Adaptive Basis Selection for Exponential Family Smoothing Splines with Application in Joint Modeling of Multiple Sequencing Samples

Ping Ma, Nan Zhang, Jianhua Z. Huang and Wenxuan Zhong

University of Georgia, Fudan University and Texas A&M University

Supplementary Material

In this supplementary material, we further present some properties of our proposed estimator, and provide the proofs for related lemmas and theorems. Lastly, we show the derivation of generalized approximate cross-validation for choosing the tuning parameter.

S1 Properties of ABS estimator

Since our basis selection algorithm involves the response variable, the standard argument for the asymptotic analysis of smoothing splines does not apply. We first present some theoretical properties which shed light on how the adaptive basis sampling algorithm works and facilitate our asymptotic analysis.

Consider the estimation of $\mathrm{E}\{\psi(X, Y)\}$ based on n i.i.d. observations

$\{(x_i, y_i)\}_{i=1}^n$, where $\psi(x, y) \in \mathcal{L}^2(\mathcal{X}, \mathcal{Y})$ is a generic multivariate function. Two notations are introduced as the standard sample mean estimator and a mean estimator based on a subset of samples which is adaptively selected by our proposed method,

$$\mathrm{E}_n(\psi) = \frac{1}{n}\sum_{i=1}^{n}\psi(x_i, y_i), \tag{S1.1}$$

$$\mathbb{E}_n^*(\psi) = \sum_{k=1}^{K}\frac{|S_k|}{n}\left\{\frac{1}{n_k}\sum_{j=1}^{n_k}\psi(x_j^{*(k)}, y_j^{*(k)})\right\}. \tag{S1.2}$$

The following lemma shows the new estimator based on a subsample provides a good approximation to that based on all observations.

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

$$\mathrm{var}\{\mathbb{E}_n^*(\psi)|\{(x_i, y_i)\}_{i=1}^n\} \leqslant \frac{K}{n^*}\,\frac{1}{n}\sum_{i=1}^{n}\psi^2(x_i, y_i) \tag{S1.3}$$

and

$$\mathrm{E}\{\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)\}^2 \leqslant \frac{K}{n^*}\,\mathrm{E}(\psi^2). \tag{S1.4}$$

This lemma implies $\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)$ converges to zero in probability if $n^* \to \infty$ for ψ with $E\{\psi^2(X, Y)\} < \infty$. In other words, the subsample estimator, $\mathbb{E}_n^*(\psi)$, is a good surrogate of the usual estimator $\mathrm{E}_n(\psi)$.

To understand the behavior of $\hat{\eta}_A$, the smoothing spline estimator computed using the adaptive basis selection algorithm, we refer to two important properties of the effective model space $\mathcal{H}_E$.

Lemma 2. *For any function outside the effective model space, its evaluations at selected samples* $\{x_j^*\}_{j=1}^{n^*}$ *are all zeros, i.e. for* $h \in \mathcal{H} \ominus \mathcal{H}_E$,

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

Lemma 3. *Under Condition 1, 2, and 4, as* $\lambda \to 0$ *and* $n^*\lambda^{2/r} \to \infty$, *if unction* h *is not in the effective model space, i.e.,* $h \in \mathcal{H} \ominus \mathcal{H}_E$, *we have*

$$V(h) = o_p\{\lambda J(h)\}.$$

S2 Proofs

Proof of Lemma 1 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 $\psi(x_j^{*(k)}, y_j^{*(k)})$ given the data is

$$\mathrm{E}\{\psi(x_j^{*(k)}, y_j^{*(k)})|\{(x_i, y_i)\}_{i=1}^n\} = \frac{1}{|S_k|}\sum_{i=1}^n \psi(x_i, y_i)\mathbf{1}(y_i \in S_k). \qquad \text{(S2.1)}$$

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

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

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 $\psi(x_j^{*(k)}, y_j^{*(k)})$ given the data is bounded by its second order conditional moment whose explicit form can be obtained by replacing ψ by ψ^2 in (S2.1), i.e.

$$\mathrm{var}\{\psi(x_j^{*(k)}, y_j^{*(k)})|\{(x_i, y_i)\}_{i=1}^n\} \leqslant \mathrm{E}\{\psi^2(x_j^{*(k)}, y_j^{*(k)})|\{(x_i, y_i)\}_{i=1}^n\} = \frac{1}{|S_k|}\sum_{i=1}^n \psi^2(x_i, y_i)\mathbf{1}(y_i \in S_k). \tag{S2.2}$$

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

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

The right hand side of the above has an upper bound due to (S2.2)

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

which in turn is upper bounded by

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

with the fact that $n_k = n^*/K$ and $|S_k|/n \leqslant 1$. We thus have proved (S1.3).

The condition mean of $\mathbb{E}_n^*(\psi)$ given the data has been proved to be $\mathrm{E}_n(\psi)$. Recall the definition of conditional variance, we have

$$\mathrm{var}\{\mathbb{E}_n^*(\psi)|\{(x_i, y_i)\}_{i=1}^n\} = \mathrm{E}[\{\mathbb{E}_n^*(\psi) - \mathrm{E}_n(\psi)\}^2|\{(x_i, y_i)\}_{i=1}^n].$$

We obtain (S1.4) immediately by taking expectation on both sides of the above, i.e.

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

Before proving the main result, we first present two useful lemmas in Gu (2013).

Lemma 4. *Under Condition 2, as* $\lambda \to 0$*, one has*

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

This is part of Lemma 9.1 in Gu (2013).

Lemma 5. *Under Condition 1, 2 and 4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$*,*

$$\frac{1}{n}\sum_{i=1}^n g(x_i)h(x_i)w(\eta_0(x_i); y_i) = V(g,h) + o_p(\{(V+\lambda J)(g)(V+\lambda J)(h)\}^{1/2})$$

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

This is Lemma 9.16 in Gu (2013).

Proof of Lemma 2 See the supplementary material of Ma et al. (2015).

Proof of Lemma 3 By Lemma 2, given the selected samples $\{x_j^*\}_{j=1}^{n^*}$, for any $h \in \mathcal{H} \ominus \mathcal{H}_E$, we have

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

Note that $\{x_j^*\}_{j=1}^{n^*}$ is the collection of $\{x_j^{*(k)}\}_{j=1}^{n_k}$ from $k = 1, \ldots, K$ slices, hence

$$\mathbb{E}_n^*\{h^2(X)w(\eta_0(X); Y)\} = \sum_{k=1}^{K} \frac{|S_k|}{n} \left\{ \frac{1}{n_k} \sum_{j=1}^{n_k} h^2(x_j^{*(k)}) w(\eta_0(x_j^{*(k)}); y_j^{*(k)}) \right\} = 0.$$

It follows that

$$V(h) = \int_{\mathcal{X}} h^2(x) v_{\eta_0}(x) f_X(x)\, dx = \mathrm{E}\{h(X)^2 v_{\eta_0}(X)\} - \mathbb{E}_n^*\{h^2(X)w(\eta_0(X); Y))\}. \tag{S2.3}$$

By Condition 1, there exist a collection of functions $\phi_\nu \in \mathcal{H}$ and a sequence of nonnegative ρ_ν such that V and J are simultaneously diagonalized, i.e., $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu\, \delta_{\nu\mu}$. Use ϕ_ν's as basis functions and expand h as $h = \sum_\nu h_\nu \phi_\nu$, where $h_\nu = V(h, \phi_\nu)$. Then, (S2.3) can be written as

$$V(h) = \mathrm{E}\left\{ \left(\sum_\nu h_\nu \phi_\nu(X) \right)^2 v_{\eta_0}(X) \right\} - \mathbb{E}_n^* \left\{ \left(\sum_\nu h_\nu \phi_\nu(X) \right)^2 w(\eta_0(X); Y) \right\}.$$

Due to the fact that $\mathrm{E}(\cdot)$ and $\mathbb{E}_n^*(\cdot)$ are both linear operators, we have

$$V(h) = \sum_\nu \sum_\mu h_\nu h_\mu \big[\mathrm{E}\{\phi_\nu(X)\phi_\mu(X) v_{\eta_0}(X)\} - \mathbb{E}_n^*\{\phi_\nu(X)\phi_\mu(X) w(\eta_0(X); Y)\} \big].$$

Applying the Cauchy-Schwarz inequality to obtain

$$V(h) \leqslant I^{1/2} \cdot \left\{ \sum_{\nu} \sum_{\mu} h_{\nu}^2 h_{\mu}^2 (1+\lambda\rho_{\nu})(1+\lambda\rho_{\mu}) \right\}^{1/2} \tag{S2.4}$$

$$= I^{1/2} \cdot \sum_{\nu} h_{\nu}^2 (1+\lambda\rho_{\nu}) \tag{S2.5}$$

where

$$I = \sum_{\nu} \sum_{\mu} \frac{1}{1+\lambda\rho_{\nu}} \frac{1}{1+\lambda\rho_{\mu}} \big[\mathrm{E}\{\phi_{\nu}(X)\phi_{\mu}(X) v_{\eta_0}(X)\} - \mathbb{E}_n^*\{\phi_{\nu}(X)\phi_{\mu}(X) w(\eta_0(X); Y)\} \big]^2. \tag{S2.6}$$

Since ϕ_{ν}'s simultaneously diagonalize V and J,

$$\sum_{\nu} h_{\nu}^2 (1+\lambda\rho_{\nu}) = (V + \lambda J)(h). \tag{S2.7}$$

In light of (S2.4), to bound $V(h)$, we need to investigate the magnitude of I whose expression is given in (S2.6).

First, by inserting

$$\mathrm{E}_n\{\phi_{\nu}(X)\phi_{\mu}(X) w(\eta_0(X); Y)\} = \frac{1}{n} \sum_{i=1}^{n} \phi_{\nu}(x_i)\phi_{\mu}(x_i) w(\eta_0(x_i); y_i)$$

into the squared term in (S2.6) and applying the inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$, we obtain

$$\begin{aligned} I \leqslant\ & 2 \sum_{\nu} \sum_{\mu} \frac{1}{1+\lambda\rho_{\nu}} \frac{1}{1+\lambda\rho_{\mu}} \big[\mathrm{E}\{\phi_{\nu}(X)\phi_{\mu}(X) v_{\eta_0}(X)\} - \mathrm{E}_n\{\phi_{\nu}(X)\phi_{\mu}(X) w(\eta_0(X); Y)\} \big]^2 \\ & + 2 \sum_{\nu} \sum_{\mu} \frac{1}{1+\lambda\rho_{\nu}} \frac{1}{1+\lambda\rho_{\mu}} \big[\mathrm{E}_n\{\phi_{\nu}(X)\phi_{\mu}(X) w(\eta_0(X); Y)\} - \mathbb{E}_n^*\{\phi_{\nu}(X)\phi_{\mu}(X) w(\eta_0(X); Y)\} \big]^2 \\ \triangleq\ & 2I_1 + 2I_2. \end{aligned}$$

Next, we examine the magnitudes of I_1 and I_2 one by one.

Order of I_1. Recall that $\mathrm{E}\{w(\eta_0(x);y)\} = v_{\eta_0}(x)$, then

$$\mathrm{E}\big[\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\}\big] = \mathrm{E}\{\phi_\nu(X)\phi_\mu(X)v_{\eta_0}(X)\}$$

and

$$\mathrm{var}\big[\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\}\big] = \frac{1}{n}\,\mathrm{var}\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\}.$$

Therefore, the expectation of I_1 is

$$\begin{aligned}\mathrm{E}\, I_1 &= \sum_\nu \sum_\mu \frac{1}{1+\lambda\rho_\nu}\frac{1}{1+\lambda\rho_\mu}\mathrm{E}\big[\mathrm{E}\{\phi_\nu(X)\phi_\mu(X)v_{\eta_0}(X)\} - \mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\}\big]^2 \\ &= \sum_\nu \sum_\mu \frac{1}{1+\lambda\rho_\nu}\frac{1}{1+\lambda\rho_\mu}\,\frac{1}{n}\,\mathrm{var}\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\}.\end{aligned}$$

By Condition 4, $\mathrm{var}\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\} \leqslant c_3$ for some constant $c_3 < \infty$. Hence, by Lemma 4,

$$\mathrm{E}\, I_1 \leqslant \frac{c_3}{n}\left(\sum_\nu \frac{1}{1+\lambda\rho_\nu}\right)^2 = O(n^{-1}\lambda^{-2/r}). \tag{S2.8}$$

Order of I_2. The expectation of I_2 is

$$\mathrm{E}\, I_2 = \sum_\nu \sum_\mu \frac{1}{1+\lambda\rho_\nu}\frac{1}{1+\lambda\rho_\mu}\,\mathrm{E}\big[\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\} - \mathbb{E}_n^*\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\big]^2.$$

As in Lemma 1, we assume $n_k = n^*/K$ for all k and substitute $\psi(x,y)$ by

$\phi_\nu(x)\phi_\mu(x)w(\eta_0(x);y)$ in (S1.4) to obtain

$$\begin{aligned}
&\mathrm{E}\big[\mathrm{E}_n\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\} - \mathbb{E}_n^*\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\big]^2 \\
&\quad \leqslant \frac{K}{n^*}\mathrm{E}\{\phi_\nu^2(X)\phi_\mu^2(X)w^2(\eta_0(X);Y)\} \\
&\quad \leqslant \frac{K}{n^*}\,(c_3+1),
\end{aligned}$$

where the constant c_3 is the bound of $\mathrm{var}\{\phi_\nu(X)\phi_\mu(X)w(\eta_0(X);Y)\}$ in Condition 4. Again, by Lemma 4,

$$\mathrm{E}\, I_2 \leqslant \frac{K(c_3+1)}{n^*}\left(\sum_\nu \frac{1}{1+\lambda\rho_\nu}\right)^2 = O(n^{*-1}\lambda^{-2/r}). \tag{S2.9}$$

Putting (S2.8) and (S2.9) 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 desired result follows from the fact $n^{*-1/2}\lambda^{-1/r} \to 0$.

Proof of Theorem 1 in the main paper By the representer theorem, $\hat\eta$, the minimizer of (2.1) in the main paper, has an explicit form as in (2.2) of the main paper. Given the effective model space $\mathcal{H}_E$, let $\hat\eta_E$ be the projection of $\hat\eta$ to $\mathcal{H}_E$ relative to the reproducing kernel Hilbert space inner product. The proposed estimator $\hat\eta_A$ uses basis functions from $\mathcal{H}_E$ while $\hat\eta$ uses the full basis from $\mathcal{H}$.

According to Theorem 9.17 in Gu (2013), $\hat{\eta}$ converges to the true function η_0 with certain rate. Notice that

$$\hat{\eta}_A - \eta_0 = (\hat{\eta}_A - \hat{\eta}_E) + (\hat{\eta}_E - \hat{\eta}) + (\hat{\eta} - \eta_0).$$

It suffices to show that both $\hat{\eta}_E - \hat{\eta}$ and $\hat{\eta}_A - \hat{\eta}_E$ converge to zero at the same or a faster rate. We achieve this in two steps.

Step 1. We show that $\hat{\eta}_E$ converges to η_0 with the same rate as $\hat{\eta}$. To this end, note that $\hat{\eta} - \hat{\eta}_E \in \mathcal{H} \ominus \mathcal{H}_E \subseteq \mathcal{H}_J$ and $\hat{\eta} \in \mathcal{H}_E$, therefore $J(\hat{\eta} - \hat{\eta}_E, \hat{\eta}_E) = 0$.

For any functions $g, h \in \mathcal{H}$, define

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

It can be easily shown that

$$\left.\frac{dA_{g,h}(\alpha)}{d\alpha}\right|_{\alpha=0} = \frac{1}{n}\sum_{i=1}^{n} u(g(x_i); y_i)h(x_i) + \lambda J(g, h). \tag{S2.10}$$

Since $\hat{\eta}$ is the minimizer of (2) in the main paper over $\mathcal{H}$, $A_{g,h}(\alpha)$ reaches its minimum at $\alpha = 0$ when $g = \hat{\eta}$ and $h = \hat{\eta} - \hat{\eta}_E$. Thus, for this choice of g and h, the derivative in (S2.10) is zero. It follows that

$$\lambda J(\hat{\eta}, \hat{\eta} - \hat{\eta}_E) = -\frac{1}{n}\sum_{i=1}^{n} u(\hat{\eta}(x_i); y_i)\{\hat{\eta}(x_i) - \hat{\eta}_E(x_i)\}. \tag{S2.11}$$

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

Thus

$$\lambda J(\hat{\eta} - \hat{\eta}_E) = -\frac{1}{n}\sum_{i=1}^{n} u(\hat{\eta}(x_i); y_i)\{\hat{\eta}(x_i) - \hat{\eta}_E(x_i)\} \triangleq S_1 + S_2, \qquad \text{(S2.12)}$$

where

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

We next study the orders of the two terms S_1 and S_2 under Conditions 1, 2 and 4, and $\lambda \to 0$, $n\lambda^{2/r} \to \infty$.

For S_1, since $u(\eta(x), y)$ is differentiable with respect to $\eta(x)$, it follows by the mean value theorem and Condition 3 that there exists a constant $\gamma \in [c_1, c_2]$ such that

$$S_1 = -\frac{\gamma}{n}\sum_{i=1}^{n} w(\eta_0(x_i); y_i)\{\hat{\eta}(x_i) - \eta_0(x_i)\}\{\hat{\eta}(x_i) - \hat{\eta}_E(x_i)\}.$$

Applying Lemma 5 to the right hand side of the above, we have

$$|S_1| = \gamma\, V(\hat{\eta} - \eta_0, \hat{\eta} - \hat{\eta}_E) + \{(V + \lambda J)(\hat{\eta} - \eta_0)(V + \lambda J)(\hat{\eta} - \hat{\eta}_E)\}^{1/2}\, o_p(1)$$
$$= \{(V + \lambda J)(\hat{\eta} - \eta_0)(V + \lambda J)(\hat{\eta} - \hat{\eta}_E)\}^{1/2}\, O_p(1)$$

For S_2, 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 $\hat{\eta} - \hat{\eta}_E = \sum_\nu (\hat{\eta} - \hat{\eta}_E)_\nu \phi_\nu$, where $(\hat{\eta} - \hat{\eta}_E)_\nu = V(\hat{\eta} - \hat{\eta}_E, \phi_\nu)$. Plugging it in

S_2 and applying Cauchy-Schwarz inequality, we have

$$\begin{aligned}|S_2| &= \left|\sum_\nu (\hat{\eta}-\hat{\eta}_E)_\nu \left\{\frac{1}{n}\sum_{i=1}^n u(\eta_0(x_i);y_i)\phi_\nu(x_i)\right\}\right| \\ &\leqslant \left\{\sum_\nu \frac{\beta_\nu^2}{1+\lambda\rho_\nu}\right\}^{1/2}\left\{\sum_\nu (\hat{\eta}-\hat{\eta}_E)_\nu^2(1+\lambda\rho_\nu)\right\}^{1/2}\end{aligned}$$

where $\beta_\nu = \frac{1}{n}\sum_{i=1}^n u(\eta_0(x_i);y_i)\phi_\nu(x_i)$ possesses properties $\mathrm{E}(\beta_\nu)=0$ and $\mathrm{var}(\beta_\nu)=\sigma^2/n$. In fact

$$\mathrm{E}(\beta_\nu) = \mathrm{E}\{u(\eta_0(X);Y)\phi_\nu(X)\} = \mathrm{E}_X\big[\mathrm{E}\{u(\eta_0(X);Y)|X\}\phi_\nu(X)\big] = 0$$

and

$$\begin{aligned}\mathrm{E}(\beta_\nu^2) &= \frac{1}{n}\mathrm{E}\{u^2(\eta_0(X);Y)\phi_\nu^2(X)\} = \frac{1}{n}\mathrm{E}_X\big[\mathrm{E}\{u^2(\eta_0(X);Y)|X\}\phi_\nu^2(X)\}\big] \\ &= \frac{\sigma^2}{n}\mathrm{E}_X\{v_{\eta_0}(X)\phi_\nu^2(X)\} = \frac{\sigma^2}{n}V(\phi_\nu) = \frac{\sigma^2}{n}.\end{aligned}$$

Furthermore, by Lemma 4,

$$\mathrm{E}\left\{\sum_\nu \frac{\beta_\nu^2}{1+\lambda\rho_\nu}\right\} = \frac{\sigma^2}{n}\sum_\nu \frac{1}{1+\lambda\rho_\nu} = O(n^{-1}\lambda^{-1/r}). \tag{S2.13}$$

and it can be shown by a similar argument as in (S2.7) that

$$\sum_\nu (\hat{\eta}-\hat{\eta}_E)_\nu^2(1+\lambda\rho_\nu) = (V+\lambda J)(\hat{\eta}-\hat{\eta}_E). \tag{S2.14}$$

Combining (S2.13) and (S2.14), we obtain

$$S_2 \leqslant \{(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)\}^{1/2} O_p(n^{-1/2}\lambda^{-1/(2r)}).$$

Now we are ready to determine the order of $(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)$. By Lemma 3, $V(\hat{\eta}-\hat{\eta}_E)$ is dominated by $\lambda J(\hat{\eta}-\hat{\eta}_E)$ since $\hat{\eta}-\hat{\eta}_E \in \mathcal{H}\ominus\mathcal{H}_E$.

Thus, $(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)$ converges to zero at the same order as $\lambda J(\hat{\eta}-\hat{\eta}_E)$. Therefore, it follows (S2.12) that

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

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

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

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

Since $\hat{\eta}$ is the minimizer of (2.2) in the main paper over $\mathcal{H}$, $A_{g,h}(\alpha)$ reaches its minimum at $\alpha = 0$ when $g = \hat{\eta}$ and $h = \hat{\eta}_A - \hat{\eta}_E$. Arguing as in the proof of (S2.11), we have

$$\lambda J(\hat{\eta}, \hat{\eta}_A - \hat{\eta}_E) = -\frac{1}{n}\sum_{i=1}^{n} u(\hat{\eta}(x_i); y_i)\{\hat{\eta}_A(x_i) - \hat{\eta}_E(x_i)\}. \tag{S2.15}$$

Since $\hat{\eta}_A$ is also the minimizer of (2.2) in the main paper over $\mathcal{H}_E$, $A_{g,h}(\alpha)$ reaches its minimum at $\alpha = 0$ when $g = \hat{\eta}_A$ and $h = \hat{\eta}_A - \hat{\eta}_E$. Thus, similar

to the previous result, we have

$$\lambda J(\hat{\eta}_A, \hat{\eta}_A - \hat{\eta}_E) = -\frac{1}{n}\sum_{i=1}^{n} u(\hat{\eta}_A(x_i); y_i)\{\hat{\eta}_A(x_i) - \hat{\eta}_E(x_i)\}. \tag{S2.16}$$

We subtract (S2.15) from (S2.16) to obtain

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

Recall that $\hat{\eta}_E$ is the projection of $\hat{\eta}$ onto $\mathcal{H}_E$ and $\hat{\eta}_A - \hat{\eta}_E \in \mathcal{H}_E$, then $(\hat{\eta} - \hat{\eta}_E) \perp (\hat{\eta}_A - \hat{\eta}_E)$. Such orthogonality implies that $J(\hat{\eta} - \hat{\eta}_E, \hat{\eta}_A - \hat{\eta}_E) = 0$ and further

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

With this result, some algebra yields

$$\frac{1}{n}\sum_{i=1}^{n}\{u(\hat{\eta}_A(x_i); y_i) - u(\hat{\eta}_E(x_i); y_i)\}\{\hat{\eta}_A(x_i) - \hat{\eta}_E(x_i)\} + \lambda J(\hat{\eta}_A - \hat{\eta}_E) \tag{S2.17}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{u(\hat{\eta}(x_i); y_i) - u(\hat{\eta}_E(x_i); y_i)\}\{\hat{\eta}_A(x_i) - \hat{\eta}_E(x_i)\} \tag{S2.18}$$

By the mean value theorem, Condition 3 and Lemma 5, there exists a constant $\zeta \in [c_1, c_2]$ such that the left hand side of (S2.17) equals

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

Similarly the right hand side of (S2.17) is bounded by

$$\{(V + \lambda J)(\hat{\eta} - \hat{\eta}_E)(V + \lambda J)(\hat{\eta}_A - \hat{\eta}_E)\}^{1/2} O_p(1).$$

Combining the above two results, we obtain that

$$(V+\lambda J)(\hat{\eta}_A-\hat{\eta}_E)\{1+o_p(1)\} = \{(V+\lambda J)(\hat{\eta}-\hat{\eta}_E)(V+\lambda J)(\hat{\eta}_A-\hat{\eta}_E)\}^{1/2}O_p(1).$$

Canceling out a term from both sides to obtain

$$(V+\lambda J)(\hat{\eta}_A-\hat{\eta}_E) \asymp (V+\lambda J)(\hat{\eta}-\hat{\eta}_E) = O_p(n^{-1}\lambda^{-1/r}+\lambda^p). \quad \text{(S2.19)}$$

Putting results from Step 1 and 2 together, we conclude the proof with the convergence rate

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

S3 Derivation of generalized approximate cross-validation

The minimizer of (2.6) in the main paper satisfies the normal equation

$$\begin{pmatrix} S_w^T S_w & S_w^T R_w \\ R_w^T R_w & R_w^T R_w + (n\lambda)Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S_w^T \tilde{\mathbf{Y}}_w \\ R_w^T \tilde{\mathbf{Y}}_w \end{pmatrix}, \quad \text{(S3.1)}$$

where $S_w = \tilde{W}^{1/2}S$, $R_w = \tilde{W}^{1/2}R$, and $\tilde{\mathbf{Y}}_w = \tilde{W}^{1/2}\tilde{\mathbf{Y}}$. The normal equation of (S3.1) can be solved by the pivoted Cholesky decomposition followed by backward and forward substitutions (Kim and Gu (2004)). On the convergence of Newton iteration, the "fitted values" of $\hat{\mathbf{Y}}_w = S_w\mathbf{d} + R_w\mathbf{c}$ by minimizing (2.4) in the main paper can be written as $\hat{\mathbf{Y}}_w = A_w(\lambda)\tilde{\mathbf{Y}}_w$,

where the smoothing matrix

$$A_w(\lambda) = (S_w, R_w) \begin{pmatrix} S_w^T S_w & S_w^T R_w \\ R_w^T R_w & R_w^T R_w + (n\lambda)Q \end{pmatrix}^+ \begin{pmatrix} S_w^T \\ R_w^T \end{pmatrix},$$

where $S_w = \tilde{W}^{1/2} S$, $R_w = \tilde{W}^{1/2} R$, $\tilde{\mathbf{Y}}_w = \tilde{W}^{1/2} \tilde{\mathbf{Y}}$, and $\mathbf{C}^+$ denotes the Moore-Penrose inverse of $\mathbf{C}$ satisfying $\mathbf{C}\mathbf{C}^+\mathbf{C} = \mathbf{C}$, $\mathbf{C}^+\mathbf{C}\mathbf{C}^+ = \mathbf{C}^+$, $(\mathbf{C}\mathbf{C}^+)^T = \mathbf{C}\mathbf{C}^+$ and $(\mathbf{C}^+\mathbf{C})^T = \mathbf{C}^+\mathbf{C}$.

A data-driven approach for the selection of the tuning parameter λ (including θ) is to choose λ which minimizes the generalized approximate cross-validation score (Gu and Xiang (2001)), one version of which was derived by Gu and Xiang (2001) and is of the following form,

$$GACV(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\{Y_i\hat{\eta}_A(x_i) - b(\hat{\eta}_A(x_i))\} + \frac{\mathrm{tr}(A_w\tilde{W}^{-1})}{n - \mathrm{tr}A_w}\frac{1}{n}\sum_{i=1}^{n} Y_i(Y_i - \hat{\mu}(x_i)). \tag{S3.2}$$

One may employ standard nonlinear optimization algorithms to minimize the generalized approximate cross-validation score. In particular, we use the modified Newton algorithm developed by Dennis and Schnabel (1996) to find the minimizer. $\hat{\eta}_A$ and $\hat{\mu}$ are evaluated at the minimizer of (2.1) in the main paper with fixed tuning parameters, and A_w and $\tilde{W}$ are evaluated at the values given at the convergence of the Newton iterations.

References

Dennis, J. E. and R. B. Schnabel (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia: SIAM. Corrected reprint of the 1983 original.

Gu, C. (2013). *Smoothing Spline ANOVA Models* (2nd ed.), Volume 297. New York: Springer.

Gu, C. and D. Xiang (2001). Cross-validating non-Gaussian data: Generalized approximate cross-validation revisited. *Journal of Computational and Graphical Statistics 10*, 581–591.

Kim, Y.-J. and C. Gu (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B 66*, 337–356.

Ma, P., J. Z. Huang, and N. Zhang (2015). Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika 102*(3), 631–645.