n\times K}$ *is a matrix of basis functions with each* $B(x) \in \mathbb{R}^K$ *being a vector of basis functions of order q at x.*
Then

$$||G_{n,q}^{-1}||_\infty = O(h^{-1})$$

Proof of Lemma A3

This Lemma is adapted from Lemma 6.3 of Zhou et al. (1998) and the key idea is to show that the elements of $G_{n,q}^{-1}$ decay exponentially and of order h^{-1}. We provide the proof here for convenience.

Let $\lambda_{\max}$ and $\lambda_{\min}$ be the maximum and minimum eigenvalues of $G_{n,q}$ respectively. Since $G_{n,q}$ is a band matrix, Theorem 2.2 of Demko (1977) is used. First, we need to satisfy the conditions of the theorem.

Note that

$$\begin{aligned}||\lambda_{\max}^{-1}G_{n,q}||_2 &= \lambda_{\max}^{-1}||G_{n,q}||_2\\ &= \lambda_{\max}^{-1}\max_{\sum_{i=1}^{K} z_i^2=1}||G_{n,q}\underline{\boldsymbol{z}}||_2\\ &\leq 1\end{aligned}$$

Where the $\max$ term in the second equality gives some eigenvalue that is at most $\lambda_{\max}^{-1}$.

Also,

$$\begin{aligned}||\lambda_{\max}G_{n,q}^{-1}||_2 &= \frac{\lambda_{\max}}{\lambda_{\min}}\left|\left|\lambda_{\min}G_{n,q}^{-1}\right|\right|_2\\ &\leq \frac{\lambda_{\max}}{\lambda_{\min}}\end{aligned}$$

Lemma 6.2 of Zhou et al. (1998) provides bounds on the eigenvalues of $G_{n,q}$. In particular, for large n, there exist constants c_1 and c_2 such that

$$c_1h/2 \leq \lambda_{\min} \leq \lambda_{\max} \leq 2c_2h$$

Therefore by Theorem 2.2 of Demko (1977), there exists constants $c > 0$ and $\gamma \in (0, 1)$ which depend only on c_1, c_2 and q such that:

$$|\lambda_{\max}g_{ij}| \leq c\gamma^{|i-j|} \tag{A4}$$

where g_{ij} is the (i, j)th element of $G_{n,q}^{-1}$.

From equation (A4),

$$|g_{ij}| \leq c\lambda_{\max}^{-1}\gamma^{|i-j|} \leq 2(c/c_1)h^{-1}\gamma^{|i-j|}$$

which means that $G_{n,q}^{-1} = O(h^{-1})$. This completes the proof of Lemma A3.

LEMMA A4. *Suppose* $\underline{\boldsymbol{\gamma}} = G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{f}/n$ *and* $\boldsymbol{P_m}$ *is the penalty matrix for the penalized spline estimator in* (3)*,*

then

$$\underline{\boldsymbol{\gamma}}^T\boldsymbol{P_m}\underline{\boldsymbol{\gamma}} = O(1)$$

Proof of Lemma A4

Again, this Lemma is adapted from Lemma 8.4 of Xiao (2019) which puts a bound on the penalty matrix of the penalized spline estimator. The proof follows closely from the proof in Xiao (2019) with more clarity.

Observe that we can write

$$\begin{aligned}\underline{\boldsymbol{\gamma}} = G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{f}/n &= G_{n,q}^{-1}\boldsymbol{B}^T(\boldsymbol{f}-\boldsymbol{s_f})/n + G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{s_f}/n\\ &= G_{n,q}^{-1}\boldsymbol{B}^T(\boldsymbol{f}-\boldsymbol{s_f})/n + \underline{\boldsymbol{\beta}}\\ &= G_{n,q}^{-1}\underline{\boldsymbol{\alpha}} + \underline{\boldsymbol{\beta}}\end{aligned}$$

where $\underline{\boldsymbol{\beta}} = G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{s_f}/n$ and $\underline{\boldsymbol{\alpha}} = \boldsymbol{B}^T(\boldsymbol{f}-\boldsymbol{s_f})/n.$

Since $\boldsymbol{P_m}$ is positive semi-definite, we can use the Cauchy-Schwarz inequality defined for an inner product $\langle x, y\rangle_{\boldsymbol{P_m}} = x^T\boldsymbol{P_m}y$ and write:

$$\left(\underline{\boldsymbol{\gamma}}\boldsymbol{P_m}\underline{\boldsymbol{\gamma}}\right)^{\frac{1}{2}} \leq \left(\underline{\boldsymbol{\alpha}}^TG_{n,q}^{-1}\boldsymbol{P_m}G_{n,q}^{-1}\underline{\boldsymbol{\alpha}}\right)^{\frac{1}{2}} + \left(\underline{\boldsymbol{\beta}}^T\boldsymbol{P_m}\underline{\boldsymbol{\beta}}\right)^{\frac{1}{2}} \tag{A5}$$

By Assumption, $\underline{\boldsymbol{\beta}}^T\boldsymbol{P_m}\underline{\boldsymbol{\beta}} = O(1)$, therefore showing that the first term in A5 is O(1) completes the proof.

In the following, we use the following matrix relations. Let $A \in \mathbb{R}^{m\times n}$, then

$$n^{-1/2}||A||_\infty \leq ||A||_2 \leq m^{1/2}||A||_\infty \tag{A6}$$

Also, let $P_m^{\frac{1}{2}}$ be a square symmetric matrix such that $\boldsymbol{P_m} = P_m^{\frac{1}{2}} P_m^{\frac{1}{2}}$.
Observe that

$$\begin{aligned}
\underline{\boldsymbol{\alpha}}^T G_{n,q}^{-1} \boldsymbol{P_m} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} &= \left(P_m^{\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}}\right)^T \left(P_m^{\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}}\right) && \text{(A7)}\\
&= ||P_m^{\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}}||_2^2 && \text{(A8)}\\
&\leq ||\underline{\boldsymbol{\alpha}}||_2^2 ||P_m^{\frac{1}{2}} G_{n,q}^{-1}||_2^2 && \text{(A9)}\\
&\leq ||\underline{\boldsymbol{\alpha}}||_2^2 ||P_m^{\frac{1}{2}}||_2^2 ||G_{n,q}^{-1}||_2^2 && \text{(A10)}\\
&\leq ||\underline{\boldsymbol{\alpha}}||_2^2 ||P_m^{\frac{1}{2}}||_2^2 K ||G_{n,q}^{-1}||_\infty^2 && \text{(A11)}\\
&= o(h^{2p+2}) O(h^{1-2m}) O(h^{-1}) O(h^{-2}) && \text{(A12)}\\
&= o(h^{2p-2m}) && \text{(A13)}\\
&= O(1) && \text{(A14)}
\end{aligned}$$

for $p \geq m$. The first and second inequalities are by Cauchy Schwartz inequality, and we have used the matrix identity in (A6) in the third inequality. Also, we have used the result by Agarwal & Studden (1980) for $||\underline{\boldsymbol{\alpha}}||_2^2$ and the assumption that $||\boldsymbol{P_m}||_2 = O(h^{1-2m})$. Finally, we have used Lemma A3 in the third inequality for $||G_{n,q}^{-1}||_\infty$ as well.

REFERENCES

AGARWAL, G. G. & STUDDEN, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *The Annals of Statistics* **8**, 1307–1325.

BARROW, D. & SMITH, P. (1979). Efficient L^2 approximation by splines. *Numerische Mathematik* **33**, 101–114.

CHARNIGO, R., HALL, B. & SRINIVASAN, C. (2011). A generalized cp criterion for derivative estimation. *Technometrics* **53**, 238–253.

CHAUDHURI, P. & MARRON, J. S. (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association* **94**, 807–823.

CLAESKENS, G., KRIVOBOKOVA, T. & OPSOMER, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika* **96**, 529–544.

CRAVEN, P. & WAHBA, G. (1978). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377–403.

DAI, W., TONG, T. & GENTON, M. G. (2016). Optimal estimation of derivatives in nonparametric regression. *Journal of Machine Learning Research* **17**, 1–25.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer New York.

DE BRABANTER, K., BRABANTER, J. D., SUYKENS, J. A. & MOOR, B. D. (2011). Kernel regression in the presence of correlated errors. *Journal of Machine Learning Research* **12**, 1955–1976.

DE BRABANTER, K., DE BRABANTER, J., DE MOOR, B. & GIJBELS, I. (2013). Derivative estimation with local polynomial fitting. *J. Mach. Learn. Res.* **14**, 281–301.

DEMKO, S. (1977). Inverses of band matrices and local convergence of spline projections. *SIAM Journal on Numerical Analysis* **14**, 616–619.

EILERS, P. H. C. & MARX, B. D. (1996). Flexible smoothing with b -splines and penalties. *Statist. Sci.* **11**, 89–121.

EILERS, P. H. C. & MARX, B. D. (2010). Splines, knots, and penalties. *WIREs Computational Statistics* **2**, 637–653.

EUBANK, R. L. & SPECKMAN, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association* **88**, 1287–1301.

FAN, J. & GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

FISHER, J. D., JOHNSON, D. S., SMEEDING, T. M. & THOMPSON, J. P. (2020). Estimating the marginal propensity to consume using the distributions of income, consumption, and wealth. *Journal of Macroeconomics* **65**, 103218.

HILDENBRAND, W. (1989). Facts and ideas in microeconomic theory. *European Economic Review* **33**, 251–276.

HÄRDLE, W., HART, J., MARRON, J. S. & TSYBAKOV, A. B. (1992). Bandwidth choice for average derivative estimation. *Journal of the American Statistical Association* **87**, 218–226.

MÜLLER, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data. [electronic resource]*. Lecture Notes in Statistics: 46. Springer New York.

MÜLLER, H.-G., STADTMÜLLER, U. & SCHMITT, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika* **74**, 743–749.

Park, C. & Kang, K.-H. (2008). Sizer analysis for the comparison of regression curves. *Computational Statistics & Data Analysis* **52**, 3954–3970.

Ramsay, J. O. & Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90**, 1257–1270.

Ruppert, D. & Wand, M. P. (1994). Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics* **22**, 1346 – 1370.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge Mathematical Library. Cambridge University Press, 3rd ed.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics* **12**, 898–916.

Simpkin, A. & Newell, J. (2013). An additive penalty p-spline approach to derivative estimation. *Comput. Stat. Data Anal.* **68**, 30–43.

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

Wahba, G. (1985). A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. *The Annals of Statistics* **13**, 1378 – 1402.

Wahba, G. & Wang, Y. (1990). When is the optimal regularization parameter insensitive to the choice of the loss function? *Communications in Statistics - Theory and Methods* **19**, 1685–1700.

Wood, S. N. (2017). *Generalizd Additive Models, second edition*. Texts in Statistical Science. CRC Press.

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

Xiao, L., Li, Y., Apanasovich, T. V. & Ruppert, D. (2012). Local asymptotics of p-splines.

Zhou, S., Shen, X. & Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics* **26**, 1760–1782.

Zhou, S. & Wolfe, D. (2000). On derivative estimation in spline regression. *Statistica Sinica* **10**, 93–108.