Biometrika, **xx**, x, *pp*. 1–9

Printed in Great Britain

Supplementary material for "Efficient computation of smoothing splines via adaptive basis sampling"

BY PING MA

Department of Statistics, University of Georgia, 101 Cedar Street, Athens, Georgia 30602, U.S.A.

pingma@uga.edu

JIANHUA Z. HUANG AND NAN ZHANG

Department of Statistics, Texas A&M University, 155 Ireland Street, College Station, Texas 77843, U.S.A.

jianhua@stat.tamu.edu nanzhang@stat.tamu.edu

1. BAYESIAN CONFIDENCE INTERVALS

Bayesian confidence intervals (Wahba, 1983) have certain across-the-function coverage property (Nychka, 1988). We need modify Wabha's formulation slightly to take into account the fact that the basis used adaptive sampling is not a full basis.

Analogous to Wahba (1983), we decompose $\eta = \eta_0 + \eta_1$, where η_0 has a diffuse prior in the space $\mathcal{N}_J$ and η_1 has an independent Gaussian process prior with mean zero and covariance

$$\mathrm{E}\{\eta_1(x_k)\eta_1(x_l)\} = \frac{\sigma^2}{n\lambda} R_J(x_k, x^{*T}) R_{**}^{+} R_J(x^*, x_l),$$

where $x^* = (x_1^*, \ldots, x_{n^*}^*)$, and $R_J(x_k, x^{*T})$ and $R_J(x^*, x_l)$ denote respectively the row and column vectors $R_J(x_k, x^{*T}) = (R_J(x_k, x_1^*), \ldots, R_J(x_k, x_{n^*}^*))$ and $R_J(x^*, x_l) = R_J(x_l, x^{*T})^T$.

With the priors for η specified above, the posterior mean of $\eta(x)$ has the following expression,

$$\mathrm{E}\{\eta(x) \mid y\} = \xi(x)^\top \hat{d}_A + r(x)^\top \hat{c}_A,$$

where $\xi(x) = (\xi_1(x), \ldots, \xi_m(x))^\top$ is a $m \times 1$ vector, $r(x) = R_J(x^*, x)$ is a $n^* \times 1$ vector, and $\hat{d}_A$ and $\hat{c}_A$ are solutions of (13) in the main paper. The posterior variance has the following expression

$$\begin{aligned}\frac{n\lambda}{\sigma^2}\mathrm{var}\{\eta(x) \mid y\} &= r(x)^\top R_{**}^{+} r(x) + \xi(x)^\top (S^\top W_*^{-1} S)^{-1} \xi(x) \\ &\quad - 2\xi^\top (S^\top W_*^{-1} S)^{-1} S^\top W_*^{-1} R_* R_{**}^{+} r(x) \\ &\quad - r(x)^\top R_{**}^{+} R_*^\top (W_*^{-1} - W_*^{-1} S (S^\top W_*^{-1} S)^{-1} S^\top W_*^{-1}) R_* R_{**}^{+} r(x),\end{aligned}$$

where $W_* = R_* R_{**}^{+} R_*^\top + n\lambda I$. Then we construct the $100(1-\alpha)\%$ Bayesian confidence interval as $\mathrm{E}\{\eta(x) \mid y\} \pm \Phi^{-1}(1-\alpha/2)[\mathrm{var}\{\eta(x) \mid y\}]^{1/2}$, where $\Phi^{-1}(1-\alpha/2)$ is the $100(1-\alpha/2)$ percentile of the standard Gaussian distribution.

2. BASIC THEORETICAL PROPERTIES OF ADAPTIVE BASIS SAMPLING

This section presents some basic convergence properties of adaptive basis sampling to help gain some insights how it works. Since adaptive basis sampling involves the response variable,

the standard argument for the asymptotic analysis of smoothing splines does not apply. The results in this section facilitate our study of asymptotic performance of the approximated smoothing spline estimator via adaptive basis sampling.

Consider the estimation of $\mathrm{E}\{\psi(X)\}$ for a generic function $\psi \in \mathcal{L}^2(\mathcal{X})$, based on n independent and identically distributed observations $\{(x_i, y_i)\}_{i=1}^n$. The classical estimator is the sample average $\mathrm{E}_n(\psi) = n^{-1}\sum_{i=1}^n \psi(x_i)$. Suppose we use only a subsample by applying adaptive basis sampling. In the following, we shall study the asymptotic behavior of the subsample estimator. For simplicity in notation, we sometimes use either x_i or y_i to refer to (x_i, y_i) since the response and predictor variables come in pairs.

Adaptive basis sampling works as follows. First, we divide the range of $\{y_i\}_{i=1}^n$ into K slices. The number of observations in the k-th slice, $|S_k|$, is a random variable and it can be written as a sum of indicator functions, i.e. $|S_k| = \sum_{i=1}^n 1(y_i \in S_k)$. Next, n_k samples are drawn with replacement from the k-th slice. Then, we estimate $\mathrm{E}\{\psi(X)\}$ using the aggregated subsamples $\{\psi(x_i^*)\}_{i=1}^{n^*} = \bigcup_{k=1}^K \{\psi(x_j^{*(k)})\}_{j=1}^{n_k}$, where $n^* = \sum_{k=1}^K n_k$. The estimator is a weighted average and is written as

$$\mathbb{E}_n^*(\psi) = \sum_{k=1}^K \frac{|S_k|}{n}\left\{\frac{1}{n_k}\sum_{j=1}^{n_k}\psi(x_j^{*(k)})\right\}. \tag{S1}$$

Here, we use the operator notation $\mathbb{E}_n^*(\psi)$ to indicate that the sampling scheme works for a generic function ψ.

The linear operator $\mathbb{E}_n^*(\cdot)$ maps an element in $\mathcal{L}^2(\mathcal{X})$ to a random variable. Adaptive basis sampling implies that $\mathbb{E}_n^*(\psi)$ depends on the data $\{(x_i, y_i)\}_{i=1}^n$. In the following, we shall derive the conditional mean and variance of $\mathbb{E}_n^*(\psi)$ given the data and determine the magnitude of the distance of $\mathbb{E}_n^*(\psi)$ from $\mathrm{E}_n(\psi)$.

For each k, $1 \le k \le K$, $\{x_j^{*(k)}\}_{j=1}^{n_k}$ is a random draw from the k-th slice S_k. Thus, for $j = 1, \ldots, n_k$, the conditional mean of $\psi(x_j^{*(k)})$ given the data is

$$\mathrm{E}[\psi(x_j^{*(k)}) \mid \{(x_i, y_i)\}_{i=1}^n] = \frac{1}{|S_k|}\sum_{i=1}^n \psi(x_i)1(y_i \in S_k). \tag{S2}$$

It follows that the conditional mean of $\mathbb{E}_n^*(\psi)$ given the data is

$$\begin{aligned}
&\mathrm{E}[\mathbb{E}_n^*(\psi) \mid \{(x_i, y_i)\}_{i=1}^n] \\
&\quad = \mathrm{E}\left[\sum_{k=1}^K \frac{|S_k|}{n}\left\{\frac{1}{n_k}\sum_{j=1}^{n_k}\psi(x_j^{*(k)})\right\}\middle| \{(x_i, y_i)\}_{i=1}^n\right] \\
&\quad = \frac{1}{n}\sum_{k=1}^K\sum_{i=1}^n \psi(x_i)1(y_i \in S_k) = \frac{1}{n}\sum_{i=1}^n \psi(x_i) = \mathrm{E}_n(\psi).
\end{aligned} \tag{S3}$$

Hence $\mathbb{E}_n^*(\psi)$ and $\mathrm{E}_n(\psi)$ have the same mean value, $\mathrm{E}(\psi)$.

In the k-th slice, for $j = 1, \ldots, n_k$, the conditional variance of $\psi(x_j^{*(k)})$ given the data is

$$\begin{aligned}
&\mathrm{var}[\psi(x_j^{*(k)}) \mid \{(x_i, y_i)\}_{i=1}^n] \\
&\quad = \frac{1}{|S_k|}\sum_{i=1}^n \psi^2(x_i)1(y_i \in S_k) - \frac{1}{|S_k|^2}\left\{\sum_{i=1}^n \psi(x_i)1(y_i \in S_k)\right\}^2.
\end{aligned}$$

Noticing that samples from the same slice and from different slices are mutually independent, we obtain that

$$\begin{aligned}
&\mathrm{var}[\mathbb{E}_n^*(\psi) \mid \{(x_i, y_i)\}_{i=1}^n] \\
&\quad = \mathrm{var}\left[\sum_{k=1}^K \frac{|S_k|}{n}\left\{\frac{1}{n_k}\sum_{j=1}^{n_k}\psi(x_j^{*(k)})\right\}\middle|\{(x_i, y_i)\}_{i=1}^n\right] \\
&\quad = \sum_{k=1}^K \frac{|S_k|^2}{n^2}\frac{1}{n_k}\mathrm{var}[\psi(x_j^{*(k)})|\{(x_i, y_i)\}_{i=1}^n].
\end{aligned} \tag{S4}$$

LEMMA S1. *Suppose $n_k = n^*/K$, for $k = 1, \ldots, K$, then under adaptive sampling, the conditional variance of $\mathbb{E}_n^*(\psi)$ is bounded*

$$\mathrm{var}\{\mathbb{E}_n^*(\psi) \mid \{(x_i, y_i)\}_{i=1}^n\} \leqslant \frac{K}{n^*}\,\frac{1}{n}\sum_{i=1}^n \psi^2(x_i). \tag{S5}$$

Consequently,

$$\mathrm{E}\{\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)\}^2 \leqslant \frac{K}{n^*}\,\mathrm{E}(\psi^2). \tag{S6}$$

This lemma implies $\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi) \xrightarrow{p} 0$ if $n^* \to \infty$ for ψ with $E\{\psi^2(X)\} < \infty$. In other words, the subsample estimator, $\mathbb{E}_n^*(\psi)$, is a good surrogate of the usual estimator $\mathrm{E}_n(\psi)$. It is instructive to distinguish the convergence rate presented in this lemma from those presented in §4.2 of the main paper, where we show that the approximated smoothing spline estimator based on adaptive sampling can achieve the optimal convergence rate for function estimation.

*Proof of Lemma S*1. Since the variance of a random variable is bounded by its second moment, (S4) implies that

$$\mathrm{var}[\mathbb{E}_n^*(\psi) \mid \{(x_i, y_i)\}_{i=1}^n] \leqslant \sum_{k=1}^K \frac{|S_k|^2}{n^2}\,\frac{1}{n_k}\,\mathrm{E}[\psi^2(x_j^{*(k)}) \mid \{(x_i, y_i)\}_{i=1}^n].$$

Applying (S2) where ψ is replaced by ψ^2, we obtain that right-hand side of the above inequality equals

$$\sum_{k=1}^K \frac{|S_k|}{n^2}\,\frac{1}{n_k}\sum_{i=1}^n \psi^2(x_i)1(y_i \in S_k),$$

which in turn is bounded from above by

$$\sum_{k=1}^K \frac{1}{n}\frac{1}{n^*/K}\sum_{i=1}^n \psi^2(x_i)1(y_i \in S_k) = \frac{K}{n^*}\,\frac{1}{n}\sum_{i=1}^n \psi^2(x_i),$$

using $n_k = n^*/K$ and $|S_k|/n \leqslant 1$. We thus have proved (S5).

To prove (S6), applying (S3) to obtain

$$\mathrm{var}\{\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)\} = \mathrm{E}\{\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)\}^2$$

and

$$\mathrm{var}\{\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)\} = \mathrm{E}(\mathrm{var}[\mathbb{E}_n^*(\psi) \mid \{(x_i, y_i)]_{i=1}^n\}).$$

We obtain (S6) by taking expectations on both sides of (S5).

3. PROOFS OF THEORETICAL RESULTS

3·1. *Three ancillary lemmas.*

This subsection presents three lemmas that are useful in the proofs of the main results in §4 of the main paper.

LEMMA S2. *Under Condition C.2, as $\lambda \to 0$, we have*

$$\sum_{\nu} \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} = O(\lambda^{-1/r}), \tag{S7}$$

$$\sum_{\nu} \frac{1}{(1+\lambda\rho_\nu)^2} = O(\lambda^{-1/r}), \tag{S8}$$

$$\sum_{\nu} \frac{1}{1+\lambda\rho_\nu} = O(\lambda^{-1/r}). \tag{S9}$$

The proof of this lemma can be found in Section 8.2 of Gu (2013).

LEMMA S3. *Under Conditions C.1, C.2, C.3, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i)h(x_i) = V(g,h) + o_p(\{(V+\lambda J)(g)(V+\lambda J)(h)\}^{1/2})$$

for all g and h in $\mathcal{H}$.

This is Lemma 8.17 of Gu (2013).

LEMMA S4. *For any function outside the effective model space, its evaluations at selected samples $\{x_j^*\}_{j=1}^{n^*}$ are all zeros, i.e. for $h \in \mathcal{H} \ominus \mathcal{H}_E$,*

$$h(x_j^*) = 0, \qquad j = 1, \ldots, n^*.$$

Proof. According to the construction algorithm of the effective model space,

$$\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(x_j^*, \cdot),\ j = 1, \ldots, n^*\}.$$

For $h \in \mathcal{H} \ominus \mathcal{H}_E$, $h \perp g$ for $g \in \mathcal{H}_E$. Since $R_J(x_j^*, \cdot) \in \mathcal{H}_E$, we have $\langle h(\cdot), R_J(x_j^*, \cdot)\rangle_{\mathcal{H}} = 0$ for $j = 1, \ldots, n^*$. On the other hand, $\mathcal{N}_J \subseteq \mathcal{H}_E$ implies $h \in \mathcal{H} \ominus \mathcal{N}_J = \mathcal{H}_J$. It then follows from the reproducing property of $R_J(\cdot,\cdot)$ on $\mathcal{H}_J$ that $h(x_j^*) = \langle h(\cdot), R_J(x_j^*, \cdot)\rangle_{\mathcal{H}_J}$.

Noticing that the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_J}$ can be obtained by restricting $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ on $\mathcal{H}_J$, we have

$$h(x_j^*) = \langle h(\cdot), R_J(x_j^*, \cdot)\rangle_{\mathcal{H}_J} = \langle h(\cdot), R_J(x_j^*, \cdot)\rangle_{\mathcal{H}} = 0,$$

for all $j = 1, \ldots, n^*$. □

3·2. *Proof of Lemma 1*

By Lemma S4, given any $h \in \mathcal{H} \ominus \mathcal{H}_E$, $h(x_j^*) = 0$ for all $j = 1, \ldots, n^*$. Since $\{x_j^{*(k)}\}_{j=1}^{n_k}$ is a subset of $\{x_j^*\}_{j=1}^{n^*}$ which are sampled from the k-th slice, we have

$$\mathbb{E}_n^*(h^2) = \sum_{k=1}^{K} \frac{|S_k|}{n}\left\{\frac{1}{n_k}\sum_{j=1}^{n_k} h^2(x_j^{*(k)})\right\} = 0.$$

It follows that

$$V(h) = \int_{\mathcal{X}} h^2(x) f_X(x)\, dx = \mathrm{E}(h^2) - \mathbb{E}_n^*(h^2). \tag{S10}$$

By Condition C.1, there exist a collection of functions $\phi_\nu \in \mathcal{H}$ and nonnegative sequence ρ_ν such that V and J are simultaneously diagonalized, i.e., $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \, \delta_{\nu\mu}$. Using ϕ_ν's as basis functions, we expand $h(x)$ as $h(x) = \sum_\nu h_\nu \phi_\nu(x)$, where $h_\nu = V(h, \phi_\nu)$. Then we have that (S10) can be written as

$$V(h) = \mathrm{E}\left\{ \left(\sum_\nu h_\nu \phi_\nu \right)^2 \right\} - \mathbb{E}_n^* \left\{ \left(\sum_\nu h_\nu \phi_\nu \right)^2 \right\}.$$

By the fact that $\mathrm{E}(\cdot)$ and $\mathbb{E}_n^*(\cdot)$ are linear operators, simple algebra yields that

$$V(h) = \sum_\nu \sum_\mu h_\nu h_\mu \big[\mathrm{E}(\phi_\nu \phi_\mu) - \mathbb{E}_n^*(\phi_\nu \phi_\mu) \big].$$

Applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned} V(h) &\leqslant I^{1/2} \cdot \left\{ \sum_\nu \sum_\mu h_\nu^2 h_\mu^2 (1 + \lambda\rho_\nu)(1 + \lambda\rho_\mu) \right\}^{1/2} \\ &= I^{1/2} \cdot \sum_\nu h_\nu^2 (1 + \lambda\rho_\nu), \end{aligned} \tag{S11}$$

where

$$I = \sum_\nu \sum_\mu \frac{1}{1 + \lambda\rho_\nu} \frac{1}{1 + \lambda\rho_\mu} \{ \mathrm{E}(\phi_\nu \phi_\mu) - \mathbb{E}_n^*(\phi_\nu \phi_\mu) \}^2. \tag{S12}$$

Since ϕ_ν's simultaneously diagonalize V and J,

$$\sum_\nu h_\nu^2 (1 + \lambda\rho_\nu) = (V + \lambda J)(h). \tag{S13}$$

In light of (S11), to bound $V(h)$, we need to investigate the magnitude of I whose expression is given in (S12).

First, by inserting $\mathrm{E}_n(\phi_\nu \phi_\mu)$ into the squared term in (S12) and applying the inequality $(a + b)^2 \leqslant 2a^2 + 2b^2$, we obtain

$$\begin{aligned} I &\leqslant 2 \sum_\nu \sum_\mu \frac{1}{1 + \lambda\rho_\nu} \frac{1}{1 + \lambda\rho_\mu} \{ \mathrm{E}(\phi_\nu \phi_\mu) - \mathrm{E}_n(\phi_\nu \phi_\mu) \}^2 \\ &\quad + 2 \sum_\nu \sum_\mu \frac{1}{1 + \lambda\rho_\nu} \frac{1}{1 + \lambda\rho_\mu} \{ \mathrm{E}_n(\phi_\nu \phi_\mu) - \mathbb{E}_n^*(\phi_\nu \phi_\mu) \}^2 \\ &= 2I_1 + 2I_2. \end{aligned}$$

Next, we examine the magnitude of I_1 and I_2 separately.

Order of I_1. The expectation of I_1 is

$$\begin{aligned} \mathrm{E}\, I_1 &= \sum_\nu \sum_\mu \frac{1}{1 + \lambda\rho_\nu} \frac{1}{1 + \lambda\rho_\mu} \mathrm{E} \{ \mathrm{E}(\phi_\nu \phi_\mu) - \mathrm{E}_n(\phi_\nu \phi_\mu) \}^2 \\ &= \sum_\nu \sum_\mu \frac{1}{1 + \lambda\rho_\nu} \frac{1}{1 + \lambda\rho_\mu} \frac{1}{n} \operatorname{var} \{ \phi_\nu(X) \phi_\mu(X) \}. \end{aligned}$$

By Condition C.3, $\text{var}\{\phi_\nu(X)\phi_\mu(X)\} \leqslant C$ for some constant C. Therefore, by (S9),

$$\mathrm{E}\, I_1 \leqslant \frac{C}{n}\left(\sum_\nu \frac{1}{1+\lambda\rho_\nu}\right)^2 = O(n^{-1}\lambda^{-2/r}). \tag{S14}$$

Order of I_2. The expectation of I_2 is

$$\mathrm{E}\, I_2 = \sum_\nu \sum_\mu \frac{1}{1+\lambda\rho_\nu}\frac{1}{1+\lambda\rho_\mu}\mathrm{E}\{\mathrm{E}_n(\phi_\nu\phi_\mu) - \mathbb{E}_n^*(\phi_\nu\phi_\mu)\}^2.$$

Assuming $n_k = n^*/K$ for all k, applying Lemma S1, and substituting ψ by the product $\phi_\nu\phi_\mu$ in (S6), we obtain

$$\mathrm{E}\{\mathrm{E}_n(\phi_\nu\phi_\mu) - \mathbb{E}_n^*(\phi_\nu\phi_\mu)\}^2 \leqslant \frac{K}{n^*}\mathrm{E}(\phi_\nu^2\phi_\mu^2) \leqslant \frac{K}{n^*}\, C, \tag{S15}$$

where the constant C is the bound for $\mathrm{E}(\phi_\nu^2\phi_\mu^2)$ in Condition C.3. Therefore, by (S9),

$$\mathrm{E}\, I_2 \leqslant \frac{K\cdot C}{n^*}\left(\sum_\nu \frac{1}{1+\lambda\rho_\nu}\right)^2 = O(n^{*-1}\lambda^{-2/r}). \tag{S16}$$

Putting (S14) and (S16) together and noticing $n^* \ll n$, we obtain

$$\mathrm{E}\, I \leqslant 2\,\mathrm{E}\, I_1 + 2\,\mathrm{E}\, I_2 = O(n^{*-1}\lambda^{-2/r}) + O(n^{-1}\lambda^{-2/r}) = O(n^{*-1}\lambda^{-2/r}).$$

Therefore, $I = O_p(n^{*-1}\lambda^{-2/r})$ and

$$V(h) \leqslant (V+\lambda J)(h)\cdot O_p(n^{*-1/2}\lambda^{-1/r}).$$

The assumption $n^*\lambda^{2/r} \to \infty$ implies that $n^{*-1/2}\lambda^{-1/r} \to 0$. The desired result thus follows from the above inequality.

3·3. *Proof of Theorem 3*

The smoothing spline estimator $\hat\eta$ is the minimizer of

$$\mathrm{PLS}(\eta) = \frac{1}{n}\sum_{i=1}^n \{y_i - \eta(x_i)\}^2 + \lambda\, J(\eta), \tag{S17}$$

over $\mathcal{H}$. By the representer theorem, it can be written as $\hat\eta(x) = \sum_{i=1}^m \hat d_\nu\phi_\nu(x) + \sum_{i=1}^n \hat c_i R_J(x_i, x)$. Let $\hat\eta_P(x)$ be the projection of $\hat\eta(x)$ to $\mathcal{H}_E$ relative to the reproducing kernel Hilbert space inner product.

According to Theorem 2 in the main paper, $\hat\eta$ converges to the true function η_0 with certain rate. Notice that $\hat\eta_A - \eta_0 = (\hat\eta_A - \hat\eta_P) + (\hat\eta_P - \hat\eta) + (\hat\eta - \eta_0)$. We shall show that both $\hat\eta_P - \hat\eta$ and $\hat\eta_A - \hat\eta_P$ converge to zero at the same or a faster rate. We achieve this in two steps.

Step 1. We show that $\hat\eta_P$ converges to η_0 in the same rate as $\hat\eta$. To this end, note that $\hat\eta - \hat\eta_P \perp \mathcal{H}_E$, and thus $J(\hat\eta_P, \hat\eta - \hat\eta_P) = 0$. For any functions $g, h \in \mathcal{H}$, we define

$$A_{g,h}(\alpha) = \frac{1}{n}\sum_{i=1}^n \{y_i - (g+\alpha h)(x_i)\}^2 + \lambda J(g+\alpha h). \tag{S18}$$

It can be easily shown that

$$\left.\frac{dA_{g,h}(\alpha)}{d\alpha}\right|_{\alpha=0} = -\frac{2}{n}\sum_{i=1}^n (y_i - g(x_i))h(x_i) + 2\lambda J(g,h). \tag{S19}$$

Since $\hat{\eta}$ is the minimizer of (S17) over $\mathcal{H}$, $A_{g,h}(\alpha)$ reaches its minimum at $\alpha = 0$ when $g = \hat{\eta}$ and $h = \hat{\eta} - \hat{\eta}_P$. Thus, for this choice of g and h, the derivative in (S19) is zero. It follows that

$$\lambda J(\hat{\eta}, \hat{\eta} - \hat{\eta}_P) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{\eta}(x_i)\}\{\hat{\eta}(x_i) - \hat{\eta}_P(x_i)\}. \tag{S20}$$

The fact that $J(\hat{\eta}_P, \hat{\eta} - \hat{\eta}_P) = 0$ implies that $J(\hat{\eta} - \hat{\eta}_P)$ is equal to $J(\hat{\eta}, \hat{\eta} - \hat{\eta}_P)$. Thus

$$\lambda J(\hat{\eta} - \hat{\eta}_P) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{\eta}(x_i)\}\{\hat{\eta}(x_i) - \hat{\eta}_P(x_i)\} \triangleq S_1 + S_2, \tag{S21}$$

where

$$S_1 = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \eta_0(x_i)\}\{\hat{\eta}(x_i) - \hat{\eta}_P(x_i)\},$$
$$S_2 = \frac{1}{n}\sum_{i=1}^{n}\{\eta_0(x_i) - \hat{\eta}(x_i)\}\{\hat{\eta}(x_i) - \hat{\eta}_P(x_i)\}.$$

We now study the orders of the two terms S_1 and S_2 under Conditions C.1, C.2, C.3, and $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$.

Recall $\phi_\nu \in \mathcal{H}$ are eigenfunctions which simultaneously diagonalize V and J such that $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu\mu}$. Write $\hat{\eta} - \hat{\eta}_P = \sum_\nu (\hat{\eta} - \hat{\eta}_P)_\nu \phi_\nu$, where $(\hat{\eta} - \hat{\eta}_P)_\nu = V(\hat{\eta} - \hat{\eta}_P, \phi_\nu)$. It follows that

$$S_1 = \sum_\nu (\hat{\eta} - \hat{\eta}_P)_\nu \left[\frac{1}{n}\sum_{i=1}^{n}\{y_i - \eta_0(x_i)\}\phi_\nu(x_i)\right].$$

Applying the Cauchy-Schwarz inequality, we have

$$S_1^2 \leqslant \left\{\sum_\nu (\hat{\eta} - \hat{\eta}_P)_\nu^2 (1 + \lambda\rho_\nu)\right\} \sum_\nu \frac{1}{1 + \lambda\rho_\nu}\left[\frac{1}{n}\sum_{i=1}^{n}\{y_i - \eta_0(x_i)\}\phi_\nu(x_i)\right]^2.$$

Similar to (S13), the first term on the above right hand side equals

$$\sum_\nu (\hat{\eta} - \hat{\eta}_P)_\nu^2 (1 + \lambda\rho_\nu) = (V + \lambda J)(\hat{\eta} - \hat{\eta}_P). \tag{S22}$$

Let

$$Z_\nu = \frac{1}{n}\sum_{i=1}^{n}(y_i - \eta_0(x_i))\phi_\nu(x_i) = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\phi_\nu(x_i),$$

then the second term in the upper bound of S_1^2 can be written as $\sum_\nu \frac{1}{1+\lambda\rho_\nu} Z_\nu^2$. Noting that

$$\mathrm{E}(Z_\nu^2) = \mathrm{E}\left\{\frac{1}{n^2}\sum_{i=1}^{n}\epsilon_i^2\phi_\nu(x_i)^2\right\} = \frac{1}{n}\mathrm{E}(\epsilon_i^2)\mathrm{E}\{\phi_\nu(x_i)^2\} = \frac{\sigma^2}{n},$$

and by Lemma S2, we have

$$\mathrm{E}\left(\sum_\nu \frac{1}{1 + \lambda\rho_\nu} Z_\nu^2\right) = \sum_\nu \frac{1}{1 + \lambda\rho_\nu}\mathrm{E}(Z_\nu^2) = O(n^{-1}\lambda^{-1/r}). \tag{S23}$$

Combining (S22) and (S23), we obtain

$$S_1 \leqslant \{(V+\lambda J)(\hat{\eta}-\hat{\eta}_P)\}^{1/2} O_p(n^{-1/2}\lambda^{-1/(2r)}). \tag{S24}$$

We next determine the order of S_2. Applying Lemma S3 with $g=\eta_0-\hat{\eta}$ and $h=\hat{\eta}-\hat{\eta}_P$, we obtain that

$$S_2 = V(\eta_0-\hat{\eta},\hat{\eta}-\hat{\eta}_P) + \{(V+\lambda J)(\eta_0-\hat{\eta})(V+\lambda J)(\hat{\eta}-\hat{\eta}_P)\}^{1/2}\, o_p(1)$$

The Cauchy-Schwarz inequality implies that

$$\begin{aligned} V(\eta_0-\hat{\eta},\hat{\eta}-\hat{\eta}_P) &\leqslant \{V(\eta_0-\hat{\eta})V(\hat{\eta}-\hat{\eta}_P)\}^{1/2} \\ &\leqslant \{(V+\lambda J)(\eta_0-\hat{\eta})(V+\lambda J)(\hat{\eta}-\hat{\eta}_P)\}^{1/2}. \end{aligned}$$

Thus

$$S_2 \leqslant \{(V+\lambda J)(\eta_0-\hat{\eta})(V+\lambda J)(\hat{\eta}-\hat{\eta}_P)\}^{1/2}\,(1+o_p(1)).$$

Now we are ready to determine the order of $(V+\lambda J)(\hat{\eta}-\hat{\eta}_P)$. Since $\hat{\eta}-\hat{\eta}_P \in \mathcal{H}\ominus\mathcal{H}_E$, $V(\hat{\eta}-\hat{\eta}_P)$ is dominated by $\lambda J(\hat{\eta}-\hat{\eta}_P)$, by Lemma 1. Thus, $(V+\lambda J)(\hat{\eta}-\hat{\eta}_P)$ converges to zero at the same order as $\lambda J(\hat{\eta}-\hat{\eta}_P)$. Therefore

$$\begin{aligned} &(V+\lambda J)(\hat{\eta}-\hat{\eta}_P) \asymp \lambda J(\hat{\eta}-\hat{\eta}_P) = S_1+S_2 \\ &\leqslant \{(V+\lambda J)(\hat{\eta}-\hat{\eta}_P)\}^{1/2} O_p(n^{-1/2}\lambda^{-1/2r}) \\ &\quad + \{(V+\lambda J)(\eta_0-\hat{\eta})(V+\lambda J)(\hat{\eta}-\hat{\eta}_P)\}^{1/2}\,(1+o_p(1)). \end{aligned}$$

After canceling out $\{(V+\lambda J)(\hat{\eta}-\hat{\eta}_P)\}^{1/2}$ and taking squares on both sides, we obtain

$$\begin{aligned} (V+\lambda J)(\hat{\eta}-\hat{\eta}_P) &\asymp O_p(n^{-1}\lambda^{-1/r}) + (V+\lambda J)(\hat{\eta}-\eta_0) \\ &\asymp (V+\lambda J)(\hat{\eta}-\eta_0) \\ &= O_p(n^{-1}\lambda^{-1/r}+\lambda^p). \end{aligned} \tag{S25}$$

Step 2. We show that $\hat{\eta}_A$, the smoothing spline estimator via adaptive sampling, converges to η_0 with the same convergence rate as $\hat{\eta}_P$.

Since $\hat{\eta}$ is the minimizer of (S17) over $\mathcal{H}$, $A_{g,h}(\alpha)$ reaches its minimum at $\alpha=0$ when $g=\hat{\eta}$ and $h=\hat{\eta}_A-\hat{\eta}_P$. Arguing as in the proof of (S20), we have

$$\lambda J(\hat{\eta},\hat{\eta}_A-\hat{\eta}_P) = \frac{1}{n}\sum_{i=1}^{n}\{y_i-\hat{\eta}(x_i)\}\{\hat{\eta}_A(x_i)-\hat{\eta}_P(x_i)\}. \tag{S26}$$

Since $\hat{\eta}_A$ is the minimizer of (S17) over $\mathcal{H}_E$, $A_{g,h}(\alpha)$ reaches its minimum at $\alpha=0$ when $g=\hat{\eta}_A$ and $h=\hat{\eta}_A-\hat{\eta}_P$. Thus, similar to the previous result, we have

$$\lambda J(\hat{\eta}_A,\hat{\eta}_A-\hat{\eta}_P) = \frac{1}{n}\sum_{i=1}^{n}\{y_i-\hat{\eta}_A(x_i)\}\{\hat{\eta}_A(x_i)-\hat{\eta}_P(x_i)\}. \tag{S27}$$

We subtract (S26) from (S27) and obtain

$$\lambda J(\hat{\eta}_A-\hat{\eta},\hat{\eta}_A-\hat{\eta}_P) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\eta}(x_i)-\hat{\eta}_A(x_i)\}\{\hat{\eta}_A(x_i)-\hat{\eta}_P(x_i)\}.$$

Because $\hat{\eta}_P$ is the projection of $\hat{\eta}$ onto $\mathcal{H}_E$ and $\hat{\eta}_A - \hat{\eta}_P \in \mathcal{H}_E$, the orthogonality implies that $J(\hat{\eta}_P - \hat{\eta}, \hat{\eta}_A - \hat{\eta}_P) = 0$. Thus

$$J(\hat{\eta}_A - \hat{\eta}_P) = J(\hat{\eta}_A - \hat{\eta}_P, \hat{\eta}_A - \hat{\eta}_P) = J(\hat{\eta}_A - \hat{\eta}, \hat{\eta}_A - \hat{\eta}_P).$$

Combining the above two displays, we obtain

$$\lambda J(\hat{\eta}_A - \hat{\eta}_P) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\eta}(x_i) - \hat{\eta}_A(x_i)\}\{\hat{\eta}_A(x_i) - \hat{\eta}_P(x_i)\}.$$

With this result, some algebra yields

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\{\hat{\eta}_A(x_i) - \hat{\eta}_P(x_i)\}^2 + \lambda J(\hat{\eta}_A - \hat{\eta}_P) \\ = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\eta}(x_i) - \hat{\eta}_P(x_i)\}\{\hat{\eta}_A(x_i) - \hat{\eta}_P(x_i)\}.\end{aligned} \tag{S28}$$

Applying Lemma S3, we see that the left hand side of (S28) equals

$$V(\hat{\eta}_A - \hat{\eta}_P) + o_p\{(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_P)\} + \lambda J(\hat{\eta}_A - \hat{\eta}_P) = (V + \lambda J)(\hat{\eta}_A - \hat{\eta}_P)\{1 + o_p(1)\}.$$

Another application of Lemma S3 yields that the right hand side of (S28) equals

$$\begin{aligned}&V(\hat{\eta}_A - \hat{\eta}_P, \hat{\eta} - \hat{\eta}_P) + o_p\{(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_P)(V + \lambda J)(\hat{\eta} - \hat{\eta}_P)\}^{1/2} \\ &\quad\leqslant \{V(\hat{\eta}_A - \hat{\eta}_P)V(\hat{\eta} - \hat{\eta}_P)\}^{1/2} + o_p\{(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_P)(V + \lambda J)(\hat{\eta} - \hat{\eta}_P)\}^{1/2} \\ &\quad\leqslant \{(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_P)(V + \lambda J)(\hat{\eta} - \hat{\eta}_P)\}^{1/2}\{1 + o_p(1)\}.\end{aligned}$$

Combining the above two results, we obtain that

$$\begin{aligned}&(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_P)\{1 + o_p(1)\} \\ &\qquad= \{(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_P)(V + \lambda J)(\hat{\eta} - \hat{\eta}_P)\}^{1/2}\{1 + o_p(1)\}.\end{aligned}$$

Canceling out a term from both sides, we obtain

$$(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_P)\{1 + o_p(1)\} = (V + \lambda J)(\hat{\eta} - \hat{\eta}_P)\{1 + o_p(1)\},$$

combining which with the result from Step 1 yields

$$(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_P) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p). \tag{S29}$$

Putting (S25) and (S29) together, we conclude the proof with the convergence rate

$$(V + \lambda J)(\hat{\eta}_A - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p).$$

REFERENCES

GU, C. (2013). *Smoothing Spline ANOVA Models*. New York: Springer, 2nd ed.

NYCHKA, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.* **83**, 1134–1143.

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45**, 133–150.