Electronic Journal of Statistics
Vol. 4 (2010) 1411–1442
ISSN: 1935-7524
DOI: 10.1214/10-EJS588

Convergence rate and Bahadur type representation of general smoothing spline M-estimates

Zuofeng Shang

Department of Statistics,
University of Wisconsin Madison
e-mail: shang@stat.wisc.edu

Abstract: This work was motivated by Cox and O'Sullivan (1990) who derived the optimal convergence rates for smoothing spline estimates when the loss function is sufficiently smooth. However, the study of statistical estimates resulting from nonsmooth criteria functions has become popular in recent years. In this paper, we will study the asymptotic properties of the smoothing spline estimates when the criteria functions are insufficiently smooth. Here, the smoothing spline estimate is defined as an approximate solution to an M-estimating equation. We prove that if the derivative of loss function is Lipschitz, then the convergence rate and Bahadur type representation of the estimate can be derived simultaneously. For a specific class of loss functions with discontinuous derivatives, the Bahadur type representation is also presented provided that we know the convergence rate. Examples are given when Huber's robust loss and median loss are employed.

AMS 2000 subject classifications: Primary 62G08, 62G35; secondary 62G20.
Keywords and phrases: Smoothing spline estimate, M-estimating equation, optimal convergence rate, Bahadur type representation, Sobolev spaces, Fréchet derivative, quantile loss, Huber's loss.

Received May 2010.

1. Introduction

Consider the bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which form an independent and identical sample from population (X, Y). We assume that the data are linked by the following nonparametric regression model

$$Y_i = \theta_0(X_i) + e_i, \ i = 1, \ldots, n, \tag{1.1}$$

where X_i's take values in $\mathbb{I} = [0, 1]$, e_i's are iid noises independent of X_i's. It is of particular interest to estimate the unknown functional parameter θ_0, a sufficiently smooth element in some Soblev space. Usually, the estimation problem under model (1.1) is ill-posed since we allow the space of parameters to be infinite-dimensional. However, we can overcome this ill-posedness by using a penalty functional, i.e., we estimate θ_0 by optimizing a penalized criteria

function. This estimation procedure is called the method of penalization (see Wahba (1990) for a detailed review and related references).

When the criteria function is sufficiently smooth, the convergence rate of the penalized estimates has been studied by many authors under various settings. The most important literature includes Chen (1991); Cox (1983, 1988); Cox and O'Sullivan (1990); Gu and Qiu (1993); Gu and Ma (2005); O'Sullivan (1993, 1995); Silverman (1982), and the references therein. For non-penalized optimization, Wong and Severini (1991) obtained optimal convergence rates for ϵ-MLE.

However, in practice, the estimates resulted from a nonsmooth loss function l are also important. Examples include (1) $l(s) = s^2 1(|s| \leq c)/2 + (c|s| - c^2/2)1(|s| > c)$ for some $c > 0$; (2) $l(s) = |s|$. The function $\varphi(s) = \frac{d}{ds} l(s)$ is called a moment function. The corresponding estimates will have unique features (such as robustness), and hence, their theoretical properties need to be explored. Shen (1998) used a penetrating method to obtain the optimal convergence rates of spline quantile estimate. As far as we know, there seems to be little theory treating the infinite-dimensional smoothing spline estimates resulted from general nonsmooth φ. Related references include Shen and Wong (1994) who studied asymptotic properties of a sieve MLE, and Chen and Pouzo (2008) who obtained optimal convergence rates for sieve estimates identified by a class of moment estimating equations with general nonsmooth moment functions.

In this paper, we will study the asymptotic behavior of penalized estimates in two general situations. In the first, by assuming that φ is Lipschitz, we simultaneously obtain the optimal convergence rates and Bahadur type representation for the estimates, provided that we know that the estimate is consistent. In the second, we allow φ to be discontinuous, and obtain a Bahadur type representation for the estimates provided that we know the convergence rate. On the one hand, our results are not only about the convergence rate but also about the Bahadur type representation, which is different from the previous contributions mainly addressing the problem of convergence rate. On the other hand, the penalized estimate considered in this paper is obtained directly over the infinite-dimensional parameter space, which is different from a sieve estimate obtained over a sequence of finite-dimensional sub-spaces approaching the entire parameter space.

Since Bahadur type representation is part of our work, it seems necessary to review some relevant literature about this issue. Most of the existing results have been built under finite-dimensional situations, i.e., θ_0 is a finite-dimensional parameter. Relevant references include He and Shao (1996), Wu (2005, 2007), and the reference therein. But, when the parameter space is infinite-dimensional, e.g., a Sobolev space, little result has been gained. Portnoy (1997) derived a Bahadur type representation for quantile smoothing spline estimates which pointwise holds on some selected knots in $\mathbb{I}$ (Portnoy, 1997, Theorem 2.2). While it still remains open whether the Bahadur type representation holds when general insufficiently smooth loss functions and usual Sobolev norms have been employed, which is also a motivation of this work.

The rest of this paper is organized as follows. In Section 2, notation and assumptions which are needed for the statement of the main results are introduced. Section 3 contains the main results of this paper. In Section 4, examples illustrating the applications of the main results are presented. Section 5 contains some additional theoretical framework which facilitates the technical arguments. Section 6 contains some concluding remarks and future work. Proofs of the main results can be found in Appendix A, and the preliminary results are proved in Appendix B.

2. Notation and assumptions

Let X be a random variable taking values in $\mathbb{I} := [0,1]$ and have distribution with (Lebesgue) density f supported on $\mathbb{I}$. Let $\Theta_1 = H^m(\mathbb{I})$ be a Sobolev space of order m, i.e.,

$$H^m(\mathbb{I}) = \{\theta : I \mapsto \mathbb{R} | \theta^{(j)} \text{ are absolutely continuous, } j = 0, \ldots, m-1, \theta^{(m)} \in L^2(\mathbb{I})\},$$

where $\theta^{(m)}$ denotes the m-order derivative of θ. Let $V(\theta, \tilde{\theta}) = E_X\{\theta(X)\tilde{\theta}(X)\} = \int_0^1 \theta(x)\tilde{\theta}(x)f(x)dx$ and $J(\theta, \tilde{\theta}) = \frac{1}{2}\langle \theta^{(m)}, \tilde{\theta}^{(m)} \rangle_{L^2}$, where $\langle \cdot, \cdot \rangle_{L^2}$ denotes the usual $L^2(\mathbb{I})$-inner product. Note that V defines a norm on $L^2(\mathbb{I})$ by $\| \cdot \|_0 = \sqrt{V(\cdot, \cdot)}$.

Define the inner product $\langle \cdot, \cdot \rangle_1$ on Θ_1 to be $\langle \theta, \tilde{\theta} \rangle_1 = V(\theta, \tilde{\theta}) + J(\theta, \tilde{\theta})$, and denote the corresponding norm to be $\| \cdot \|_1$. By a standard calculation, for any $\theta \in \Theta_1$, the Fréchet derivative of $J(\theta) := J(\theta, \theta)$ at θ, which is denoted by $DJ(\theta)$, satisfies $DJ(\theta)\Delta\theta = 2J(\theta, \Delta\theta)$ for any $\Delta\theta \in \Theta_1$. So, by Riesz's representation theorem, we could view $DJ(\theta)$ as an element in Θ_1 satisfying $\langle \Delta\theta, DJ(\theta) \rangle_1 = 2J(\theta, \Delta\theta)$. Consequently, $W : \theta \mapsto \frac{1}{2}DJ(\theta)$ is a well defined bounded linear operator from Θ_1 to Θ_1 satisfying $< \Delta\theta, W\theta >_1 = J(\theta, \Delta\theta)$ for any $\theta, \Delta\theta \in \Theta_1$.

With W well defined, we can define the estimate of θ_0. Let $X_1, \ldots, X_n$ be *iid* samples drawn from density f. The responses Y_i's and covariates X_i's are linked by the nonparametric model (1.1). We estimate θ_0 by finding the solution $\hat{\theta}_{n,\lambda_n}$ to the following approximate penalized M-estimating equation

$$\|S_{n,\lambda_n}(\theta)\|_1 = o(\delta_n), \tag{2.1}$$

where $S_{n,\lambda_n}(\theta) = \sum_{i=1}^n \varphi(Y_i - \theta(X_i))K_{X_i} + \lambda_n W\theta$, φ is an either smooth or nonsmooth moment function, δ_n is a positive sequence, K is a bivariate kernel function defined on $\mathbb{I} \times \mathbb{I}$ satisfying certain properties and $K_x(\cdot) = K(x, \cdot)$. Hereafter, unless otherwise explicitly stated, we drop the subscript from λ_n. Suppose $\hat{\theta}_{n,\lambda}$ satisfies equation (2.1), then $\hat{\theta}_{n,\lambda}$ is an estimate of θ_0, and is called an M-estimate. The main results in this paper are about the asymptotic properties of $\hat{\theta}_{n,\lambda}$. Before proceeding further, we introduce several technical assumptions which will be used to prove the main results. Firstly, we assume that the bilinear functionals V and J satisfy the following assumption.

Assumption A.1. *V is completely continuous w.r.t. $V + J$.*

Assumptions A.1 was originally introduced by Gu and Qiu (1993) to simultaneously diagonalize V and J. This assumption means that for any $\epsilon > 0$, there are linear functionals $l_1, \ldots, l_k$ such that if $l_j(\theta) = 0$, $j = 1, \ldots, k$, then $V(\theta, \theta) \leq \epsilon(V + J)(\theta, \theta)$ for any $\theta \in \Theta_1$. Assumption A.1 is not a restrictive assumption and holds under general settings (see, e.g., (Weinberger, 1974, Theorem 2.9)). A direct consequence from Assumption A.1 is that the bilinear functionals V and J can be simultaneously diagonalized, which is the following result.

Proposition 2.1. *(Weinberger, 1974, Theorem 3.1) There is a sequence of eigenvalues γ_μ and corresponding eigenvectors h_μ such that*

$$V(h_\mu, h_\nu) = \delta_{\mu\nu}, \quad J(h_\mu, h_\nu) = \gamma_\mu \delta_{\mu\nu}, \ \mu, \nu \in \mathbb{Z}, \tag{2.2}$$

where $\delta_{\mu\nu}$ is the Kronecker's notation. Any $\theta \in \Theta_1$ admits the Fourier expansion $\theta = \sum_\mu V(\theta, h_\mu) h_\mu$ with the convergence held under $\|\cdot\|_1$.

We let $\{h_\mu\}$ be the sequence satisfying (2.2). By Proposition 2.1, we may rewrite $\langle \cdot, \cdot \rangle_1$ as $\langle \theta, \tilde{\theta} \rangle_1 = \sum_\mu (1 + \gamma_\mu) V(\theta, h_\mu) V(\tilde{\theta}, h_\mu)$, $\theta, \tilde{\theta} \in \Theta_1$. Therefore, any two elements $\theta = \sum_\mu \theta_\mu h_\mu$ and $\tilde{\theta} = \sum_\mu \tilde{\theta}_\mu h_\mu$ in Θ_1 admit the following expansion

$$\langle \theta, \tilde{\theta} \rangle_1 = \sum_{\mu \in \mathbb{Z}} \theta_\mu \tilde{\theta}_\mu (1 + \gamma_\mu). \tag{2.3}$$

By Proposition 2.2, the sequence $\{h_\mu\}$ forms a basis in Θ_1. To facilitate the proofs of our main results, we assume, throughout this work, the following assumption for $\{h_\mu\}$.

Assumption A.2. *$\{h_\mu\}$ forms an orthonormal basis in $(L^2(\mathbb{I}), \|\cdot\|_0)$ which satisfies $\sup_\mu \|h_\mu\|_{sup} < \infty$, where $\|h_\mu\|_{sup} := \sup_{x \in \mathbb{I}} |h_\mu(x)|$ denotes the supremum norm of h_μ. Furthermore, the sequence $\{\gamma_\mu\}_{\mu \in \mathbb{Z}}$ satisfies $\gamma_\mu \approx \mu^{2m}$, where $\alpha_n \approx \beta_n$ means that there exist positive constants d_1 and d_2 such that $d_1 < \alpha_n / \beta_n < d_2$ when n goes to ∞.*

Remark 2.1. When X_i's are uniformly generated from $\mathbb{I}$, i.e., the density function of X_i's is $f(x) = 1(x \in \mathbb{I})$, the system $\{h_\mu\} := \{\exp(2\pi\sqrt{-1}\mu x)\}_{\mu \in \mathbb{Z}}$ will satisfy property (2.2) and Assumption A.2. The resulting sequence $\gamma_\mu = 2(2\pi\mu)^{2m}$ for $\mu \in \mathbb{Z}$. So, as $|\mu|$ tends to infinity, $\gamma_\mu \approx \mu^{2m}$.

To define a Sobolev space with a different order, for $b \geq 0$, we let $\|\theta\|_b^2 = \sum_{\mu \in \mathbb{Z}} (1 + \gamma_\mu^b) |V(\theta, h_\mu)|^2$. Let Θ_b be the completion of $\{\theta \in \Theta_1 | \ \|\theta\|_b < \infty\}$ under $\|\cdot\|_b$. According to Theorem 3.2 in Cox (1988), Θ_b is a Sobolev space with order mb under the inner product $\langle \theta, \tilde{\theta} \rangle_b = \sum_{\mu \in \mathbb{Z}} (1 + \gamma_\mu^b) V(\theta, h_\mu) V(\tilde{\theta}, h_\mu)$.

Let $C(\mathbb{I})$ be the Banach space of continuous functions defined on $\mathbb{I}$ endowed with supremum norm $\|\cdot\|_{sup}$. Define $S(\theta) = E\{\varphi(Y - \theta(X)) K_X\}$, where the expectation is taken with respect to (Y, X), and define $S_\lambda(\theta) = nS(\theta) + \lambda W(\theta)$. Let $E\{\varphi(e - u)\} = \zeta(u)$, for $u \in \mathbb{R}$, where e denotes the error in model (1.1).

Assumption A.3. *There is a $\gamma \in (0,1]$, a positive number α and a constant $C_\varphi > 0$ such that*

$$\sup_{s_1,s_2\in\mathbb{R},0<|s_1-s_2|\le\alpha} \frac{|\varphi(s_1)-\varphi(s_2)|}{|s_1-s_2|^\gamma} \le C_\varphi.$$

Assumption A.3′. *$\|\varphi\|_{sup} < \infty$. There is a neighborhood $\mathcal{A}$ of θ_0 in $C(\mathbb{I})$ and constants $C_\mathcal{A}$ and δ_0 such that for any $0 < \delta \le \delta_0$,*

$$\sup_{\tilde\theta\in\mathcal{A}} E\left\{\sup_{\|\theta-\tilde\theta\|_{sup}\le\delta} |\varphi(Y-\theta(X))-\varphi(Y-\tilde\theta(X))|^2\right\} \le C_\mathcal{A}\delta.$$

Assumption A.4. *For some $d > 1$, $\theta_0 \in \Theta_d$; there exists a θ_0-neighborhood $N_0 \subset C(\mathbb{I})$ such that θ_0 is the unique root of $S(\theta)$ in N_0.*

Assumption A.5. *Model errors e_i's are independent of X_i's.*

Assumption A.3′ is the so-called stochastic equi-continuity condition (see Pollard (1982)), and has been adopted by a number of authors (see Chen *et al.* (2003); Chen and Pouzo (2008); He and Shao (1996)). Assumption A.3′ is satisfied by quantile loss. Assumption A.3 is satisfied by several commonly used robust loss functions such as the Huber's loss function. Assumption A.4 essentially requires two things. First, θ_0 is smoother than the elements in Θ_1. As demonstrated later, the estimate $\hat\theta_{n,\lambda}$ will be obtained in space Θ_1, so we need this assumption to guarantee that the true parameter θ_0 is smoother than $\hat\theta_{n,\lambda}$. Second, θ_0 is identifiable since it is assumed to be the unique root of $S(\theta) = 0$. Assumption A.5 requires that the random design values X_i's are independent of the model errors e_i's. This is only a technical assumption which facilitates the proofs. We may use Assumption A.5 to rewrite the functional $S(\theta)$ as $S(\theta) = E_X\{E_e\{\varphi(e-(\theta-\theta_0)(X))|X\}K_X\} = E_X\{\zeta((\theta-\theta_0)(X))K_X\}$, $\forall\theta\in\Theta_0$.

Assumption A.6. *ζ is twice continuously differentiable and both ζ' and ζ'' are upper bounded. Furthermore, there is a neighborhood I of zero such that $\inf_{u\in I}\zeta'(u) > 0$, and $\sup_{u\in I} E\{|\varphi(e-u)|^2\} < \infty$.*

Assumption A.6 is satisfied by some commonly adopted function φ (see examples in Section 5). Under Assumption A.6, we have the following result which demonstrates that the zero of S_λ is sufficiently close to θ_0.

Proposition 2.2. *Suppose that Assumptions A.1, A.2, A.4–A.6 are satisfied. Let φ satisfy either Assumption A.3 or A.3′. Let $m > 1/2$, $0 \le b \le 1$ and $d > 2b + 1/(2m)$. If $\lambda/n \to 0$, then there exists a unique $\theta_\lambda^b \in \Theta_b$ such that $S_\lambda(\theta_\lambda^b) = 0$ and*

$$\|\theta_\lambda^b - \theta_0\|_b = O\left((\lambda/n)^{(d-b)/2}\right). \tag{2.4}$$

Furthermore, if $0 \le b' < b \le 1$, then $\theta_\lambda^{b'} = \theta_\lambda^b$ under $\|\cdot\|_{b'}$-norm. This is the local uniqueness of the solution to S_λ.

The proof of Proposition 2.2 is given in Appendix B. Proposition 2.2 is similar to Theorem 3.1 of Cox and O'Sullivan (1990). However, unlike the latter, Proposition 2.2 guarantees the uniqueness of θ_λ, i.e., θ_λ is locally fixed when b changes. If we consider θ_λ as the "target" of $\hat{\theta}_{n,\lambda}$, then changing the parameter space Θ_b for $0 \leq b \leq 1$ will not move this target, this means that θ_λ is somewhat "identifiable".

When $d > 2 + 1/(2m)$, it follows by fixing $b = 1$ in Proposition 2.2 that there exists $\theta_\lambda := \theta_\lambda^1 \in \Theta_1$ such that $S_\lambda(\theta_\lambda) = 0$ and $\|\theta_\lambda - \theta_0\|_1 = o(1)$. Consequently, we may assume that any element $\theta_* \in \mathcal{K} \equiv \{\theta_\lambda\} \cup \{\theta_0\}$ satisfies $\|\theta_* - \theta_0\|_{sup} \in I$, with I indicated by Assumption A.6. Recall here that $\lambda = \lambda_n$ is a sequence of penalty parameters indexed by n. On the other hand, if $0 \leq b < 1$, then $\|\theta_\lambda - \theta_0\|_b = O\left((\lambda/n)^{(d-b)/2}\right)$.

Assumption A.7. *$\hat{\theta}_{n,\lambda} \in \Theta_1$ and there exists some positive sequence δ_n such that $\|S_{n,\lambda}(\hat{\theta}_{n,\lambda})\|_1 = o_p(\delta_n)$ and $\|\hat{\theta}_{n,\lambda} - \theta_\lambda\|_1 = o_p(1)$.*

Assumption A.7′. *$\hat{\theta}_{n,\lambda} \in \Theta_1$ and there exists some positive sequences δ_n and s_n, such that $\|S_{n,\lambda}(\hat{\theta}_{n,\lambda})\|_1 = o_p(\delta_n)$, $\|\hat{\theta}_{n,\lambda} - \theta_0\|_1 = o_p(1)$ and $\|\hat{\theta}_{n,\lambda} - \theta_0\|_{sup} = O_p(s_n)$.*

Both Assumptions A.7 and A.7′ do not require that $\hat{\theta}_{n,\lambda}$ is a root of $S_{n,\lambda}$. However, we need the assumption that $\hat{\theta}_{n,\lambda}$ is as smooth as θ_λ. In some special cases, $\hat{\theta}_{n,\lambda}$ can be taken as an approximate MLE, e.g., ϵ-MLE introduced by Wong and Severini (1991), and the consistency of $\hat{\theta}_{n,\lambda}$ can be proved under the assumption of the uniform continuity of the likelihood and the relative compactness of the parameter space (see (Wong and Severini, 1991, Theorem 1)). $\|\hat{\theta}_{n,\lambda} - \theta_\lambda\|_1 = o_p(1)$ thus follows from the consistency of $\hat{\theta}_{n,\lambda}$ and the fact that $\|\theta_\lambda - \theta_0\|_1 = o(1)$.

3. Main results

Our main results consist of two parts. Section 3.1 focuses on the convergence rates of $\hat{\theta}_{n,\lambda}$, while Section 3.2 includes the Bahadur type representations for $\hat{\theta}_{n,\lambda}$.

3.1. Convergence rate

The following result demonstrates when $\hat{\theta}_{n,\lambda}$ attains the optimal convergence rate.

Theorem 3.1. *Let Assumptions A.1–A.7 be satisfied and δ_n be given in Assumption A.7. Assume that $\gamma = 1$ in Assumption A.3, $d > 2 + 1/(2m)$ and $1/(2m) < b < 1 - 1/(2m)$. If the penalty parameter λ is selected such that the following (i)-(iii) hold,*

(i) $\lambda = o(n)$.
(ii) $\delta_n = O(n^{1/2}(\lambda/n)^{(1-1/(2m))/2})$.

(iii) $\max\{(\lambda/n)^{-(1+b)/2}, (\lambda/n)^{-(2b+1/(2m))}\} = o(na_n^{-1})$, *where* $a_n = (n\log\log n)^{\kappa}$ *and* $\kappa = \frac{m}{2m-1}$.

Then

$$\|\hat{\theta}_{n,\lambda} - \theta_0\|_b = O_p\left(n^{-1/2}(\lambda/n)^{-(b+1/(2m))/2} + (\lambda/n)^{(d-b)/2}\right).$$

The proof of Theorem 3.1 is given in Appendix A.

Remark 3.1. The quantities m, b, d have a clear statistical interpretation. m and d respectively represent the degree of smoothness of the estimate $\hat{\theta}_{n,\lambda}$ and the true parameter θ_0, and b indicates the norm under which the bias of $\hat{\theta}_{n,\lambda}$ is measured. In practice, it is possible to regularize the values of d, m, b and λ so that condition (iii) in Theorem 3.1 is satisfied. One way is to let d be relatively large, for instance, d, m, b and λ satisfy

$$\max\left\{\kappa + \frac{m(1+b)}{2md+1}, \kappa + \frac{4mb+1}{2md+1}\right\} < 1 \text{ and } \lambda/n = n^{-2m/(2md+1)}. \quad (3.1)$$

Under (3.1), it can be verified according to Theorem 3.1 that the convergence rate of $\|\hat{\theta}_{n,\lambda} - \theta_0\|_b$ is $n^{-m(d-b)/(2md+1)}$.

We claim that this convergence rate is optimal. This is based on the following considerations. Since θ_0 is md-times differentiable, and $\hat{\theta}_{n,\lambda}^{(mb)}$ is clearly an estimate of $\theta_0^{(mb)}$, where $\theta^{(mb)}$ denotes the mb-order derivative of θ, then by Stone (1982), the optimal convergence rate for $\|\hat{\theta}_{n,\lambda}^{(mb)} - \theta_0^{(mb)}\|_{L^2(\mathbb{I})}$ is $n^{-m(d-b)/(2md+1)}$. On the other hand, we notice that $\|\hat{\theta}_{n,\lambda} - \theta_0\|_b \geq \|\hat{\theta}_{n,\lambda}^{(mb)} - \theta_0^{(mb)}\|_{L^2(\mathbb{I})}$, which means that $\|\hat{\theta}_{n,\lambda} - \theta_0\|_b$ cannot converge faster than $\|\hat{\theta}_{n,\lambda}^{(mb)} - \theta_0^{(mb)}\|_{L^2(\mathbb{I})}$. So $n^{-m(d-b)/(2md+1)}$ is the optimal convergence rate for $\|\hat{\theta}_{n,\lambda} - \theta_0\|_b$, and this rate is achieved when $\lambda/n = n^{-2m/(2md+1)}$.

The technical arguments in the proofs are valid only for a suitably large d. When d is small, which implies that θ_0 is not smooth enough, our approach cannot result in an optimal convergence rate. In such situations, we leave the achievability of the optimal convergence rate as an open problem.

Remark 3.2. Wong and Severini (1991) obtained an optimal convergence rate for the infinite-dimensional nonpenalized MLE under the assumption that the loss function is twice uniformly and continuously differentiable. Here, we briefly introduce their way of proof. They first established an important result which they call "Basic Lemma". This result states that the bias of the estimate is controlled by two terms. Then they obtained the optimal convergence rates by balancing these two terms. However, the second order derivative of the likelihood is needed for the proof of their "Basic Lemma".

In the proof of Theorem 3.1, we cannot establish such "Basic Lemma" as we do not assume that the second order derivative of the likelihood exists. Instead, we first establish Lemma A.1 which states that the variation of $S_{n,\lambda}(\theta)$ is stochastically controlled by the variation of θ. Using Lemma A.1, one can obtain

the desired results without finding the Fréchet derivative of $S_{n,\lambda}$. He and Shao (1996) used the same idea to establish the Bahadur representation for finite-dimensional estimates. Here, we have actually generalized their result to the infinite-dimensional setting using the techniques introduced by Kosorok (2008). We then use Lemma A.1 to establish a quadratic inequality for the term $\|\hat{\theta}_{n,\lambda} - \theta_0\|_b$, and use this inequality to obtain the convergence rate for $\hat{\theta}_{n,\lambda}$.

Remark 3.3. Cox and O'Sullivan (1990) obtained the convergence rate for the penalized estimates under the assumption that the likelihood is three times Fréchet differentiable. They controlled the bias of the estimate by two terms which they called the systematic error and stochastic error. The optimal convergence rates were obtained by balancing these two error terms. Their proof also relies on the sufficient smoothness of penalized likelihood.

One of the commonly used parameter spaces is $H^2(\mathbb{I})$ which corresponds to the case $m = 2$. For instance, Gu and Qiu (1993) have considered this parameter space for purpose of estimating spline densities, which is different from our problem. For this specific situation, it can be shown by Theorem 3.1 that the following result holds.

Corollary 3.2. *Let Assumptions A.1–A.7 be satisfied and δ_n be given in Assumption A.7. Assume that $\gamma = 1$ in Assumption A.3. Suppose we choose b and d such that $1/4 < b < 3/4$ and $d > \max\{9/4, (12b+1)/2, (6b+5)/4\}$. Let $\lambda/n \approx n^{-4/(4d+1)}$ and $\delta_n = O(n^{(2d-1)/(4d+1)})$. Then $\|\hat{\theta}_{n,\lambda} - \theta_0\|_b = O_p(n^{-2(d-b)/(4d+1)})$.*

3.2. Bahadur type representation

The purpose of establishing a Bahadur representation is to approximate an estimate by a sum of independent random variables. Generally speaking, the remainder in a Bahadur representation should be of higher orders than usual statistical bias. This might be the reason why this sort of result attracts so many authors. In this section, we attempt to establish a Bahadur type representation for $\hat{\theta}_{n,\lambda}$. In the following result, we consider the case that φ is Lipschitz, i.e., satisfying Assumption A.3.

Theorem 3.3. *Let the assumptions in Theorem 3.1 be satisfied. Assume further that $\delta_n = O(a_n n^{-m(d-b)/(2md+1)})$. If d, m, b and λ satisfy (3.1), then the following representation holds,*

$$\|\hat{\theta}_{n,\lambda} - \theta_0 + DS_\lambda(\theta_0)^{-1} S_{n,\lambda}(\theta_0)\|_b = O_p\left(a_n n^{-\frac{3md-2mb-m+1}{2md+1}}\right). \tag{3.2}$$

The proof of Theorem 3.3 is based on arguments similar to the proof of Theorem 3.1 and can also be found in Appendix A. However, we should mention that the convergence rate in (3.2) might be suboptimal. By Theorem 3.3, if we fix b, then the convergence rate of the remainder term $\hat{\theta}_{n,\lambda} - \theta_0 + DS_\lambda(\theta_0)^{-1} S_{n,\lambda}(\theta_0)$ under $\|\cdot\|_b$-norm could be arbitrarily close to $n^{-1}(\log\log n)^{1/2}$ when m and d are large enough, which could be even faster than the optimal convergence

rate of $\hat{\theta}_{n,\lambda}$ discussed in Remark 3.1. By Theorems 3.1 and 3.3, when φ is Lipschitz, it is possible to derive the optimal convergence rate and (suboptimal) Bahadur type representation simultaneously from the assumption that $\hat{\theta}_{n,\lambda}$ is consistent. Unfortunately, when φ is not Lipschitz, or even not continuous, the derivation of optimal convergence rate becomes complicated. However, if the convergence rate of $\hat{\theta}_{n,\lambda}$ is known a priori, we may derive a suboptimal Bahadur type representation.

Theorem 3.4. *Let Assumptions A.1, A.2, A.4–A.6 and A.7′ be satisfied.*

(i) If Assumption A.3 is satisfied, then $\hat{\theta}_{n,\lambda}$ satisfies

$$\|\hat{\theta}_{n,\lambda} - \theta_0 + DS_\lambda(\theta_0)^{-1}S_{n,\lambda}(\theta_0)\|_b = O_p(R_n), \tag{3.3}$$

where

$$\begin{aligned} R_n &= (\lambda/n)^{-(b+1/(2m))/2}s_n^2 + n^{-1}(\lambda/n)^{-(1+b)/2} \\ &\quad \left\{(\log\log n)^{1/2}(n^{1/2}s_n^{\gamma-1/(2m)}+1) + \delta_n\right\}. \end{aligned} \tag{3.4}$$

In particular, if $\gamma = 1$ and

$$\begin{aligned} &s_n \approx (n/\log n)^{-md/(2md+1)}, \lambda/n \approx n^{-2m/(2md+1)}, \\ &\delta_n = O\left((\log\log n)^{1/2}(n^{1/2}s_n^{1-1/(2m)}+1)\right), \end{aligned} \tag{3.5}$$

then

$$R_n = n^{-\frac{4md-2mb-1}{2(2md+1)}} + n^{-\frac{(4m-1)(d-1)+2m(1-b)}{2(2md+1)}}(\log n)^{(2m-1)d/(2(2md+1))}(\log\log n)^{1/2}. \tag{3.6}$$

(ii) If Assumption A.3′ is satisfied and $s_n(\log\log n)^m = o(1)$, then (3.3) holds with

$$\begin{aligned} R_n &= (\lambda/n)^{-(b+1/(2m))/2}s_n^2 + n^{-1}(\lambda/n)^{-(1+b)/2} \\ &\quad \left\{n^{1/2}s_n^{1/2-1/(2m)} + s_n^{-1/m} + \delta_n\right\}. \end{aligned} \tag{3.7}$$

In particular, if (3.5) holds and $\delta_n = O(n^{1/2}s_n^{1/2-1/(2m)} + s_n^{-1/m})$, then

$$\begin{aligned} R_n &= n^{-\frac{4md-2mb-1}{2(2md+1)}}(\log n)^{\frac{2md}{2md+1}} \\ &\quad + n^{-\frac{(3m-1)(d-1)+m(1-2b)}{2(2md+1)}}(\log n)^{(m-1)d/(2(2md+1))}. \end{aligned}$$

The proof of Theorem 3.4 is given in Appendix A which relies on Lemma A.3.

Remark 3.4. When $S_{n,\lambda}$ is Fréchet differentiable, we might still be able to obtain a result similar to Theorem 3.4 without using Lemma A.3. However, when $S_{n,\lambda}$ is not Fréchet differentiable, Lemma A.3 plays an important role in the proof of Theorem 3.4. Actually, Lemma A.3 somewhat overcomes the difficulty caused by the nonsmoothness of $S_{n,\lambda}(\theta)$.

Stone (1982) proved that the optimal convergence rate for a nonparametric estimate under the supremum norm is $(n/\log n)^{-md/(2md+1)}$. As demonstrated by Stone (1982), this optimal convergence rate is achievable under certain conditions.

Portnoy (1997) obtained a Bahadur type representation for quantile smoothing spline estimates (see (Portnoy, 1997, Theorem 2.2)). This representation holds at the breakpoints (a discrete set of points) in $\mathbb{I}$. While, the representation (3.3) or (3.7) holds under Sobolev norms. On the other hand, the proof of the result by Portnoy (1997) strongly relies on the property of quantile loss. Actually, the quantile smoothing spline estimate has to be piecewise linear. While, the result in Theorem 3.4 is valid for a general class of φ.

4. Examples

This section contains several illustrative examples. In these examples, we let $h_\mu(x) = \exp\left(2\pi\sqrt{-1}\mu x\right)$, $\mu \in \mathbb{Z}$, be the sequence of orthonormal basis in $L^2(\mathbb{I})$ under $L^2(\mathbb{I})$-norm. Suppose X_i's are independent and uniformly distributed on $\mathbb{I}$. Then Assumptions A.1 and A.2 follow straightforwardly. Assume that the true parameter $\theta_0 \in \Theta_d$ with $d > 2 + 1/(2m)$. In the following examples, we assume that the density function f_e of the noise e satisfies the following assumption.

Assumption A.8. *f_e is symmetric, strictly positive around zero, and having a bounded derivative.*

It can be verified that in these examples, Assumption A.6 follows from Assumption A.8.

Example 4.1. Consider the Huber's loss which corresponds to the following φ_1

$$\varphi_1(s) = \begin{cases} -s, & |s| \le c \\ -c \cdot sgn(s), & |s| > c, \end{cases}$$

where $c > 0$ is a constant. It is easy to verify that Assumption A.3 holds for $C_{\varphi_1} = \gamma = 1$.

Let N_0 be a subset of $C(\mathbb{I})$ such that any $\theta \in N_0$ satisfies $\|\theta - \theta_0\|_{sup} \in I$, where I is indicated in Assumption A.6. Suppose $\theta \in N_0$ satisfies $S(\theta) = 0$. By Fubini's theorem, for any $\mu \in \mathbb{Z}$,

$$\begin{aligned} \langle S(\theta), h_\mu\rangle_1 &= E_X\{\langle \zeta((\theta-\theta_0)(X))K_X, h_\mu\rangle_1\} \\ &= E_X\{\zeta((\theta-\theta_0)(X))h_\mu(X)\} = 0, \end{aligned} \tag{4.1}$$

which implies $\zeta((\theta-\theta_0)(x)) = 0$ for any $x \in \mathbb{I}$. Therefore, $\theta = \theta_0$ by monotonicity of ζ over I. This verifies the identifiability of θ_0, i.e, Assumption A.4.

By Corollary 3.2, when $m = 2$, for some suitable b, d and λ which satisfy the assumptions in Corollary 3.2, $\hat{\theta}_{n,\lambda}$ achieves the optimal convergence rate under $\|\cdot\|_b$, and the following Bahadur type representation holds

$$\|\hat{\theta}_{n,\lambda} - \theta_0 + DS_\lambda(\theta_0)^{-1}S_{n,\lambda}(\theta_0)\|_b = O_p\left(n^{-\frac{10d-12b-5}{3(4d+1)}}(\log\log n)^{2/3}\right).$$

Example 4.2. We consider the negative sign function $\varphi_2(e) := -sgn(e)$, then $\hat{\theta}_{n,\lambda}$ corresponds to the median loss. By an argument similar to the proof of Theorem 3.2 in He and Shao (1996), Assumption A.3′ holds. Similar to Example 4.1, it can be shown that both Assumptions A.4 and A.6 hold. Consequently, when the convergence rate of $\hat{\theta}_{n,\lambda}$ is available, a representation of type (3.3) holds with the convergence rate of R_n indicated by (3.7).

5. Theoretical framework

In the previous sections, we have introduced the main results in this paper and several sufficient conditions which are used to derive these results. In this section, we will give some additional framework which is useful for us to continue the theoretical derivation. The proofs of all the propositions in this section can be found in Appendix B.

5.1. About the kernel function K and operator W

To prove our main results, the function K, which is used to define the equation (2.1), has to satisfy certain desired properties. The following result guarantees the existence of K which satisfies all such desired properties.

Proposition 5.1. *For any $\beta \in (1/(2m), 1]$, there is a bivariate kernel function $K(\cdot,\cdot)$ defined on $\mathbb{I}\times\mathbb{I}$ satisfying the properties:*

(i) $\forall x \in \mathbb{I}$, $K_x := K(x,\cdot) \in \Theta_{2-\beta}$;
(ii) $\forall \theta \in \Theta_\beta$, $\langle K_x, \theta\rangle_1 = \theta(x)$;
(iii) *There is a constant* $C_K > 0$ *such that* $\sup_{x\in\mathbb{I}} \|K_x\|_{2-\beta} \le C_K$.

Throughout this paper, we assume that for some fixed $\beta \in (1/(2m), 1]$, $K : \mathbb{I}\times\mathbb{I} \to \mathbb{R}$ satisfies the properties (i)–(iii) in Proposition 5.1. This requirement for β is related to the dimension of $\mathbb{I}$, which is 1 in the present situation.

The original domain of W is Θ_1. To facilitate the technical proofs, it is useful to extend this domain to be a larger space, say Θ_b for $0 \le b \le 1$. There are two equivalent ways to achieve this extension. One way is based on the fact $Wh_\mu = \frac{\gamma_\mu}{1+\gamma_\mu} h_\mu$, $\forall \mu \in \mathbb{Z}$. To see this, for any μ, since $Wh_\mu \in \Theta_1$, we have $Wh_\mu = \sum_\nu \theta_\nu h_\nu$ for some sequence θ_ν. Then $\gamma_\nu \delta_{\mu\nu} = \langle Wh_\mu, h_\nu\rangle_1 = \theta_\nu(1+\gamma_\nu)$. And hence, $\theta_\mu = \frac{\gamma_\mu}{1+\gamma_\mu}$ and $\theta_\nu = 0$ for $\nu \ne \mu$, which leads to $Wh_\mu = \frac{\gamma_\mu}{1+\gamma_\mu} h_\mu$. Using this fact, one can define $W\sum_\mu \theta_\mu h_\mu = \sum_\mu \frac{\gamma_\mu}{1+\gamma_\mu}\theta_\mu h_\mu$, which is a bounded linear operator from Θ_b to Θ_b for any $0 \le b \le 1$. The other way is through Lemma 2.1 in Cox and O'Sullivan (1990). By such extension, W is a well defined bounded linear operator from Θ_b to Θ_b for any $0 \le b \le 1$.

To conclude this subsection, we assert that, by the above properties of K and W, $S_{n,\lambda}(\theta)$ (defined in Section 2) is exactly the Fréchet derivative of

$$l_n(\theta) = \sum_{i=1}^{n} \rho(Y_i - \theta(X_i)) + \lambda J(\theta),$$

where $J(\theta)$ is defined in the beginning of Section 2 and ρ is the loss function satisfying $\rho' = \varphi$. We leave the details of the verification to the interested readers. So, finding the exact solution to $S_{n,\lambda}(\theta)$ is equivalent to minimizing $l_n(\theta)$. This demonstrates the credibility of estimating θ_0 through finding the approximate solution to $S_{n,\lambda}(\theta)$.

5.2. Fréchet derivatives and their applications

Our technical proofs rely on the exact calculations of the Fréchet derivatives. The following results summarize these calculations.

Proposition 5.2. *Suppose Assumptions A.5 and A.6 are satisfied. Let φ satisfy either Assumption A.3 or Assumption A.3′. If $0 \le b \le 1$, then we have*

(i) For any $\theta \in \Theta_b$, $S(\theta) \in \Theta_{2-\beta}$;

(ii) The Fréchet derivative of S at $\theta \in \Theta_b$ is given by

$$DS(\theta)\xi = E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)K_X\},\ \forall \xi \in \Theta_b. \tag{5.1}$$

Therefore $DS(\theta) \in \mathcal{B}(\Theta_b, \Theta_{2-\beta})$, where $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ is the collection of bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$.

(iii) The second order Fréchet derivative of S at $\theta \in \Theta_b$ is given by

$$D^2S(\theta)\xi\eta = E_X\{\zeta''((\theta-\theta_0)(X))\xi(X)\eta(X)K_X\},\ \forall \xi,\eta \in \Theta_b. \tag{5.2}$$

Therefore $D^2S(\theta) \in \mathcal{B}(\Theta_b, \mathcal{B}(\Theta_b, \Theta_{2-\beta}))$.

Taking $\theta = \theta_0$ in (5.1), we shall get that $\langle DS(\theta_0)\xi, \xi\rangle_1 = \zeta'(0)V(\xi,\xi)$ for any $\xi \in \Theta_1$. Thus, $DS(\theta_0)$ and V define "equivalent" norms. The following result says that when θ is "close" to θ_0, $DS(\theta)$ and V are still "equivalent".

Proposition 5.3. *Suppose Assumptions A.5 and A.6 are satisfied. Let φ satisfy either Assumption A.3 or Assumption A.3′. Suppose $\theta \in C(\mathbb{I})$ with $\|\theta-\theta_0\|_{sup} \in I$, where I is indicated by Assumption A.6, then we have*

$$\inf_{\xi\in\Theta_1/\{0\}} \frac{\langle DS(\theta)\xi,\xi\rangle_1}{V(\xi,\xi)} \ge \inf_{u\in I}\zeta'(u), \quad \text{and} \quad \sup_{\xi\in\Theta_1/\{0\}} \frac{\langle DS(\theta)\xi,\xi\rangle_1}{V(\xi,\xi)} \le \|\zeta'\|_{sup},$$

where $\|\cdot\|_{sup}$ indicates the supremum norm.

An important application of Proposition 5.3 can actually guarantee the invertibility of $DS_\lambda(\theta_0)$. To see this, we note that by Proposition 5.3, if θ is close to θ_0 in supremmum norm, then $DS(\theta)$ is "equivalent" to V. Thus, $DS_\lambda(\theta_0)$ is bounded linear strictly coercive from Θ_1 to Θ_1, i.e., $\langle DS_\lambda(\theta_0)\xi,\xi\rangle_1 \ge cV(\xi,\xi)$ for some positive constant c. By Lax-Milgram theorem (see Adams (1975)), $DS_\lambda(\theta_0)$ has to be bounded invertible from Θ_1 to Θ_1. We denote $DS_\lambda(\theta_0)^{-1}$ to be the inverse operator of $DS_\lambda(\theta_0)$.

By Proposition 5.2, $DS_\lambda(\theta_0) \in \mathcal{B}(\Theta_b, \Theta_b)$ for any $0 \le b \le 1$. Originally, $DS_\lambda(\theta_0)^{-1}$ was defined on Θ_1. However, the following result asserts that the domain of $DS_\lambda(\theta_0)^{-1}$ can actually be extended from Θ_1 to Θ_b for any $0 \le b < 1$.

Proposition 5.4. *For any $0 \le b < 1$, $DS_\lambda(\theta_0)^{-1}$ can be extended as an element in $\mathcal{B}(\Theta_b, \Theta_b)$, and this extension is exactly the inverse of $DS_\lambda(\theta_0) \in \mathcal{B}(\Theta_b, \Theta_b)$.*

By Proposition 5.3, when θ is sufficiently close to θ_0, there is certain similarity between the norms defined by V and $\langle DS(\theta)\cdot, \cdot\rangle_1$. However, the basis $\{h_\mu\}$ might no longer be orthogonal under the latter norm. Therefore, it is desired to seek a new basis which is orthonormal under the latter norm. For this purpose, we define $V_*(\theta, \tilde{\theta}) = \langle DS(\theta_*)\theta, \tilde{\theta}\rangle_1$ for any $\theta_* \in \mathcal{K}$ and $\theta, \tilde{\theta} \in \Theta_1$. Thanks to Proposition 5.3 and Assumption A.1, if $\theta_* \in \mathcal{K}$, then V_* is completely continuous with respect to $V_* + J$. So there are eigenvalues $\{\gamma_{*\mu}\}$ and eigenvectors $\{h_{*\mu}\} \subseteq \Theta_1$ satisfying

$$V_*(h_{*\mu}, h_{*\nu}) = \delta_{\mu\nu}, \quad \text{and } J(h_{*\mu}, h_{*\nu}) = \gamma_{*\mu}\delta_{\mu\nu}, \quad \mu, \nu \in \mathbb{Z}. \tag{5.3}$$

Furthermore, an application of Courant-Weyl's principle (Weinberger, 1974, Theorem 5.2) shows that there are positives c_1 and c_2 (independent of $\theta_* \in \mathcal{K}$) such that

$$c_2\gamma_\mu \le \gamma_{*\mu} \le c_1\gamma_\mu, \ \mu \in \mathbb{Z}. \tag{5.4}$$

For $0 \le b \le 1$, define

$$\langle \theta, \tilde{\theta}\rangle_{*b} = \sum_\mu (1 + \gamma_{*\mu}^b) V_*(\theta, h_{*\mu}) V_*(\tilde{\theta}, h_{*\mu}), \tag{5.5}$$

and let $\|\theta\|_{*b} = \sqrt{\langle \theta, \theta\rangle_{*b}}$. Let Θ_{*b} be the completion of $\{\theta \in \Theta_1 | \, \|\theta\|_{*b} < \infty\}$ under $\|\cdot\|_{*b}$. By Proposition 5.3, it can be verified that $\Theta_{*0} = \Theta_0$ and $\Theta_{*1} = \Theta_1$. Applying the interpolation approach introduced in Cox (1988), $\Theta_{*b} = H^{mb}(\mathbb{I})$, where the equality means set equality and norm equivalence. It can be further shown that this equivalence is uniform for $\theta_* \in \mathcal{K}$, i.e., the following result holds.

Proposition 5.5. *For any $0 \le b \le 1$, there are positive constants d_1 and d_2 (independent of $\theta_* \in \mathcal{K}$) such that*

$$d_1\|\theta\|_b \le \|\theta\|_{*b} \le d_2\|\theta\|_b, \ \theta \in \Theta_b. \tag{5.6}$$

6. Conclusion and future work

Theoretical properties of smoothing spline estimates have been studied in this paper. Precisely, we have demonstrated both optimal convergence rates and Bahadur type representations for a smoothing spline estimate which is an approximate root of an M-estimating equation. In particular, we assume that the moment function φ, which plays a role in characterizing the M-estimating equation, may be either Lipschitz or discontinuous.

We have considered only unidimensional splines. Both the associated editor and one anonymous reviewer have suggested to consider the multidimensional situation. This is a valuable but complicated problem. We conjecture that the

techniques used in unidimensional case can also be applied to multidimensional situations, i.e., the penalty functional becomes

$$J(\theta) = \sum_{\substack{\alpha_1,\ldots,\alpha_p \in \mathbb{N}^* \\ \alpha_1+\cdots+\alpha_p=m}} \frac{m!}{\alpha_1!\cdots\alpha_p!} \int_{[0,1]^p} \left| \frac{\partial^k}{\partial x_1^{\alpha_1}\cdots\partial x_p^{\alpha_p}} \theta(x_1,\ldots,x_p) \right|^2 dx_1\cdots dx_p,$$

where $\mathbb{N}^*$ is the set of nonnegative integers and $\theta \in H^2([0,1]^p)$. However, the generalizations to multidimensional situations might involve more complicated notation and mathematical derivations. We intend to leave this problem as future work.

As suggested by an anonymous reviewer, another issue we intend to explore in future is to generalize the model framework. In this paper, the model where samples were drawn is the classical nonparametric model $y = \theta(x) + e$. This framework restricts the applications and needs to be extended. One extension is to assume that samples (X_i, Y_i) are *iid* drawn from an unknown distribution $P(X, Y)$. This new setting does not require a regression model to link X and Y, and thus, allows more flexibility. One particular example is the support vector machines in which Y takes values 1 or -1 indicating positive and negative classes respectively. A classifier θ, which belongs to a Sobolev space, could be found by minimizing $\sum_{i=1}^n (1 - Y_i\theta(X_i))_+ + \lambda J(\theta)$, with $(a)_+ = a$ if $a < 0$ and 0 otherwise, and with J being the penalty functional. The results in this paper cannot be directly applied to this situation, and we intend to explore their extensions in future work.

Appendix A: Proofs of the results in Section 3

In this section, proofs of the results in Section 3 will be given. Entropy theory will be used in the proofs. Let B_b denote the unit ball in Θ_b, i.e., $B_b = \{\theta \in \Theta_b |\ \|\theta\|_b \leq 1\}$. Define $D(\epsilon, \|\cdot\|_{sup})$ to be the packing number of B_1 under supremum norm, i.e., the maximal number of the elements that can be fit in B_1 while maintaining a distance greater than ϵ between all elements. Then from Cucker and Smale (2002), there is a positive constant $c > 0$ such that $\log_2 D(\epsilon, \|\cdot\|_{sup}) \leq c\epsilon^{-1/m}$. We refer to Zhou (2002) for entropy theory in general reproducing kernel Hilbert spaces. To facilitate the technical proofs, we assume without loss of generality that $\zeta'(0) = 1$. All the arguments can be applied without much revision to the case that $\zeta'(0) \neq 1$.

Before proving Theorem 3.1, we need several technical lemmas. Let $W_i(\theta) = \varphi(Y_i - \theta(X_i))K_{X_i} - S(\theta)$, for $i = 1, 2, \ldots, n$. Let $T_{\lambda,i}(\theta) = W_i(\theta + \theta_\lambda) - W_i(\theta_\lambda)$ and $T^0_{\lambda,i}(\theta) = W_i(\theta + \theta_0) - W_i(\theta_0)$. Define $Z_n(\theta) = \sum_{i=1}^n T_{\lambda,i}(\theta)$ and $Z^0_n(\theta) = \sum_{i=1}^n T^0_{\lambda,i}(\theta)$.

Lemma A.1. *Under Assumption A.3, let* $\kappa = \frac{m\gamma}{2m\gamma-1}$, *then*

(i) $\sup_{\|\theta\|_1 \leq 1} \frac{\|Z_n(\theta)\|_1 + \|Z^0_n(\theta)\|_1}{n^{1/2}\|\theta\|_{sup}^{\gamma-1/(2m)}+1} = O_p((\log\log n)^{1/2})$.

(ii) $\sup_{\|\theta\|_1\le 1} \frac{\|Z_n(\theta)\|_1+\|Z_n^0(\theta)\|_1}{(n\log\log n)^{\kappa}\|\theta\|_{sup}^{\gamma}+1} = O_p(1)$.

Proof of Lemma A.1. Proofs of parts (i) and (ii) are similar, so we focus on part (i) primarily and briefly discuss the proof of part (ii).

Proof of (i). We only prove

$$\sup_{\|\theta\|_1\le 1} \frac{\|Z_n(\theta)\|_1}{n^{1/2}\|\theta\|_{sup}^{\gamma-1/(2m)}+1} = O_p((\log\log n)^{1/2}). \tag{A.1}$$

Indeed, by a similar argument, one can verify that (A.1) is true if Z_n is replaced by Z_n^0, then part (i) holds.

Since $W_i(\theta)$, $i=1,\dots,n$, are *iid* random variables with zero mean, the sequence of sums $\sum_{i=1}^{j} T_{\lambda,i}(\theta)$, $j=1,\dots,n$, is a martingale (with respect to natural filtration) with zero mean. By Assumption A.3, esssup$\|T_{\lambda,i}(\theta)\|_1 \le \tilde{C}_\varphi\|\theta\|_{sup}^{\gamma}$, where $\tilde{C}_\varphi$ is a universal constant and "esssup" is the essential supremum bound of a random variable. By Theorem 3.5 of Pinelis (1994), for any $\theta,\xi\in B_1$ and $t\ge 0$

$$P(\|Z_n(\theta)-Z_n(\xi)\|_1 > t) \le 2\exp\left(-\frac{t^2}{2\tilde{C}_\varphi^2 n\|\theta\|_{sup}^{2\gamma}}\right).$$

By Lemma 8.1 in Kosorok (2008), we have $\|\|Z_n(\theta)-Z_n(\xi)\|_1\|_\psi \le \tilde{C}_\varphi\sqrt{2n}\|\theta-\xi\|_{sup}^{\gamma}$, where $\psi(s)=\exp(s^2)-1$ and $\|\cdot\|_\psi$ is the Orlicz norm defined by $\|\mathcal{X}\|_\psi \equiv \inf\{c>0|E\psi(|\mathcal{X}|/c)\le 1\}$.

By continuity of φ and separability of B_1 under supremum norm, the process Z_n is separable, i.e., there is a countable set $\mathcal{T}\subset B_1$ such that $\sup_{\theta\in B_1}\inf_{\xi\in\mathcal{T}}\times \|Z_n(\theta)-Z_n(\xi)\|_1 = 0$ almost surely. The packing number of B_1 under supremum norm has a bound $\log_2 D(\delta,\|\cdot\|_{sup}) \le c_0\delta^{-1/m}$ for some universal c_0. By the proof of Theorem 8.4 in Kosorok (2008), we can show that there exists some universal constant K such that for any positive δ,

$$\left\| \sup_{\theta\in B_1,\|\theta\|_{sup}\le\delta} \|Z(\theta)\|_1 \right\|_\psi \le K\sqrt{n}\delta^{\gamma-1/(2m)}. \tag{A.2}$$

By (A.2) and Lemma 8.1 in Kosorok (2008), we have the following exponential inequality

$$P\left(\sup_{\theta\in B_1,\|\theta\|_{sup}\le\delta}\|Z_n(\theta)\|_1 > v\right) \le 2\exp\left(-\frac{v^2}{K^2 n\delta^{2\gamma-1/m}}\right),\ \forall v,\delta\ge 0. \tag{A.3}$$

Let $\epsilon = n^{-1/2}$, $Q_\epsilon = [-\log_2\epsilon - 1]$, $\alpha=\gamma-1/(2m)$ and $t=C(\log\log n)^{1/2}$. By Sobolev's inequality (Adams (1975)), there exists a constant c_1 such that for any $\theta\in\Theta_1$, $\|\theta\|_{sup}\le c_1\|\theta\|_1$. So $\sup_{\theta\in B_1,\|\theta\|_{sup}\le c_1}\frac{\|Z_n(\theta)\|_1}{1+n^{1/2}\|\theta\|_{sup}^{\alpha}} \ge \sup_{\theta\in B_1}\frac{\|Z_n(\theta)\|_1}{1+n^{1/2}\|\theta\|_{sup}^{\alpha}}$.

It follows from (A.3) that

$$\begin{aligned}
& P\left(\sup_{\theta\in B_1,\|\theta\|_{sup}\le c_1}\frac{\|Z_n(\theta)\|_1}{1+n^{1/2}\|\theta\|_{sup}^{\alpha}}\ge t\right)\\
\le\ & P\left(\sup_{\theta\in B_1,\|\theta\|_{sup}\le c_1\epsilon^{1/\alpha}}\|Z_n(\theta)\|_1\ge t\right)\\
& +\sum_{l=0}^{Q_\epsilon}P\left(\sup_{\theta\in B_1,c_1(2^l\epsilon)^{1/\alpha}\le\|\theta\|_{sup}\le c_1(2^{l+1}\epsilon)^{1/\alpha}}\frac{\|Z_n(\theta)\|_1}{1+n^{1/2}\|\theta\|_{sup}^{\alpha}}\ge t\right)\\
\le\ & P\left(\sup_{\theta\in B_1,\|\theta\|_{sup}\le c_1\epsilon^{1/\alpha}}\|Z_n(\theta)\|_1\ge t\right)\\
& +\sum_{l=0}^{Q_\epsilon}P\left(\sup_{\theta\in B_1,\|\theta\|_{sup}\le c_1(2^{l+1}\epsilon)^{1/\alpha}}\|Z_n(\theta)\|_1\ge t\{1+c_1^{\alpha}n^{1/2}2^l\epsilon\}\right)\\
\le\ & 2\exp\left(-\frac{t^2}{K^2c_1^{2\alpha}}\right)+2\sum_{l=0}^{Q_\epsilon}\exp\left(-\frac{t^2(1+c_1^{\alpha}n^{1/2}2^l\epsilon)^2}{K^2n[c_1(2^{l+1}\epsilon)^{1/\alpha}]^{2\gamma-1/m}}\right) \qquad \text{(A.4)}\\
\le\ & 2\exp\left(-\frac{t^2}{K^2c_1^{2\alpha}}\right)+2(Q_\epsilon+1)\exp\left(-\frac{t^2}{4K^2}\right)\to 0,
\end{aligned}$$

where the limit is taken by fixing a large C and by letting $n\to\infty$. This completes the proof of part (i).

Proof of (ii). Let $a_n=(n\log\log n)^{\kappa}$, $\epsilon=a_n^{-1}$ and $Q_\epsilon=[-\log_2\epsilon-1]$. By an argument similar to (A.4), one can show that

$$\begin{aligned}
& P\left(\sup_{\theta\in B_1,\|\theta\|_{sup}\le c_1}\frac{\|Z_n(\theta)\|_1}{1+a_n\|\theta\|_{sup}^{\gamma}}\ge t\right)\\
\le\ & 2\exp\left(-\frac{t^2}{K^2c_1^{2\alpha}}\log\log n\right)+2\sum_{l=0}^{Q_\epsilon}\exp\left(-\frac{t^2(1+c_1^{\gamma}a_n2^l\epsilon)^2}{K^2c_1^{2\alpha}n(2^{l+1}\epsilon)^{2-1/(m\gamma)}}\right).
\end{aligned}$$

By Young's inequality, i.e., $ab\le a^p/p+b^q/q$ holds for any $a,b\ge 0$ and positive p,q with $1/p+1/q=1$ (Hardy, Littlewood and Pólya, 1952, page 113), we have $1+u^2\ge \text{const}\cdot u^{2-1/(\gamma m)}$ for any $u>0$. Thus, the above sum is bounded by

$$2\exp\left(-\frac{t^2}{K^2c_1^{2\alpha}}\log\log n\right)+2(Q_\epsilon+1)\exp\left(-\frac{\text{const}\cdot t^2}{K^22^{2\alpha/\gamma}}\log\log n\right).$$

If we preselect t such that $\text{const}\cdot t^2>K^22^{2\alpha/\gamma}$, then the above sum converges to zero as $n\to\infty$. This completes the proof of part (ii). □

Lemma A.2. *Under the assumptions in Proposition 2.2, if $0\le b\le 1$, then there exists a neighborhood N_0 of zero in Θ_1 and λ_0 such that $\forall\lambda\ge\lambda_0$,*

$$\inf_{\eta\in\Theta_b}\inf_{\xi\in N_0}\{\langle DS(\theta_\lambda)\xi,\xi\rangle_1-|\langle D^2S(\theta_\lambda+\eta)\xi\xi,\xi\rangle_1|\}\ge 0.$$

Proof of Lemma A.2. By Proposition 2.2, $\|\theta_\lambda - \theta_0\|_1 \to 0$, which implies that $\|\theta_\lambda - \theta_0\|_{sup} \to 0$. So there exists λ_0 such that $\forall \lambda \geq \lambda_0$ and $x \in [0,1]$, $(\theta_\lambda - \theta_0)(x) \in I$, where I is the interval indicated in Assumption A.6. Thus, there exists a constant $C_0 > 0$ such that $\forall \lambda \geq \lambda_0$ and $x \in [0,1]$, $\zeta'((\theta_\lambda - \theta_0)(x)) > C_0$. Let $C_{\zeta'} = \inf_{u \in I} \zeta'(u)$. Therefore, if $\|\xi\|_1$ is sufficiently small such that $\|\xi\|_{sup} \leq C_0/(C_{\zeta'}\|\zeta''\|_{sup})$, then

$$\begin{aligned} & \langle DS(\theta_\lambda)\xi, \xi\rangle_1 \pm \langle D^2 S(\theta_\lambda + \eta)\xi\xi, \xi\rangle_1 \\ = \;& E\{\zeta'((\theta_\lambda - \theta_0)(x))|\xi(x)|^2 \pm \zeta''((\theta_\lambda + \eta - \theta_0)(x))|\xi(x)|^3\} \\ = \;& E\{|\xi(x)|^2(\zeta'((\theta_\lambda - \theta_0)(x)) \pm \zeta''((\theta_\lambda + \eta - \theta_0)(x))|\xi(x)|)\} \\ \geq \;& 0, \end{aligned}$$

where C_K is given in Proposition 5.1. This completes the proof of Lemma A.2. □

The following lemma is used to prove Theorem 3.4 part (ii).

Lemma A.3. *Let assumptions A.3′, A.5, A.6, and A.7′ be satisfied, and $\hat{\theta}_{n,\lambda}$ and s_n be given in A.7′. Further assume that $s_n(\log\log n)^m = o(1)$. Then for any $A > 0$, the empirical process $Z_n(\theta)$ has the following stochastic bound*

$$\sup_{\|\theta\|_1 \leq 1, \|\theta\|_{sup} \leq A s_n} \|Z_n^0(\theta)\|_1 = O_p\left(n^{1/2} s_n^{1/2 - 1/(2m)} + s_n^{-1/m}\right). \tag{A.5}$$

Proof of Lemma A.3. Without out loss of generality, we assume $A = 1$. For $A \neq 1$, similar arguments can be performed by factoring out A. We use a chaining argument (see Alexander (1984)) to prove this result. $\|\cdot\|_{sup}$ entropy on B_1 will be used. Let $\epsilon = n^{-z}$ with $z > \max\{2m/(2+m), m\}$, and $k = [\log_2 s_n \epsilon^{-1}]$. Then it can be checked that $n\epsilon^{1/2} \ll n^{1/2}\epsilon^{1/2 - 1/(2m)} + \epsilon^{-1/m}$. Define $u_n = n^{1/2}(s_n)^{1/2 - 1/(2m)} + (s_n)^{-1/m}$. For $j = 0, \ldots, k$, let $\mathcal{T}_j$ be a subset in B_1 with cardinality $N_j := D(2^{k-j}\epsilon, \|\cdot\|_{sup})$ such that the distance between any two elements in $\mathcal{T}_k$ is greater than $2^{k-j}\epsilon$, where $D(\epsilon, \|\cdot\|_{sup})$ denotes the packing number. By the discussions in the beginning of Section 6, $\log_2 N_j \leq c(2^{k-j}\epsilon)^{-1/m}$ for some constant $c > 0$. Then the $(2^{k-j}\epsilon)$-balls with centers in $\mathcal{T}_j$ will cover B_1. Consequently, any $\xi \in B_1$ corresponds a chain $\{\xi_j\}_{j=0}^k$ which satisfies $\xi_j \in \mathcal{T}_j$, $\|\xi - \xi_k\|_{sup} \leq \epsilon$ and $\|\xi_j - \xi_{j+1}\|_{sup} \leq 2^{k-j}\epsilon$. It follows that if $\|\xi\|_{sup} \leq s_n$, then $\|\xi_0\|_{sup} \leq 2^{k+1}\epsilon + s_n = s_n(1 + o(1))$. Therefore,

$$\begin{aligned} \sup_{\xi \in B_1, \|\xi\|_{sup} \leq s_n} \|Z_n^0(\xi)\|_1 \;\leq\; & \max_{\xi_k \in \mathcal{T}_k} \sup_{\|\xi - \xi_k\|_{sup} \leq \epsilon} \|Z_n^0(\xi) - Z_n^0(\xi_k)\|_1 \\ & + \sum_{j=0}^{k-1} \max_{\xi_j \in \mathcal{T}_j, \xi_{j+1} \in \mathcal{T}_{j+1} \atop \|\xi_j - \xi_{j+1}\|_{sup} \leq 2^{k-j}\epsilon} \|Z_n^0(\xi_j) - Z_n^0(\xi_{j+1})\|_1 \\ & + \max_{\xi_0 \in \mathcal{T}_0 \atop \|\xi_0\|_{sup} \leq 2^{k+1}\epsilon + s_n} \|Z_n^0(\xi_0)\|_1. \end{aligned}$$

To simplify the notation, we define

$$t_0 = Cu_n,$$
$$t_{j+1} = C\left\{(2^{k-j}\epsilon)^{1/2-1/(2m)}n^{1/2} + (2^{k-j}\epsilon)^{-1/m}\right\},\ j = 0,\ldots,k.$$

It is easy to see that $\sum_{j=0}^{k+1} t_j = C(n^{1/2}(2^k\epsilon)^{1/2-1/(2m)} + (2^k\epsilon)^{-1/m}) = t_0(1 + o(1))$. And hence,

$$\begin{aligned}
& P\left(\sup_{\xi\in B_1,\|\xi\|_{sup}\le s_n}\|Z_n^0(\xi)\|_1 > 2t_0\right)\\
\le\ & P\left(\max_{\xi_k\in\mathcal{T}_k}\sup_{\|\xi-\xi_k\|_{sup}\le\epsilon}\|Z_n^0(\xi) - Z_n^0(\xi_k)\|_1 > 2t_{k+1}\right)\\
& +\sum_{j=0}^{k-1} P\left(\max_{\substack{\xi_j\in\mathcal{T}_j,\xi_{j+1}\in\mathcal{T}_{j+1}\\ \|\xi_j-\xi_{j+1}\|_{sup}\le 2^{k-j}\epsilon}}\|Z_n^0(\xi_j) - Z_n^0(\xi_{j+1})\|_1 > t_{j+1}\right)\\
& +P\left(\max_{\substack{\xi_0\in\mathcal{T}_0\\ \|\xi_0\|_{sup}\le 2^{k+1}\epsilon+s_n}}\|Z_n^0(\xi_0)\|_1 > t_0\right)\\
\le\ & \sum_{\xi_k\in\mathcal{T}_k} P\left(\sup_{\|\xi-\xi_k\|_{sup}\le\epsilon}\|Z_n^0(\xi) - Z_n^0(\xi_k)\|_1 > 2t_{k+1}\right)\\
& +\sum_{j=0}^{k-1}\sum_{\substack{\xi_j\in\mathcal{T}_j,\xi_{j+1}\in\mathcal{T}_{j+1}\\ \|\xi_j-\xi_{j+1}\|_{sup}\le 2^{k-j}\epsilon}} P\left(\|Z_n^0(\xi_j) - Z_n^0(\xi_{j+1})\|_1 > t_{j+1}\right)\\
& +\sum_{\substack{\xi_0\in\mathcal{T}_0\\ \|\xi_0\|_{sup}\le 2^{k+1}\epsilon+s_n}} P\left(\|Z_n^0(\xi_0)\|_1 > t_0\right)\\
\equiv\ & \sum_{\xi_k\in\mathcal{T}_k} P_{\xi_k} + \sum_{j=0}^{k-1}\sum_{\substack{\xi_j\in\mathcal{T}_j,\xi_{j+1}\in\mathcal{T}_{j+1}\\ \|\xi_j-\xi_{j+1}\|_{sup}\le 2^{k-j}\epsilon}} P_{\xi_j,\xi_{j+1}} + \sum_{\substack{\xi_0\in\mathcal{T}_0\\ \|\xi_0\|_{sup}\le 2^{k+1}\epsilon+s_n}} P_{\xi_0}.
\end{aligned}$$

We first complete the approximations of P_{ξ_k}. Denote $V_{\xi_k,i} = \sup_{\|\xi-\xi_k\|_{sup}\le\epsilon}\times \|T_{\lambda,i}^0(\xi) - T_{\lambda,i}^0(\xi_k)\|_1$, where $T_{\lambda,i}^0(\theta)$ is defined in the beginning of Section 6. Since φ is bounded, $M := \max_{\xi_k}\max_{1\le i\le n} V_{\xi_k,i} < \infty$. By the selection of ϵ and Assumption A.3′, $\max_{\xi_k}\sum_{i=1}^n E\{V_{\xi_k,i}\} = O(n\epsilon^{1/2}) = o(t_{k+1})$. Therefore, by Freedman's inequality (Freedman (1975)),

$$\begin{aligned}
P_{\xi_k} &\le P\left(\sum_{i=1}^n [V_{\xi_k,i} - E\{V_{\xi_k,i}\}] \ge t_{k+1}\right)\\
&\le 2\exp\left(-\frac{C^2(\epsilon^{1/2-1/(2m)})^2 n}{4C_{\mathcal{A}}n\epsilon}\right) + 2\exp\left(-\frac{C^2\epsilon^{-1/m}}{8M}\right)\\
&= 2\exp\left(-\frac{C^2\epsilon^{-1/m}}{4C_{\mathcal{A}}}\right) + 2\exp\left(-\frac{C^2\epsilon^{-1/m}}{8M}\right),
\end{aligned}$$

where $C_{\mathcal{A}}$ is constant defined in Assumption A.3′. Thus, by choosing a large C and letting $n \to \infty$, $\sum_{\xi_k \in \mathcal{T}_k} P_{\xi_k} \to 0$.

Before proceeding further, we introduce the following variant of Bernstein type inequality, which is proved by Yurinskiĭ (1976).

Lemma A.4. *Let $\mathcal{H}$ be a Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$. If $\xi_1, \ldots, \xi_n \in \mathcal{H}$ are i.i.d. random elements satisfying $E\xi_i = 0$ and $\|\xi\|_{\mathcal{H}} \le M$ a.s., then for any $x > 0$,*

$$\begin{aligned} P\left(\|\sum_{i=1}^{n} \xi_i\|_{\mathcal{H}} \ge x\right) &\le 2\exp\left(-\frac{x^2}{2[nE\{\|\xi_1\|_{\mathcal{H}}^2\} + Mx]}\right) \\ &\le 2\exp\left(-\frac{x^2}{4nE\{\|\xi_1\|_{\mathcal{H}}^2\}}\right) + 2\exp\left(-\frac{x}{4M}\right). \end{aligned}$$

Next, we complete the approximations of $P_{\xi_j,\xi_{j+1}}$ and P_{ξ_0}. Let M' be the bound for $\max_{0\le j\le k-1} \max_{1\le i\le n} \|T_{\lambda,i}(\xi_j) - T_{\lambda,i}(\xi_{j+1})\|_1$. By Lemma A.4,

$$\begin{aligned} P_{\xi_j,\xi_{j+1}} &\le 2\exp\left(-\frac{t_{j+1}^2}{4C_{\mathcal{A}} n(2^{k-j}\epsilon)}\right) + 2\exp\left(-\frac{t_{j+1}}{4M'}\right) \\ &\le 2\exp\left(-\frac{C^2(2^{k-j}\epsilon)^{-1/m}}{4C_{\mathcal{A}}}\right) + 2\exp\left(-\frac{C(2^{k-j}\epsilon)^{-1/m}}{4M'}\right). \end{aligned}$$

Therefore, by choosing $C > \max\{[4(c+1)C_{\mathcal{A}}]^{1/2}, 4M'(c+1)\}$, we get that

$$\begin{aligned} &\sum_{j=0}^{k-1} \sum_{\substack{\xi_j\in\mathcal{T}_j,\xi_{j+1}\in\mathcal{T}_{j+1} \\ \|\xi_j-\xi_{j+1}\|_{sup}\le 2^{k-j}\epsilon}} P_{\xi_j,\xi_{j+1}} \\ \le\ &\sum_{j=0}^{k-1} N_j \left\{2\exp\left(-\frac{C^2(2^{k-j}\epsilon)^{-1/m}}{4C_{\mathcal{A}}}\right) + 2\exp\left(-\frac{C(2^{k-j}\epsilon)^{-1/m}}{4M'}\right)\right\} \\ \le\ &\sum_{j=0}^{k-1} \exp\left(-(2^{k-j}\epsilon)^{-1/m}\right) \\ \le\ &2k\exp\left(-s_n^{-1/m}\right) \to 0, \end{aligned} \tag{A.6}$$

where the last limit holds when $n \to \infty$. By an argument similar to (A.6), we can show that as $n \to \infty$

$$\sum_{\substack{\xi_0\in\mathcal{T}_0 \\ \|\xi_0\|_{sup}\le 2^{k+1}\epsilon+s_n}} P_{\xi_0} \le 2\exp\left(-s_n^{-1/m}\right) \to 0.$$

This completes the proof of Lemma A.3. □

Proof of Theorem 3.1. For simplicity, denote $\theta = \hat{\theta}_{n,\lambda} - \theta_\lambda$. It follows by Assumption A.7 that $\|\theta\|_1 \leq 1$ with a large probability as n is large. By Lemma A.1 (ii), for large n,

$$\|S_{n,\lambda}(\theta+\theta_\lambda) - S_{n,\lambda}(\theta_\lambda) - S_\lambda(\theta+\theta_\lambda) + S_\lambda(\theta_\lambda)\|_1 \leq M(a_n\|\theta\|_{sup}+1). \quad \text{(A.7)}$$

It follows by Lemma A.2 that,

$$\langle DS(\theta_\lambda)\theta,\theta\rangle_1 + \int_0^1\int_0^1 s\langle D^2S(\theta_\lambda+s's\theta)\theta\theta,\theta\rangle_1 ds'ds \geq \frac{1}{2}\langle DS(\theta_\lambda)\theta,\theta\rangle_1. \quad \text{(A.8)}$$

Before proceeding further, we list several approximation results (Claims I–IV in what follows) and their brief verification.

Claim I: $\|S_{n,\lambda}(\theta_\lambda)\|_1 = O_p(n^{1/2})$.

To prove Claim I, it is sufficient to show $\|S_{n,\lambda}(\theta_\lambda) - S_\lambda(\theta_\lambda)\|_1 = O_p(n^{1/2})$. For any $\mu \in \mathbb{Z}$, by Cauchy's inequality,

$$\begin{aligned} |V(S_{n,\lambda}(\theta_\lambda) - S_\lambda(\theta_\lambda), h_\mu)|^2 &= |E_Z\{S_{n,\lambda}(\theta_\lambda)(Z)h_\mu(Z) - S_\lambda(\theta_\lambda)h_\mu(Z)\}|^2 \\ &\leq E_Z\{|S_{n,\lambda}(\theta_\lambda)(Z)h_\mu(Z) - S_\lambda(\theta_\lambda)h_\mu(Z)|^2\}. \end{aligned}$$

Let $E\{\cdot\}$ and $Var\{\cdot\}$ denote the expectation and variance w.r.t. (X_i, Y_i)'s, then it follows by Fubini's theorem and Assumption A.5 that

$$\begin{aligned} & E\{|V(S_{n,\lambda}(\theta_\lambda) - S_\lambda(\theta_\lambda), h_\mu)|^2\} \\ \leq\ & E\{E_Z\{|S_{n,\lambda}(\theta_\lambda)(Z)h_\mu(Z) - S_\lambda(\theta_\lambda)(Z)h_\mu(Z)|^2\}\} \\ =\ & E_Z\{E\{|S_{n,\lambda}(\theta_\lambda)(Z)h_\mu(Z) - S_\lambda(\theta_\lambda)(Z)h_\mu(Z)|^2\}\} \\ =\ & nE_Z\{Var\{\varphi(Y-\theta_\lambda(X))K_X(Z)h_\mu(Z)\}\} \leq C_0 n, \end{aligned}$$

with some constant C_0 independent of μ. Therefore, by Fubini's theorem

$$\begin{aligned} & E\{\|S_{n,\lambda}(\theta_\lambda) - S_\lambda(\theta_\lambda)\|_1^2\} \\ =\ & E\left\{\sum_\mu \|V(S_{n,\lambda}(\theta_\lambda) - S_\lambda(\theta_\lambda), h_\mu)|^2(1+\gamma_\mu)\right\} \\ =\ & \sum_\mu E\{|V(S_{n,\lambda}(\theta_\lambda) - S_\lambda(\theta_\lambda), h_\mu)|^2\}(1+\gamma_\mu) = O(n), \end{aligned}$$

which proves Claim I. □

Using a similar argument in the proof of Proposition 5.4 (see Appendix B), we can show that for any $\theta_* = \theta_\lambda \in \mathcal{K}$, $DS_\lambda(\theta_*)^{-1}$ is a well defined element in $\mathcal{B}(\Theta_b, \Theta_b)$ for and $1/(2m) < b < 1 - 1/(2m)$.

Claim II: For any $\xi \in \Theta_1$, $\|DS_\lambda(\theta_\lambda)\xi\|_1 \geq C_3 n(\lambda/n)^{(1+b)/2}\|\xi\|_b$.

Denote $\theta_* = \theta_\lambda$. Let $\xi, \eta \in \Theta_1$ have the expansions $\xi = \sum_\mu \xi_\mu h_{*\mu}$ and $\eta = \sum_\mu \eta_\mu h_{*\mu}$. It follows from (5.3) that $\langle D_\lambda(\theta_*)h_{*\mu}, h_{*\nu}\rangle = (n+\lambda\gamma_*)\delta_{\mu\nu}$. By

the restriction $\sum_\mu \eta_\mu^2(1+\mu)=1$ and Cauchy's inequality, it can be shown that

$$\begin{aligned}\|DS_\lambda(\theta_*)\xi\|_1 &= \sup_{\|\eta\|_1=1}|\langle DS_\lambda(\theta_*)\xi,\eta\rangle_1| \\ &= \sqrt{\sum_\mu \frac{(n+\lambda\gamma_{*\mu})^2}{1+\gamma_\mu}\xi_\mu^2} \\ &\geq \sqrt{\sum_\mu \frac{(n+\lambda\gamma_\mu)^2}{(1+\gamma_\mu)(1+\gamma_\mu^b)}\xi_\mu^2(1+\gamma_\mu^b)} \\ &\geq n(\lambda/n)^{(1+b)/2}\|\xi\|_b,\end{aligned}$$

where the supremum is attained when $\eta_\mu=\xi_\mu(n+\lambda\gamma_{*\mu})/(t\sqrt{1+\gamma_\mu})$ and $t=\sqrt{\sum_\mu \xi_\mu^2(n+\lambda\gamma_{*\mu})^2/(1+\gamma_\mu)}$. This completes the proof of Claim II. □

Since $DS_\lambda(\theta_*)^{-1}$ is self-adjoint, we have the following expansion by Cox (1988),

$$\begin{aligned}DS_\lambda(\theta_*)^{-1}\xi &= \sum_\mu \langle DS_\lambda(\theta_*)^{-1}\xi, DS(\theta_*)h_{*\mu}\rangle_1 h_{*\mu} \\ &= \sum_\mu \langle \xi, DS_\lambda(\theta_*)^{-1}DS(\theta_*)h_{*\mu}\rangle_1 h_{*\mu} \\ &= \sum_\mu (n+\lambda\gamma_{*\mu})^{-1}\langle \xi, h_{*\mu}\rangle_1 h_{*\mu}. \qquad \text{(A.9)}\end{aligned}$$

Claim III: $\|DS_\lambda(\theta_\lambda)^{-1}D^2S_\lambda(\theta_\lambda+s's\theta)\theta\theta\|_b \leq C_6\|\theta\|_{sup}^2(\lambda/n)^{-(b+1/(2m))/2}$.

Claim III can be proved by replacing θ_0 by $\theta_*=\theta_\lambda$ and $\{h_\mu\}$ by $\{h_{*\mu}\}$ in the equations (B.5) and (B.6) in Appendix B, and by an application of (A.9). □

Claim IV: $\|DS_\lambda(\theta_\lambda)^{-1}S_{n,\lambda}(\theta_\lambda)\|_b = O_p(n^{-1/2}(\lambda/n)^{-(b+1/(2m))/2})$.

By the expansion in (A.9), we get that

$$\|DS_\lambda(\theta_\lambda)^{-1}S_{n,\lambda}(\theta_\lambda)\|_b^2 = \sum_\mu |\langle S_{n,\lambda}(\theta_\lambda), h_{*\mu}\rangle_1|^2(n+\lambda\gamma_{*\mu})^{-2}(1+\gamma_{*\mu}^b). \qquad \text{(A.10)}$$

By independence between X_i and e_i, Assumption A.3 and that $\|\theta_\lambda-\theta_0\|_{sup}$ is bounded uniformly for λ, we have the following approximation

$$\begin{aligned}& E\{|\langle S_{n,\lambda}(\theta_\lambda), h_{*\mu}\rangle_1|^2\} \\ =\ & E\{|\sum_\mu [\varphi(Y_i-\theta_\lambda(X_i))h_{*\mu}(X_i) - E\varphi(Y-\theta_\lambda(X))h_{*\mu}(X)]|^2\} \\ \leq\ & nE\{|\varphi(Y-\theta_\lambda(X))h_{*\mu}(X)|^2\} \\ \leq\ & 2nE\{|\varphi(Y-\theta_\lambda(X))-\varphi(e)|^2|h_{*\mu}(X)|^2\} + 2nE\{|\varphi(e)|^2|h_{*\mu}(X)|^2\} \\ \leq\ & 2nC_\varphi^2 E\{|\theta_\lambda(X)-\theta_0(X)|^2|h_{*\mu}(X)|^2\} + 2nE\{|\varphi(e)|^2\}\{|h_{*\mu}(X)|^2\} = O(n).\end{aligned}$$

Therefore, Claim IV holds. □

Next, we will use Claims I–IV and several relevant assumptions to establish an inequality for $\|\theta\|_b$, and find the optimal convergence rate.

By Claim I, the assumptions that $\|\theta\|_1 = o_p(1)$ and $\|S_{n,\lambda}(\theta+\theta_\lambda)\|_1 = o_p(\delta_n)$, and that both $n^{1/2}$ and δ_n are controlled by a_n, we can verify that with a large probability

$$\begin{aligned} \|S_\lambda(\theta+\theta_\lambda) - S_\lambda(\theta_\lambda)\|_1 &\leq 2M(a_n\|\theta\|_{sup} + \delta_n + n^{1/2} + 1) \\ &\leq 2M(a_n + \delta_n + n^{1/2}) \leq 6Ma_n. \end{aligned} \tag{A.11}$$

By (A.8), the expansion $\theta = \sum_\mu \theta_\mu h_\mu$ in Θ_1, and Assumption A.5, we have

$$\begin{aligned} & \langle S_\lambda(\theta+\theta_\lambda) - S_\lambda(\theta_\lambda), \theta\rangle_1 \\ = \;& n\langle S(\theta+\theta_\lambda) - S(\theta_\lambda), \theta\rangle_1 + \lambda\langle W\theta, \theta\rangle_1 \\ = \;& n\{\langle DS(\theta_\lambda)\theta, \theta\rangle_1 + \int_0^1\int_0^1 s\langle D^2S(\theta_\lambda + s's\theta)\theta\theta, \theta\rangle_1 ds'ds\} + \lambda\langle W\theta, \theta\rangle_1 \\ \geq \;& \frac{n}{2}\langle DS(\theta_\lambda)\theta, \theta\rangle_1 + \lambda\langle W\theta, \theta\rangle_1 \\ \approx \;& \frac{n}{2}\sum_\mu \theta_\mu^2 + \lambda\sum_\mu \theta_\mu^2\gamma_\mu \\ = \;& \sum_\mu \theta_\mu^2(n/2 + \lambda\gamma_\mu) \\ \approx \;& \sum_\mu \theta_\mu^2(1+\gamma_\mu^b)\left(\frac{n/2+\lambda\gamma_\mu}{1+\gamma_\mu^b}\right) \\ \geq \;& C_1 n(\lambda/n)^b\|\theta\|_b^2, \end{aligned} \tag{A.12}$$

where the last inequality follows from Young's inequality. From (A.11) and (A.12), with a large probability,

$$\|\theta\|_b \leq C_2 a_n^{1/2} n^{-1/2}(\lambda/n)^{-b/2}. \tag{A.13}$$

According to Claim II, Taylor's expansion of $S_\lambda(\theta+\theta_\lambda)$ at θ_λ, and the exchangeability between $DS_\lambda(\theta_\lambda)^{-1}$ and the integral, we have

$$\begin{aligned} & \|S_{n,\lambda}(\theta_\lambda) + S_\lambda(\theta+\theta_\lambda) - S_\lambda(\theta_\lambda)\|_1 \\ = \;& \|DS_\lambda(\theta_\lambda)[DS_\lambda(\theta_\lambda)^{-1}S_{n,\lambda}(\theta_\lambda) + \theta \\ & + \int_0^1\int_0^1 sDS_\lambda(\theta_\lambda)^{-1}D^2S_\lambda(\theta_\lambda + s's\theta)\theta\theta ds'ds]\|_1 \\ \geq \;& C_4 n(\lambda/n)^{(1+b)/2}\|DS_\lambda(\theta_\lambda)^{-1}S_{n,\lambda}(\theta_\lambda) + \theta \\ & + \int_0^1\int_0^1 sDS_\lambda(\theta_\lambda)^{-1}D^2S_\lambda(\theta_\lambda + s's\theta)\theta\theta ds'ds\|_b. \end{aligned} \tag{A.14}$$

By (A.7) and (A.14), there is some constant C_5 such that with a large probability,

$$\begin{aligned} & \|DS_\lambda(\theta_\lambda)^{-1}S_{n,\lambda}(\theta_\lambda)+\theta+\int_0^1\int_0^1 sDS_\lambda(\theta_\lambda)^{-1}D^2S_\lambda(\theta_\lambda+s's\theta)\theta\theta ds'ds\|_b \\ \leq\ & C_5n^{-1}(\lambda/n)^{-(1+b)/2}(a_n\|\theta\|_{sup}+\delta_n). \end{aligned} \tag{A.15}$$

By (A.15) and Claims III and IV, with a large probability,

$$\begin{aligned} \|\theta\|_b \ \leq\ & C_7\{a_nn^{-1}(\lambda/n)^{-(1+b)/2}\|\theta\|_{sup}+n^{-1}(\lambda/n)^{-(1+b)/2}\delta_n \\ & +\|\theta\|_b^2(\lambda/n)^{-(b+1/(2m))/2}+n^{-1/2}(\lambda/n)^{-(b+1/(2m))/2}\}. \end{aligned} \tag{A.16}$$

By assumption (ii) in Theorem 3.1, equation (A.16) and $b>1/(2m)$ (which implies that $\|\theta\|_{sup}\leq \text{const}\cdot\|\theta\|_b$), when n is sufficiently large,

$$\|\theta\|_b/2\leq C_7(n^{-1/2}(\lambda/n)^{-(b+1/(2m))/2}+\|\theta\|_b^2(\lambda/n)^{-(b+1/(2m))/2}). \tag{A.17}$$

Solving inequality (A.17), we get either

$$\|\theta\|_b\geq(1/(4C_7))(\lambda/n)^{(b+1/(2m))/2},\ \text{or} \tag{A.18}$$

$$\begin{aligned} \|\theta\|_b \ \leq\ & (1/(4C_7))(\lambda/n)^{(b+1/(2m))/2}-\sqrt{1/(16C_7^2)(\lambda/n)^{b+1/(2m)}-n^{-1/2}} \\ \leq\ & 4C_7n^{-1/2}(\lambda/n)^{-(b+1/(2m))/2}. \end{aligned} \tag{A.19}$$

However, (A.18) is rejected by (A.13) and assumption (iii) in Theorem 3.1. Thus, the upper bound for $\|\theta\|_b$ is given by (A.19). Combining (A.19) and $\|\theta_\lambda-\theta_0\|_b=O((\lambda/n)^{(d-b)/2})$ (Proposition 2.2), we get that

$$\|\hat\theta_{n,\lambda}-\theta_0\|_b=O_p\left((\lambda/n)^{(d-b)/2}+n^{-1/2}(\lambda/n)^{-(b+1/(2m))/2}\right).$$

This completes the proof of Theorem 3.1. □

Proof of Theorem 3.3. By a reexamination of the proofs, Lemma A.2 still holds if we replace θ_λ by θ_0. Therefore, by Lemma A.1 (ii), (A.7) holds when θ_λ is replaced by θ_0 and θ is replaced by $\hat\theta_{n,\lambda}-\theta_0$. By the assumption that $\|S_{n,\lambda}(\theta+\theta_0)\|_1=o_p(\delta_n)$, we get that with a large probability,

$$\|S_{n,\lambda}(\theta_0)+S_\lambda(\theta+\theta_0)-S_\lambda(\theta_0)\|_1\leq M(a_n\|\theta\|_{sup}+\delta_n). \tag{A.20}$$

It follows by taking $\theta_*=\theta_0$ in Claim II and an argument similar to (A.14) that for some large constant C',

$$\begin{aligned} & \|S_{n,\lambda}(\theta_0)+S_\lambda(\theta+\theta_0)-S_\lambda(\theta_0)\|_1 \\ \geq\ & C'n(\lambda/n)^{(1+b)/2}\|DS_\lambda(\theta_0)^{-1}S_{n,\lambda}(\theta_0)+\theta \\ & +\int_0^1\int_0^1 sDS_\lambda(\theta_0)^{-1}D^2S_\lambda(\theta_0+s's\theta)\theta\theta ds'ds\|_b. \end{aligned}$$

An examination of the proofs reveals that Claim III holds if we replace θ_λ by θ_0, that is, for some large constant C'', if n is sufficiently large, then for any $0 \le s, s' \le 1$,

$$\|DS_\lambda(\theta_0)^{-1}D^2S_\lambda(\theta_0+s's\theta)\theta\theta\|_b \le C''\|\theta\|_b^2(\lambda/n)^{-(b+1/(2m))/2}. \tag{A.21}$$

Following (A.20) and (A.21), and $\|\hat\theta_{n,\lambda}-\theta_0\|_b = O_p((\lambda/n)^{(d-b)/2})$ (Theorem 3.1), there exists some constant $C>0$ such that with a large probability,

$$\begin{aligned} & \|\theta + DS_\lambda(\theta_0)^{-1}S_{n,\lambda}(\theta_0)\|_b \\ \le\ & C(n^{-1}(\lambda/n)^{-(1+b)/2}a_n\|\theta\|_b + n^{-1}(\lambda/n)^{-(1+b)/2}\delta_n \\ & +(\lambda/n)^{-(b+1/(2m))/2}\|\theta\|_b^2) \\ \le\ & Cn^{-1}a_n(\lambda/n)^{-(1+2b-d)/2}\left(1+a_n^{-1}(\lambda/n)^{-(d-b)/2}\delta_n\right. \\ & \left.+na_n^{-1}(\lambda/n)^{(1+d-b-1/(2m))/2}\right). \end{aligned}$$

Since $m>1$ and $d>2b+1/(2m)$, it can be verified that $1-\kappa-\frac{m(1+d-b-1/(2m))}{2md+1} < 0$. Therefore, by $\lambda/n \approx n^{-2m/(2md+1)}$, we have $na_n^{-1}(\lambda/n)^{(1+d-b-1/(2m))/2} = o(1)$. Since $\delta_n = O(a_n n^{-m(d-b)/(2md+1)})$, we get that

$$a_n^{-1}(\lambda/n)^{-(d-b)/2}\delta_n \approx \delta_n a_n^{-1}n^{m(d-b)/(2md+1)} = O(1).$$

Consequently, for large n and with a large probability

$$\|\theta+DS_\lambda(\theta_0)^{-1}S_{n,\lambda}(\theta_0)\|_b \le 2Cn^{-1}(\lambda/n)^{-(1+2b-d)/2}a_n = 2Ca_n n^{-\frac{3md-2mb-m+1}{2md+1}}.$$

This completes the proof of Theorem 3.3. □

Proof of Theorem 3.4. We only briefly prove part (i) since (ii) can be shown similarly by using Lemma A.3. Let $\theta = \hat\theta_{n,\lambda}-\theta_0$. By Assumption A.7′, $\|S_{n,\lambda}(\theta+\theta_0)\|_1 = o_p(\delta_n)$, and with a large probability, $\|\theta\|_1 \le 1$. By Lemma A.1 (i), and arguments similar to (A.20) and (A.21), it can be shown that (3.3) holds with R_n given by (3.4). Thus, (3.6) follows from a direct calculation. □

Appendix B: Proofs of Propositions in Section 5

In this section, we list all the proofs of propositions in Section 5. Hereafter, let B_b denote the unit ball in Θ_b, i.e., $B_b = \{\theta\in\Theta_b|\ \|\theta\|_b \le 1\}$.

Proof of Proposition 5.1. The assumption $\beta > 1/(2m)$ implies that Θ_β is a reproducing kernel Hilbert space (Berlinet and Thomas-Agnan, 2004, Theorem 132). Let $\tilde K(\cdot,\cdot)$ be a reproducing kernel on Θ_β, i.e., for any $x\in\mathbb{I}$, $\tilde K_x$ is an element in Θ_β and for any $\theta\in\Theta_\beta$, $\langle\tilde K_x,\theta\rangle_\beta = \theta(x)$. Let $\tilde K_x = \sum_{\mu\in\mathbb{Z}}\tilde K_{x,\mu}h_\mu$, and define $K_{x,\mu} = \tilde K_{x,\mu}\cdot\frac{1+\gamma_\mu^\beta}{1+\gamma_\mu}$, $\forall x\in\mathbb{I},\mu\in\mathbb{Z}$. Then $K_x = \sum_{\mu\in\mathbb{Z}}K_{x,\mu}h_\mu$ is a

well defined element in $\Theta_{2-\beta}$ and for any $\theta \in \Theta_\beta$ with $\theta = \sum_{\mu\in\mathbb{Z}} \theta_\mu h_\mu$, by (2.3)

$$\begin{aligned}\langle K_x, \theta\rangle_1 &= \sum_{\mu\in\mathbb{Z}} \tilde{K}_{x,\mu} \cdot \frac{1+\gamma_\mu^\beta}{1+\gamma_\mu}\theta_\mu(1+\gamma_\mu) \\ &= \sum_{\mu\in\mathbb{Z}} \tilde{K}_{x,\mu}\theta_\mu(1+\gamma_\mu^\beta) =< \tilde{K}_x, \theta >_\beta = \theta(x).\end{aligned}$$

To see iii), note that for any $x \in \mathbb{I}$,

$$\begin{aligned}\|K_x\|_{2-\beta} &\le \text{const} \cdot \|\tilde{K}_x\|_\beta = \text{const} \cdot \sup_{\|\theta\|_\beta=1} |\langle \tilde{K}_x, \theta\rangle_\beta| \\ &= \text{const} \cdot \sup_{\|\theta\|_\beta=1} |\theta(x)| \le C_K,\end{aligned}$$

for some $C_K > 0$, where the last inequality follows from Sobolev's inequality that $\|\theta\|_{sup} \le \text{const} \cdot \|\theta\|_\beta$ when $\beta > 1/(2m)$ (Adams (1975)). □

Proof of Proposition 5.2. We only prove the results under Assumption A.3. For Assumption A.3′, the proof is similar.

(i) Suppose $\theta \in \Theta_b$, and define $\zeta_x := \zeta((\theta-\theta_0)(x))$ for $x \in \mathbb{I}$. By Assumption A.3, $E_X\{|\zeta_X|\} \le C_\varphi E\{|\theta(X)-\theta_0(X)|\} + E\{|\varphi(e)|\} < \infty$, which means ζ_X is absolutely integrable. By the boundedness of K_X and h_μ, and by Fubini's theorem, for any $\mu \in \mathbb{Z}$

$$\begin{aligned}V(S(\theta), h_\mu) &= E_Z\{E_X\{\zeta_X K_X(Z)\}h_\mu(Z)\} \\ &= E_X\{\zeta_X E_Z\{K_X(Z)h_\mu(Z)\}\} \\ &= E_X\{\zeta_X V(K_X, h_\mu)\}.\end{aligned}$$

Then $S(\theta) \in \Theta_{2-\beta}$ follows by Cauchy's inequality and (iii) of Proposition 5.1, i.e.,

$$\|S(\theta)\|_{2-\beta}^2 = \sum_\mu |V(S(\theta), h_\mu)|^2(1+\gamma_\mu^{2-\beta}) \le E_X\{\zeta_X^2\}E_X\{\|K_X\|_{2-\beta}^2\} < \infty.$$

(ii) For $\theta, \xi \in \Theta_b$, by mean value theorem, we have

$$\begin{aligned}&S(\theta+\xi) - S(\theta) - E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)K_X\} \\ =\ & E_X\{[\zeta'((\theta-\theta_0)(X) + t(X)\xi(X)) - \zeta'((\theta-\theta_0)(X))]\xi(X)K_X\},\end{aligned}$$

where $0 \le t(x) \le 1$ for any $x \in \mathbb{I}$. Denote $L(\xi, x) = \zeta'((\theta-\theta_0)(x) + t(x)\xi(x)) - \zeta'((\theta-\theta_0)(x))$ and $K_x = \sum_\mu K_{x,\mu}h_\mu$. Since ζ'' is upper bounded, $|L(\xi,x)| \le \text{const} \cdot |\xi(x)|$. By Cauchy's inequality,

$$\begin{aligned}&\|E_X\{L(\xi,X)\xi(X)K_X\}\|_{2-\beta}^2 \\ =\ & \sum_\mu |E_X\{L(\xi,X)\xi(X)K_{X,\mu}\}|^2(1+\gamma_\mu^{2-\beta}) \\ \le\ & E_X\{|L(\xi,X)|^2\}E_X\{|\xi(X)K_{X,\mu}|^2\}(1+\gamma_\mu^{2-\beta}) \\ =\ & E_X\{|L(\xi,X)|^2\}E_X\{|\xi(X)|^2\|K_X\|_{2-\beta}^2\} \\ \le\ & \text{const} \cdot \|\xi\|_0^4,\end{aligned}$$

where the last inequality follows from $E_X\{|L(\xi,X)|^2\} \le \text{const}\cdot\|\xi\|_0^2$, and $\sup_{x\in\mathbb{I}}\|K_x\|_{2-\beta}<\infty$. When $\|\xi\|_b\to 0$,

$$\|S(\theta+\xi)-S(\theta)-E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)K_X\}\|_{2-\beta}^2/\|\xi\|_b^2 \le \text{const}\cdot\|\xi\|_0^4/\|\xi\|_b^2\to 0,$$

which proves (5.1). By (5.1) and Cauchy's inequality, it can be shown that if $\xi\in\Theta_b$, then

$$\|DS(\theta)\xi\|_{2-\beta}^2 \le E_X\{|\zeta'((\theta-\theta_0)(X))\xi(X)|^2\}E_X\{\|K_X\|_{2-\beta}^2\}\le C\|\xi\|_b^2,$$

where C does not depend on ξ. This finishes the proof of part (ii).

(iii) Proof can be finished similarly to that in (ii). □

Proof of Proposition 5.3. We only show the lower bound. The proof for the upper bound is similar. Suppose $\xi\in\Theta_1$. For any $\mu\in\mathbb{Z}$, by Fubini's theorem,

$$\begin{aligned} & V(E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)K_X\},h_\mu)\\ =\ & E_Z\{E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)K_X(Z)\}h_\mu(Z)\}\\ =\ & E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)E_Z\{K_X(Z)h_\mu(Z)\}\}\\ =\ & E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)V(K_X,h_\mu)\}. \end{aligned}$$

Therefore,

$$\begin{aligned} & \langle E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)K_X\},\xi\rangle_1\\ =\ & \sum_\mu V(E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)K_X\},h_\mu)V(\xi,h_\mu)(1+\gamma_\mu)\\ =\ & \sum_\mu E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)V(K_X,h_\mu)\}V(\xi,h_\mu)(1+\gamma_\mu). \end{aligned} \tag{B.1}$$

On the other hand, by Cauchy's inequality and Proposition 5.1(iii), we have

$$\begin{aligned} & E_X\{|\zeta'((\theta-\theta_0)(X))\xi(X)|\sum_\mu |V(\xi,h_\mu)||V(K_X,h_\mu)|(1+\gamma_\mu)\}\\ \le\ & \text{const}\cdot E_X\{\|\xi\|_1\|K_X\|_1\}<\infty. \end{aligned}$$

Therefore, by dominated convergence theorem, the summation and expectation in (B.1) could be changed. Thus, by Proposition 5.1 (ii)

$$\begin{aligned} & \langle DS(\theta)\xi,\xi\rangle_1\\ =\ & \langle E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)K_X\},\xi\rangle_1\\ =\ & E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)\sum_\mu V(K_X,h_\mu)V(\xi,h_\mu)(1+\gamma_\mu)\}\\ =\ & E_X\{\zeta'((\theta-\theta_0)(X))\xi(X)\langle K_x,\xi\rangle_1\}\\ =\ & E\{\zeta'((\theta-\theta_0)(X))\xi(X)^2\}\\ \ge\ & \inf_{u\in I}\zeta'(u)V(\xi,\xi). \end{aligned}$$

This completes the proof of Proposition 5.3. □

Proof Proposition 5.4. For any $\mu \in \mathbb{Z}$, $DS_\lambda(\theta_0)h_\mu \in \Theta_1$. So by Proposition 2.1, $DS_\lambda(\theta_0)h_\mu = \sum_\nu \xi_\nu h_\nu$ holds in Θ_1 for some sequence ξ_ν. By Proposition 5.2 (ii), $\langle DS(\theta_0)h_\mu, h_\nu\rangle_1 = \delta_{\mu,\nu}$. By the fact that $\langle Wh_\mu, h_\nu\rangle_1 = \gamma_\mu\delta_{\mu,\nu}$, we have $\langle DS_\lambda(\theta_0)h_\mu, h_\nu\rangle_1 = \langle nDS(\theta_0)h_\mu + \lambda Wh_\mu, h_\nu\rangle_1 = (n+\lambda\gamma_\mu)\delta_{\mu,\nu}$. On the other hand, $\langle\sum_\nu \xi_\nu h_\nu, h_\nu\rangle_1 = \xi_\nu\langle h_\nu, h_\nu\rangle_1 = \xi_\nu(1+\gamma_\nu)$. Therefore $\xi_\nu = ((n+\lambda\gamma_\mu)/(1+\gamma_\nu))\delta_{\mu,\nu}$. In other words, $DS_\lambda(\theta_0)h_\mu = ((n+\lambda\gamma_\mu)/(1+\gamma_\mu))h_\mu$. So

$$DS_\lambda(\theta_0)^{-1}h_\mu = \frac{1+\gamma_\mu}{n+\lambda\gamma_\mu}h_\mu. \tag{B.2}$$

Define an operator T by $T\xi = \sum_\mu \xi_\mu \frac{1+\gamma_\mu}{n+\lambda\gamma_\mu}h_\mu$ for any $\xi = \sum_\mu \xi_\mu h_\mu$. It follows that $\xi \in \Theta_b$ implies $T\xi \in \Theta_b$ and $\|T\xi\|_b \le \text{const}\cdot\|\xi\|_b$. Therefore, $T \in \mathcal{B}(\Theta_b, \Theta_b)$ is an extension of $DS_\lambda(\theta_0)^{-1}$. To show that T is the inverse of $DS_\lambda(\theta_0)$, it is sufficient to show that for any $\xi \in \Theta_b$, $T(DS_\lambda(\theta_0)\xi) = \xi$, which follows from the linearity of T and $DS_\lambda(\theta_0)$, and the fact that $DS_\lambda(\theta_0)h_\mu = \frac{n+\lambda\gamma_\mu}{1+\gamma_\mu}h_\mu$ for any $\mu \in \mathbb{Z}$. □

Proof of Proposition 2.2. Following Taylor's expansion in Θ_b, we have for any $\phi \in \Theta_b$,

$$S_\lambda(\theta_0+\phi) - S_\lambda(\theta_0) = DS_\lambda(\theta_0)\phi + \int_0^1\int_0^1 sD^2S_\lambda(\theta_0+s's\phi)\phi\phi ds'ds. \tag{B.3}$$

Operating $DS_\lambda(\theta_0)^{-1}$ on both sides of (B.3), and exchanging with integral, we have

$$DS_\lambda(\theta_0)^{-1}(S_\lambda(\theta_0+\phi)-S_\lambda(\theta_0)) = \phi+\int_0^1\int_0^1 sDS_\lambda(\theta_0)^{-1}D^2S_\lambda(\theta_0+s's\phi)\phi\phi ds'ds. \tag{B.4}$$

Let $T_\lambda(\phi) = \phi - DS_\lambda(\theta_0)^{-1}S_\lambda(\theta_0+\phi)$ define a mapping from Θ_b to Θ_b. It is easy to see from (B.4) that

$$\begin{aligned}\|T_\lambda(\phi)\|_b &= \|\phi - DS_\lambda(\theta_0)^{-1}S_\lambda(\theta_0+\phi)\|_b \\ &\le \|\int_0^1\int_0^1 sDS_\lambda(\theta_0)^{-1}D^2S_\lambda(\theta_0+s's\phi)\phi\phi ds'ds\|_b \\ &\quad +\|DS_\lambda(\theta_0)^{-1}S_\lambda(\theta_0)\|_b\end{aligned}$$

By a direct calculation, for any $\mu \in \mathbb{Z}$

$$\begin{aligned}|\langle D^2S_\lambda(\theta_0+s's\phi)\phi\phi, h_\mu\rangle_1| &= n|\langle D^2S(\theta_\lambda+s's\phi)\phi\phi, h_\mu\rangle_1| \\ &\le n\|\zeta''\|_{sup}\cdot|E\{|\phi(X)|^2\langle K_X, h_\mu\rangle_1\}| \\ &\le n\|\zeta''\|_{sup}\cdot\|\phi\|^2_{sup}\|h_\mu\|_0 \\ &\le \text{const}\cdot n\|\phi\|_b^2. \end{aligned}\tag{B.5}$$

Both $DS(\theta_0)$ and $DS_\lambda(\theta_0)^{-1}$ are self-adjoint operators. It follows by (B.2) that for any $\mu \in \mathbb{Z}$, $DS_\lambda(\theta_0)^{-1}DS(\theta_0)h_\mu = \frac{1}{n+\lambda\gamma_\mu}h_\mu$. Then it follows by (B.5) that

for some constant C_1,

$$\begin{aligned}
& \|DS_\lambda(\theta_0)^{-1}D^2S_\lambda(\theta_0+s's\phi)\phi\phi\|_b^2 \\
= & \sum_\mu |\langle DS_\lambda(\theta_0)^{-1}D^2S_\lambda(\theta_0+s's\phi)\phi\phi, DS(\theta_0)h_\mu\rangle_1|^2(1+\gamma_\mu^b) \\
= & \sum_\mu |\langle D^2S_\lambda(\theta_0+s's\phi)\phi\phi, DS_\lambda(\theta_0)^{-1}DS(\theta_0)h_\mu\rangle_1|^2(1+\gamma_\mu^b) \\
= & \sum_\mu |\langle D^2S_\lambda(\theta_0+s's\phi)\phi\phi, h_\mu\rangle_1|^2(n+\lambda\gamma_\mu)^{-2}(1+\gamma_\mu^b) \\
\le & C_1^2n^2\|\phi\|_b^4\sum_\mu(n+\lambda\gamma_\mu)^{-2}(1+\gamma_\mu^b) \\
\approx & C_1^2\|\phi\|_b^4(\lambda/n)^{-(b+1/(2m))}, \qquad \text{(B.6)}
\end{aligned}$$

where the last step follows from Lemma 2.2 (iii) in Cox and O'Sullivan (1990).

On the other hand, let $\theta_0=\sum_\mu\theta_\mu^0h_\mu$, it can be shown that there exists some constant C_2 such that

$$\begin{aligned}
& \|DS_\lambda(\theta_0)^{-1}S_\lambda(\theta_0)\|_b^2 \\
= & \sum_\mu |V(DS_\lambda(\theta_0)^{-1}S_\lambda(\theta_0),h_\mu)|^2(1+\gamma_\mu^b) \\
= & \sum_\mu |\langle S_\lambda(\theta_0),h_\mu\rangle_1|^2(n+\lambda\gamma_\mu)^{-2}(1+\gamma_\mu^b) \\
= & \lambda^2\sum_\mu\gamma_\mu^2(n+\lambda\gamma_\mu)^{-2}(1+\gamma_\mu^b)|\theta_\mu^0|^2 \\
= & \lambda^2\sum_\mu\frac{\gamma_\mu^2(n+\lambda\gamma_\mu)^{-2}(1+\gamma_\mu^b)}{1+\gamma_\mu^d}|\theta_\mu^0|^2(1+\gamma_\mu^d) \\
\le & C_2(\lambda/n)^2\sum_\mu(\lambda/n)^{d-2-b}|\theta_\mu^0|^2(1+\gamma_\mu^d) \\
= & C_2(\lambda/n)^{d-b}\|\theta_0\|_d^2. \qquad \text{(B.7)}
\end{aligned}$$

By restricting ϕ such that $\|\phi\|_b\le C\cdot(\lambda/n)^{(d-b)/2}$ for some sufficiently large constant C, by the assumption that $\lambda/n\to0$, and by (B.5)–(B.7), we have $\|T_\lambda(\phi)\|_b\le C(\lambda/n)^{(d-b)/2}$. This implies that the operator T_λ maps $C\cdot(\lambda/n)^{(d-b)/2}\cdot B_b$ into $C\cdot(\lambda/n)^{(d-b)/2}\cdot B_b$.

Next, we show that the operator T_λ is a contraction mapping on $C\cdot(\lambda/n)^{(d-b)/2}\cdot B_b$. For any $\phi_1,\phi_2\in C\cdot(\lambda/n)^{(d-b)/2}\cdot B_b$, by Taylor's expansion

$$\begin{aligned}
& S_\lambda(\theta_0+\phi_1)-S_\lambda(\theta_0+\phi_2) \\
= & \int_0^1 DS_\lambda(\theta_0+\phi_1+s(\phi_2-\phi_1))(\phi_2-\phi_1)ds \\
= & DS_\lambda(\theta_0)(\phi_2-\phi_1) \\
& +\int_0^1\int_0^1 D^2S_\lambda(\theta_0+s'(\phi_1+s(\phi_2-\phi_1)))(\phi_2-\phi_1)(\phi_1+s(\phi_2-\phi_1))ds'ds.
\end{aligned}$$

By an argument similar to (B.6), and by $d > 2b + 1/(2m)$ and $\lambda/n = o(1)$, it can be verified that for sufficiently large n and some large constant C_3,

$$\begin{aligned} & \|T_\lambda(\phi_1) - T_\lambda(\phi_2)\|_b \qquad \text{(B.8)} \\ \leq & \int_0^1 \int_0^1 \|DS_\lambda(\theta_0)^{-1} D^2 S_\lambda(\theta_0 \\ & + s'(\phi_1 + s(\phi_2 - \phi_1)))(\phi_2 - \phi_1)(\phi_1 + s(\phi_2 - \phi_1))\|_b ds' ds \\ \leq & C_3(\lambda/n)^{-(b+1/(2m))/2}(\|\phi_1\|_b + \|\phi_2\|_b)\|\phi_1 - \phi_2\|_b \\ \leq & 2CC_3(\lambda/n)^{(d-2b-1/(2m))/2}\|\phi_1 - \phi_2\|_b \\ \leq & (1/2)\|\phi_1 - \phi_2\|_b, \qquad \text{(B.9)} \end{aligned}$$

which shows that T_λ is a contraction mapping on $C \cdot (\lambda/n)^{(d-b)/2} \cdot B_b$. Therefore, by contraction mapping theorem (Rudin (1991)), there exists a unique fixed point ϕ_λ with $\|\phi_\lambda\|_b \leq C \cdot (\lambda/n)^{(d-b)/2}$, i.e., $T_\lambda(\phi_\lambda) = \phi_\lambda$. Consequently, $\theta_\lambda^b := \theta_0 + \phi_\lambda$ is a unique root of S_λ in $C \cdot (\lambda/n)^{(d-b)/2} \cdot B_b$.

For $0 \leq b' < b \leq 1$, let $\theta_\lambda^{b'}$ satisfy $S_\lambda(\theta_\lambda^{b'}) = 0$ and $\|\theta_\lambda^{b'} - \theta_0\|_{b'} = O((\lambda/n)^{(d-b')/2})$. We will show that $\|\theta_\lambda^b - \theta_\lambda^{b'}\|_{b'} = 0$. Let $\phi_1 = \theta_\lambda^b - \theta_0$ and $\phi_2 = \theta_\lambda^{b'} - \theta_0$. Following the argument in (B.8), one can show that there is some constant C' such that

$$\begin{aligned} & \|\phi_1 - \phi_2\|_{b'} \\ = & \|T_\lambda(\phi_1) - T_\lambda(\phi_2)\|_{b'} \\ \leq & C_3(\lambda/n)^{-(b'+1/(2m))/2}(\|\phi_1\|_{b'} + \|\phi_2\|_{b'})\|\phi_1 - \phi_2\|_{b'} \\ \leq & C_3(\lambda/n)^{-(b'+1/(2m))/2}(\|\phi_1\|_b + \|\phi_2\|_{b'})\|\phi_1 - \phi_2\|_{b'} \\ \leq & C_3C'(\lambda/n)^{-(b'+1/(2m))/2+(d-b)/2}\|\phi_1 - \phi_2\|_{b'} \\ \leq & (1/2)\|\phi_1 - \phi_2\|_{b'}, \end{aligned}$$

where the last step follows from $\lambda/n \to 0$ and $d > b + b' + 1/(2m)$. Therefore, $\|\phi_1 - \phi_2\|_{b'} = 0$, which completes the proof. □

Proof of Proposition 5.5. Note that (5.6) holds for $b = 0, 1$. We use K-method of interpolation to build a relationship between these two norms which was also used by Cox (1988) to identify Θ_b and Θ_{*b}. Let the K-functional be defined as

$$K(u, \theta) = \inf_{\theta = \theta^0 + \theta^1 \in \Theta_{*0} + \Theta_{*1}} \left(\|\theta^0\|_{*0}^2 + u^2\|\theta^1\|_{*1}^2\right)^{1/2},$$

and the norm induced by K be defined to be

$$\|\theta\|_{*b,2} = \left(\int_0^\infty [u^{-b}K(u, \theta)]^2 du/u\right)^{1/2}.$$

The interpolation space is then $(\Theta_{*0}, \Theta_{*1})_b = \{\theta \in \Theta_{*0} | \ \|\theta\|_{*b,2} < \infty\}$. If $\theta = \theta^0 + \theta^1 \in \Theta_{*0} + \Theta_{*1}$, and if we denote $\theta_\mu^0 = V_*(\theta^0, h_{*\mu})$ and $\theta_\mu^1 = V_*(\theta^1, h_{*\mu})$,

then $\theta_\mu = V_*(\theta, h_\mu) = \theta_\mu^0 + \theta_\mu^1$. Therefore, for any $u \in (0, \infty)$

$$\begin{aligned} \|\theta^0\|_{*0}^2 + u^2\|\theta^1\|_{*1}^2 &= \sum_\mu \left(|\theta_\mu^0|^2 + u^2(1+\gamma_{*\mu})|\theta_\mu - \theta_\mu^0|^2\right) \\ &\geq \sum_\mu \frac{u^2(1+\gamma_{*\mu})}{1+u^2(1+\gamma_{*\mu})}\theta_\mu^2, \end{aligned}$$

where the lower bound in the above inequality is achieved by $\theta_\mu^0 = u^2(1+\gamma_{*\mu})\theta_\mu/(1+u^2(1+\gamma_{*\mu}))$ for any $\mu \in \mathbb{Z}$. It thus follows that

$$K(u,\theta)^2 = \sum_\mu \frac{u^2(1+\gamma_{*\mu})}{1+u^2(1+\gamma_{*\mu})}\theta_\mu^2.$$

Then

$$\|\theta\|_{*b,2} = \left(\int_0^\infty u^{-2b-1} \sum_\mu \frac{u^2(1+\gamma_{*\mu})}{1+u^2(1+\gamma_{*\mu})}\theta_\mu^2 du\right)^{1/2} \approx C_b\|\theta\|_{*b},$$

where $C_b = \left(\int_0^\infty u^{1-2b}/(1+u^2)du\right)^{1/2}$ only depends on b. The above arguments are also valid for the interpolation couple Θ_0 and Θ_1. Then (5.6) follows from the result that $\|\cdot\|_{*b}/\|\cdot\|_b$ is uniformly lower and upper bounded for $\theta_* \in \mathcal{K}$ when $b = 0, 1$. □

Acknowledgements

I thank Professor Chunming Zhang for motivations, and thank Professor Dingxuan Zhou for kind suggestions on the covering approaches between functional spaces. I also thank the Associate Editor and the referees for many valuable comments that helped to improve the presentation of this work.

References

Adams, R. A. (1975). *Sobolev Spaces*. Academic Press, New York-London Pure and Applied Mathematics, Vol. 65. MR0450957

Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041–1067. MR757769

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, MA. MR2239907

Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71** 1591–1608. MR2000259

Chen, X. and Pouzo, D. (2008). Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Moments. Preprint.

Chen, Z. H. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19** 1855–1868. MR1135152

Cox, D. D. (1983). Asymptotics for M-type smoothing splines. *Ann. Statist.* **11** 530–551. MR696065

Cox, D. D. (1988). Approximation of method of regularization estimators. *Ann. Statist.* **16** 694–712. MR947571

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18** 1676–1695. MR1074429

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)* **39** 1–49 (electronic). MR1864085

Freedman, D. A. (1975). On tail probabilities for martingales. *Ann. Probability* **3** 100–118. MR0380971

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *Ann. Statist.* **21** 217–234. MR1212174

Gu, C. and Ma, P. (2005). Optimal smoothing in nonparametric mixed-effect models. *Ann. Statist.* **33** 1357–1379. MR2195638

Hardy, G. H., Littlewood, J. E. and Pólya, G. (1952). *Inequalities*. Cambridge, at the Univ. Press 2d ed. MR0046395

He, X. and Shao, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630. MR1425971

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer: New York.

O'Sullivan, F. (1993). Nonparametric estimation in the Cox model. *Ann. Statist.* **21** 124–145. MR1212169

O'Sullivan, F. (1995). A study of least squares and maximum likelihood for image reconstruction in positron emission tomography. *Ann. Statist.* **23** 1267–1300. MR1353506

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.* **22** 1679–1706. MR1331198

Pollard, D. (1982). A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A* **33** 235–248. MR668445

Portnoy, S. (1997). Local asymptotics for quantile smoothing splines. *Ann. Statist.* **25** 414–434. MR1429932

Rudin, W. (1991). *Functional Analysis*, Second ed. *International Series in Pure and Applied Mathematics*. McGraw-Hill Inc., New York. MR1157815

Shen, X. (1998). On the method of penalization. *Statist. Sinica* **8** 337–357. MR1624410

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615. MR1292531

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810. MR663433

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR673642

Wahba, G. (1990). *Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. Philadelphia, PA. MR1045442

Weinberger, H. F. (1974). *Variational Methods for Eigenvalue Approximation*. Society for Industrial and Applied Mathematics, Philadelphia, Pa. MR0400004

Wong, W. H. and Severini, T. A. (1991). On maximum likelihood estimation in infinite-dimensional parameter spaces. *Ann. Statist.* **19** 603–632. MR1105838

Wu, W. B. (2005). On the Bahadur representation of sample quantiles for dependent sequences. *Ann. Statist.* **33** 1934–1963. MR2166566

Wu, W. B. (2007). M-estimation of linear models with dependent errors. *Ann. Statist.* **35** 495–521. MR2336857

Yurinskiĭ, V. V. (1976). Exponential inequalities for sums of random vectors. *J. Multivariate Anal.* **6** 473–499. MR0428401

Zhou, D.-X. (2002). The covering number in learning theory. *J. Complexity* **18** 739–767. MR1928805