Online Smooth Backfitting for Generalized Additive Models

S.1 Further Discussions and Extensions

S.1.1 Online inference

We discuss online generalized likelihood ratio (GLR) test for AM and GAM in this subsection. First, we briefly review the classical GLR test for AM proposed in Fan and Jiang (2005) and Chatla (2022). Consider the hypothesis testing problem

$$H_0 : \beta_d(x_d) = 0 \ \text{ v.s. } \ H_1 : \beta_d(x_d) \neq 0. \tag{1}$$

Let $RSS_1 = \sum_{i=1}^{N}\{Y_i - \widehat{\beta}_0 - \sum_{j=1}^{d}\widehat{\beta}_j(X_{ij})\}^2$ and $RSS_0 = \sum_{i=1}^{N}\{Y_i - \check{\beta}_0 - \sum_{j=1}^{d-1}\check{\beta}_j(X_{ij})\}^2$, where $\widehat{\beta}_j$ and $\check{\beta}_j$ $(j = 0, 1, \ldots, d)$ are estimates under the alternative and null hypotheses, respectively. The GLR statistic is defined as follows,

$$\lambda_n(H_0) = \frac{N}{2}\log\frac{RSS_0}{RSS_1}.$$

Asymptotic null distribution and power of GLR statistic based on the classical backfitting and smooth backfitting are studied in Fan and Jiang (2005) and Chatla (2022), respectively. In both methods, critical values are obtained by bootstrap.

Next, we discuss the online extension of GLR test for AM. Let $\widetilde{\beta}_{K,j}$ and $\check{\beta}_{K,j}$ $(j = 0, 1, \ldots, d)$ be the online estimates under the alternative and null hypotheses, respectively. Define the online GLR statistic by

$$\lambda_K(H_0) = \frac{n_K}{2}\log\frac{RSS_{K0}}{RSS_{K1}},$$

where $RSS_{K1} = \sum_{i=1}^{n_K}\{Y_{Ki} - \widetilde{\beta}_{K0} - \sum_{j=1}^{d}\widetilde{\beta}_{Kj}(X_{Kij})\}^2$ and $RSS_0 = \sum_{i=1}^{n_K}\{Y_{Ki} - \check{\beta}_{K0} - \sum_{j=1}^{d-1}\check{\beta}_{Kj}(X_{Kij})\}^2$. Note that our proposed method guarantees that the pseudo-bandwidths are of the same order as the bandwidth selected by the batch method. Let $\sigma_{N_K}, \sigma, \mu_{N_K}, d_{1N_K}$ be defined the same as σ_n, μ_n, d_n in Chatla (2022). The following results hold following Theorem 1 and 3 of Chatla (2022).

Proposition 1. *Suppose that Assumptions (A1)–(A4) hold. Then, under H_0 for the testing problem* (1)*,*

$$P\left\{\sigma_{N_K}^{-1}\left(\lambda_K(H_0) - \mu_{N_K} - (2\sigma^2)^{-1}d_{1N_K}\right) < t|Data\right\} \xrightarrow{d} N(0,1).$$

Proposition 2. *Suppose that Assumptions (A1)–(A4) hold and $\widetilde{h}_{Kj} = c_j N_K^{-2/9}$ for $j = 1, \ldots, d$ and some constants c_j. Then for the testing problem* (1)*, the GLR test can detect alternatives with rate $N_K^{-4/9}$.*

Finally, as the GLR test for GAM remains an open question, we discuss the main challenge of developing the online GLR test for GAM. Note that theoretical guarantees of the classical GLR test in Fan and Jiang (2005) and Chatla (2022) are based on the normal equations. Specifically, $\lambda_n(H_0)$ can be approximated by

$$\frac{n}{2}\frac{RSS_0 - RSS_1}{RSS_1}$$

and the normal equations facilitate the expression $RSS_0 - RSS_1 = \epsilon^\top(A_1 - A_2)\epsilon + (2\sigma^2)^{-1}d_{1n}$ for some matrices A_1, A_2. Based on this, the asymptotic normality can be established. For GAM, the estimation equations are nonlinear and the asymptotic null distribution would be difficult to derive, which is worth further investigation and we leave it for future work.

S.1.2 Extension to semiparametric models

We discuss the extension of the proposed method to semiparametric models in this subsection. Denote Θ as a finite dimensional parameter set and $\mathcal{F}$ as a function space. Let $z_i = (x_i, y_i)$ for $1 \leq i \leq N$ be the observations, and $\theta^* \in \Theta$ and $f^* \in \mathcal{F}$ be the true unknown finite and infinite dimensional parameters. Chen et al. (2003) proposed to estimate θ^* and f^* by maximizing the likelihood. Specifically, let $q(z, \theta^*, f^*)$ be the density function of z_i and

$$m(z, \theta, f) = \frac{\partial \log(q(z, \theta, f))}{\partial \theta}$$

be the score function, and the sample score is given by $S_N(\theta) = \sum_{i=1}^{N} m(z_i, \theta, f)$. Chen et al. (2003) proposed to estimate $f(\cdot|\theta)$ nonparametrically and then estimate $S_N(\theta)$ by $\widehat{S}_N(\theta)$. They proved the estimate $\widehat{\theta}$ which solves $\widehat{S}_N(\theta) = 0$ has the same asymptotic distribution as the infeasible estimate obtained with known f^*. Note that $\widehat{S}_N(\theta)$ is a smooth function with respect to θ and depends on the bandwidth of $\widehat{f}$. In the online context, one can expand $\widehat{S}_N(\theta)$ to an appropriate order and treat the coefficients as statistics which can be updated using the DCB method. Though the implementation of extension is straightforward, the analysis of asymptotic efficiency of the online estimate of θ would pose technical difficulty, which we leave for future work.

S.1.3 Results on regrets

In this subsection, we discuss the performance of the proposed method in terms of regrets which is also a popular measurement for online algorithms defined as below (Blum and Mansour, 2007; Warmuth and Kuzmin, 2008; Li et al., 2019):

$$\text{regret}_K = \sum_{k=1}^{K} \sum_{j=1}^{d} \left\{ IMSE(\widetilde{\beta}_{kj}) - IMSE(\widehat{\beta}_{kj}) \right\}.$$

Let $\ell(L) = (1 + c_1 L^{-1} + c_2 L^{-2})^{-1}$ be the lower bound of the relative efficiency derived in Theorem 3. Note that the regets are measured cumulatively over blocks and can be calculated based on the relative efficiency we have already defined, Given $IMSE(\widetilde{\beta}_{kj}) \leq \ell(L)^{-1} IMSE(\widehat{\beta}_{kj})$, it is straightforward to obtain

$$\begin{aligned}\text{regret}_K &\leq \sum_{k=1}^{K}\sum_{j=1}^{d}\left\{\left(\ell(L)^{-1}-1\right) IMSE(\widehat{\beta}_{kj})\right\} \\ &= \left(\ell(L)^{-1}-1\right)\left\{\frac{1}{4}\mu_2^2(\mathcal{K})\theta_j \sum_{k=1}^{K}\sum_{j=1}^{d}\widehat{h}_{kj}^4 + R(\mathcal{K})\sigma_j^2 \sum_{k=1}^{K}\sum_{j=1}^{d}\frac{1}{N_K \widehat{h}_{kj}}\right\} + o_p\left(\widehat{h}_{Kj}^4 + N_K^{-1}\widehat{h}_{Kj}^{-1}\right),\end{aligned}$$

where $R(\mathcal{K}) = \int \mathcal{K}(x)^2 dx$, $\mu_2(\mathcal{K}) = \int x^2 \mathcal{K}(x) dx$ and σ_j^2, θ_j are defined as in (15). Note that $\widehat{h}_{Kj} \to h_{Kj}^* = [R(\mathcal{K})\sigma_j^2/\{\mu_2(\mathcal{K})^2\theta_j N_K\}]^{\frac{1}{5}}$. Then we obtain the regret bound of the proposed method

$$\text{regret}_K \lesssim \frac{25}{4}(c_1 L^{-1} + c_2 L^{-2})\left[\sum_{j=1}^{d}\{\mu_2^2(\mathcal{K})\theta_j\}^{1/5}\{R(\mathcal{K})\sigma_j^2\}^{4/5}\right] N_K^{1/5}.$$

S.2 Numerical comparison in the simplified case

In this section, we compare the performance of the proposed method with Kong and Xia (2019), Quan and Lin (2022) and Xue and Yao (2022) when the link function is identical and there is only one component function. We abbreviate these methods as Kong, Quan and Xue, respectively. The data are generated by

$$y = \sum_{k=1}^{4} k^{-1.5}\phi_k(x) + e$$

with $\phi_1(x) = 1, \phi_{2k}(x) = \cos(2k\pi x), \phi_{2k+1}(x) = \sin(2k\pi x)$, $x \sim U(0,1)$ and $e \sim N(0,1)$. To mimic this practice, we set the block size $n_k = 100$ for $k = 1, \ldots, K_{max}$. For the proposed method, the parameters are set as follows: $G = R = 0.3$, the length of candidate sequence for the main regression $L = 20$ and for the polit estimates $L' = 10$. For the method of Kong, we set the bandwidth $\check{h}_k = (3/10)^{1/5}\widetilde{h}_k$ where $\widetilde{h}_k$ is the online bandwidth of our

proposed method, as they did not provide the empirical bandwidth selection approach. For the method of Quan, the tuning parameters are determined by the semi-data-driven strategy described in Section 2.4 in Quan and Lin (2022). For the method of Xue, we adopt cubic splines with parameter λ tuned by the generalized cross-validation method and the dynamic knots updated according to the implementation in Section 2.2 of Xue and Yao (2022) with the parameters $\alpha = 1$ and $\nu = 1/3$.

The results of these methods are presented in Figure 1. In terms of relative efficiency, our method is stably efficient and higher than the theoretical lower bound during the whole procedure; the efficiency of Kong increases at start and tends to stable at 0.84 as data accumulate; the spline based methods of Quan and Xue produce similar performances. In conclusion, all estimators yield comparable performance in the long run in this simplified case.

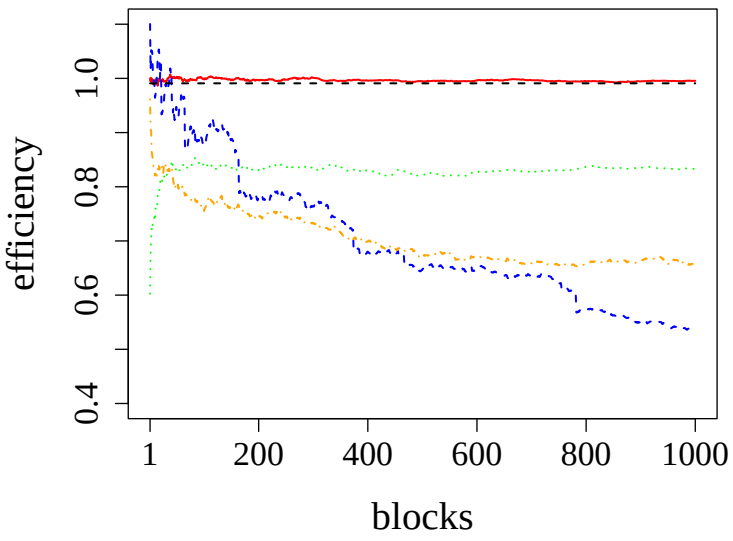

Figure 1: The relative efficiencies of the proposed method (red solid) and its theoretical lower bound (black dashed), and the methods of Kong (green dotted), Quan (blue dashed) and Xue (yellow dashed).

S.3 Proofs of the Main Results

In this section, we first derive the algorithmic convergence and asymptotic properties of the proposed method, i.e., Theorem 1 and Lemma 2–3, in Section S.3.1 and Section S.3.2. The convergence of bandwidth selection and lower bound of relative efficiencies are shown in Section S.3.3 and S.3.4.

S.3.1 Proof of Theorem 1

The proof is established by verifying the conditions of Lemma 1.

Proof. of Theorem 1. Recalling the space $\mathcal{H}(\boldsymbol{Q})$ defined in (5), we further define

$$\mathcal{H}^0(\boldsymbol{Q}) = \left\{ \boldsymbol{f} \in \mathcal{H}(\boldsymbol{Q}) : \int \{Q_{00}(\boldsymbol{x}) f_j(x_j) + Q_{0j}(\boldsymbol{x}) f_{1j}(x_j)\} d\boldsymbol{x} = 0 \text{ for } j = 1, \ldots, d \right\},$$

$$\mathcal{H}_j(\boldsymbol{Q}) = \left\{ \boldsymbol{f} = (f, f_{1j})^\top : \|\boldsymbol{f}\|_{\boldsymbol{Q},j} < \infty,\ f(\boldsymbol{x}) = f_0 + f_j(x_j), f_{1j}(\boldsymbol{x}) = g_j(x_j) \right\},$$

$$\mathcal{H}_j^0(\boldsymbol{Q}) = \left\{ \boldsymbol{f} \in \mathcal{H}_j(\boldsymbol{Q}) : \int \{Q_{00}(\boldsymbol{x}) f_j(x_j) + Q_{0j}(\boldsymbol{x}) f_{1j}(x_j)\} d\boldsymbol{x} = 0 \right\},$$

where $\|\cdot\|_{\boldsymbol{Q},j}$ is given by

$$\|\boldsymbol{f}\|_{\boldsymbol{Q},j} = \int \left\{ f_0 Q_{00} + \begin{bmatrix} f_j(x_j) & f_{1j}(x_j) \end{bmatrix} \begin{bmatrix} Q_{00} & Q_{0j} \\ Q_{j0} & Q_{jj} \end{bmatrix} \begin{bmatrix} f_j(x_j) \\ f_{1j}(x_j) \end{bmatrix} \right\} d\boldsymbol{x}.$$

Denote the projection operator from $\mathcal{H}(\boldsymbol{M}^{\lfloor 0 \rfloor})$ onto $\mathcal{H}_j^0(\boldsymbol{M}^{\lfloor 0 \rfloor})$ by $\boldsymbol{\Pi}_j$. In the following proof, we omit the superscript "$\lfloor 0 \rfloor$" in $\boldsymbol{M}$ for conciseness. Then we have that for any $\boldsymbol{\xi} \in \mathcal{H}(\boldsymbol{M})$,

$$\boldsymbol{\Pi}_j \boldsymbol{\xi} = \left(\int \begin{bmatrix} M_{00} & M_{0j} \\ M_{j0} & M_{jj} \end{bmatrix} d\boldsymbol{x}_{-j} \right)^{-1} \begin{bmatrix} \int \boldsymbol{M}_{0:} \boldsymbol{\xi} d\boldsymbol{x}_{-j} \\ \int \boldsymbol{M}_{j:} \boldsymbol{\xi} d\boldsymbol{x}_{-j} \end{bmatrix}.$$

Let $\boldsymbol{\Pi} = (\boldsymbol{\Pi}_1^\top, \ldots, \boldsymbol{\Pi}_d^\top)^\top$. The Fréchet derivative of $\widetilde{\boldsymbol{F}}_K$ at $\widetilde{\boldsymbol{\beta}}_K^{\lfloor 0 \rfloor}$ can be written as

$$\widetilde{\boldsymbol{F}}_K^{(1)}(\widetilde{\boldsymbol{\beta}}_K^{\lfloor 0 \rfloor}, \boldsymbol{\xi}) = \boldsymbol{D}\boldsymbol{A}\boldsymbol{\xi} \tag{2}$$

where

$$\boldsymbol{D} = \begin{bmatrix} \int M_{00}(\boldsymbol{x})d\boldsymbol{x} & \mathbf{0}^\top \\ \mathbf{0} & \operatorname{diag}\left(\left\{ \int \begin{bmatrix} M_{00} & M_{0j} \\ M_{j0} & M_{jj} \end{bmatrix} d\boldsymbol{x}_{-j} \right\}_{j=1,\ldots,d} \right) \end{bmatrix}, \ \boldsymbol{A} = \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \boldsymbol{\Pi} \end{bmatrix}.$$

Since $g'(\widetilde{m}_{K-1}(\cdot))^2 V(\widetilde{m}_{K-1}(\cdot))$ is bounded and $\widetilde{p}_{K-1}(\cdot)$ is bounded away from 0, one can obtain that $\boldsymbol{D}^{-1}$ is bounded. Using the same technique, we can prove that $\boldsymbol{\Pi}$ is one-to-one and onto, and $\|\boldsymbol{\Pi}\boldsymbol{\xi}\| \leq d\|\boldsymbol{\xi}\|$. Hence $\boldsymbol{\Pi}$ has a bounded inverse, which implies that $\boldsymbol{A}$ also has, i.e. there exists a constant c_1 such that $\|\widetilde{\boldsymbol{F}}_K^{(1)}(\widetilde{\boldsymbol{\beta}}_K^{\lfloor 0 \rfloor})\| < c_1$.

The other conditions are verified as follows. To prove that $\widetilde{\boldsymbol{F}}_K^{(1)}$ satisfies the Lipschitz condition $\|\widetilde{\boldsymbol{F}}_K^{(1)}\boldsymbol{\beta} - \widetilde{\boldsymbol{F}}_K^{(1)}\boldsymbol{\beta}'\| \leq c_2\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|$, by the decomposition (2), it is sufficient to prove the boundedness of $\boldsymbol{D}$ and $\boldsymbol{A}$, which is guaranteed by the smoothness and boundedness of $g'(\widetilde{m}_{K-1}(\cdot))^2 V(\widetilde{m}_{K-1}(\cdot))$ and $\widetilde{p}_{K-1}(\cdot)$. Finally, by the uniform continuity of $\widetilde{\boldsymbol{F}}_K$ and the fact $\widetilde{\boldsymbol{F}}_K\widetilde{\boldsymbol{\beta}}_K = 0$, there exists a positive constant r such that $\sup_{\boldsymbol{\beta} \in B_r(\widetilde{\boldsymbol{\beta}}_K)} \|\widetilde{\boldsymbol{F}}_K\boldsymbol{\beta}\| < (2c_1^2c_2)^{-1}$. This proves that if $\widetilde{\boldsymbol{\beta}}_K^{\lfloor 0 \rfloor} \in B_r(\widetilde{\boldsymbol{\beta}}_K)$, then $\|\{\widetilde{\boldsymbol{F}}_K^{(1)}(\widetilde{\boldsymbol{\beta}}_K^{\lfloor 0 \rfloor})\}^{-1}\widetilde{\boldsymbol{F}}_K(\widetilde{\boldsymbol{\beta}}_K^{\lfloor 0 \rfloor})\| < (2c_1c_2)^{-1}$. Note that $\boldsymbol{M}$ is close to $\overline{\boldsymbol{Q}}_K$, the two norms $\|\cdot\|_{\boldsymbol{M}}$ and $\|\cdot\|_{\overline{\boldsymbol{Q}}_K}$ are equivalent. Conclusion (1) now follows from Lemma 1. Theorem 1 (2) follows from the property of Hilbert-Schmidt operators $\boldsymbol{\Pi}_j$; see Theorem 4.B in Appendix 4 of Bickel et al. (1993) and Theorem 4 of Yu et al. (2008) for details. □

S.3.2 Proof of Lemma 2-3

We first prove Lemma 2. Recall that $\check{\boldsymbol{\xi}}^*$ is defined in (23). The proof is based on the decomposition of $\check{\boldsymbol{\xi}}^*$.

Proof. We further decompose $\check{\xi}_j^*(x_j) = \check{\xi}_j^{A*}(x_j) + \check{\xi}_j^{B*}(x_j), \check{\xi}_{1j}^*(x_j) = \check{\xi}_{1j}^{A*}(x_j) + \check{\xi}_{1j}^{B*}(x_j)$ that

$$\begin{bmatrix} \check{\xi}_j^A(x_j) \\ \check{\xi}_{1j}^A(x_j) \end{bmatrix} = \begin{bmatrix} \check{\xi}_j^*(x_j) \\ \check{\xi}_{1j}^*(x_j) \end{bmatrix} - \begin{bmatrix} E(\check{\xi}_j^*(x_j) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) \\ E(\check{\xi}_{1j}^*(x_j) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) \end{bmatrix},$$

$$\begin{bmatrix} \check{\xi}_j^B(x_j) \\ \check{\xi}_{1j}^B(x_j) \end{bmatrix} = \begin{bmatrix} E(\check{\xi}_j^*(x_j) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) \\ E(\check{\xi}_{1j}^*(x_j) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) \end{bmatrix}.$$

We omit the constants in the functions and define the spaces

$$\overline{\mathcal{H}}^0(\boldsymbol{Q}) = \Big\{ \boldsymbol{f} = (f_1(x_1), \ldots, f_d(x_d), f_{11}(x_1), f_{1d}(x_d))^\top : \int (Q_{00} f_j + Q_{0j} f_{1j}) d\boldsymbol{x} = 0 \text{ for } j = 1, \ldots, d \Big\},$$

$$\overline{\mathcal{H}}_j^0(\boldsymbol{Q}) = \{ \boldsymbol{f} \in \overline{\mathcal{H}}^0(\boldsymbol{Q}) : \boldsymbol{f} \text{ depends only on } x_j \}.$$

In this proof, we consider $2d$-dimensional functions. Let Ψ_j be the projection operator from $\overline{\mathcal{H}}^0(\boldsymbol{Q}_K^*)$ onto $\overline{\mathcal{H}}_j^0(\boldsymbol{Q}_K^*)$. Then for any $\boldsymbol{\xi} \in \overline{\mathcal{H}}^0(\boldsymbol{Q}_K^*)$,

$$\Psi_j \xi_\ell(x_\ell) = f_\ell(x_\ell), \quad \Psi_{1j} \xi_{1\ell}(x_\ell) = f_{1\ell}(x_\ell),$$

where for $\ell \neq j$, $f_\ell(x_\ell) = \xi_\ell(x_\ell)$ and $f_{1\ell}(x_\ell) = \xi_{1\ell}(x_\ell)$, and

$$f_j(x_j) = g_j(x_j) - \int g_j(x_j) \overline{Q}_{K,00}^*(\boldsymbol{x}) d\boldsymbol{x},$$

$$\begin{bmatrix} g_j(x_j) \\ f_{1j}(x_j) \end{bmatrix} = -\sum_{\ell \neq j} \left(\int \begin{bmatrix} \overline{Q}_{K,00}^* & \overline{Q}_{K,0j}^* \\ \overline{Q}_{K,j0}^* & \overline{Q}_{K,jj}^* \end{bmatrix} d\boldsymbol{x}_{-j} \right)^{-1} \int \begin{bmatrix} \overline{Q}_{K,00}^* & \overline{Q}_{K,0\ell}^* \\ \overline{Q}_{K,j0}^* & \overline{Q}_{K,j\ell}^* \end{bmatrix} \begin{bmatrix} \xi_\ell(x_\ell) \\ \xi_{1\ell}(x_\ell) \end{bmatrix} d\boldsymbol{x}_{-j}.$$

Then define $T = \Psi_1 \cdots \Psi_d$. Recall that $\omega_{ki}(\boldsymbol{x}, \boldsymbol{\beta})$ plays the same role of $K_h(\boldsymbol{X}_{ki} - \boldsymbol{x})$ in Mammen (1999). With assumption (A2) on the smoothness of link function g, conclusions of AM can be carried over to GAM and we can obtain that T is smaller than $\gamma < 1$ w.p.1. Let $\zeta_j(x_j) = -\int \sum_{k=1}^K P_{k,0}(\boldsymbol{\beta}^*; \widetilde{\boldsymbol{\eta}}_{k|K}) d\boldsymbol{x}_{-j}$, $\zeta_{1j}(x_j) = -\int \sum_{k=1}^K P_{k,j}(\boldsymbol{\beta}^*; \widetilde{\boldsymbol{\eta}}_{k|K}) d\boldsymbol{x}_{-j}$, $\zeta_{0,j}(x_j) = -\check{\xi}_0^* \int \overline{Q}_{K,00}^*(\boldsymbol{x}) d\boldsymbol{x}_{-j}$ and $\zeta_{0,1j} = -\check{\xi}_0^* \int \overline{Q}_{K,0j}^*(\boldsymbol{x}) d\boldsymbol{x}_{-j}$. Then we have the following expression for $\check{\boldsymbol{\xi}}$, $\check{\boldsymbol{\xi}} = T\check{\boldsymbol{\xi}} + \boldsymbol{\tau}$, where $\boldsymbol{\tau} = \boldsymbol{\zeta} + \boldsymbol{\zeta}_0$. By iterative applications, we have

$\check{\boldsymbol{\xi}}^s = \sum_{i=0}^{\infty} T^i \boldsymbol{\tau}^s$ for $s = A, B$. Based on these stochastic expansions, one can use the techniques in the proof of Theorem 1–2 in Mammen and Nielsen (1999) to obtain the following result: for $x_j \in (0,1)$,

$$\sup \left| \check{\xi}_j^A(x_j) - \left(\zeta_j(x_j) - \check{\xi}_0\right) \right| = o_p(N_K^{-2/5}),$$
$$\sup \left| \check{\xi}_j^B(x_j) - b_j(x_j) \right| = o_p(N_K^{-2/5}),$$

and the asymptotic distribution of $\check{\boldsymbol{\beta}}^*$ follows from the standard theory of kernel smoothing.

□

Now we present the proof of Lemma 3, which is a product of the pre-specified bandwidth order in Assumption (A5).

Proof. Noting that $\check{\boldsymbol{F}}_K$ differs from $\widehat{\boldsymbol{F}}_K$ only in the bandwidths, the proof of Lemma 6 of Yu et al. (2008) can be carried over by substituting $\rho_{Kj,i}$ for h_j^i. By Assumption (A5), we have

$$\|\check{\boldsymbol{F}}_K \check{\boldsymbol{\beta}}_K^*\|_{p,0} = o_p\left(N_K^{-2/5}\right), \quad \|\check{\boldsymbol{F}}_K \check{\boldsymbol{\beta}}_K^*\|_{\infty} = o_p\left(N_K^{-2/5}\right).$$

We note that $\widetilde{\boldsymbol{F}}_K$ is the linear approximation of $\check{\boldsymbol{F}}_K$, then $\|\check{\boldsymbol{F}}_K \check{\boldsymbol{\beta}}_K^*\| \simeq \|\widetilde{\boldsymbol{F}}_K \check{\boldsymbol{\beta}}_K^*\|$ for norms $\|\cdot\|_{p,0}$ and $\|\cdot\|_{\infty}$, which gives

$$\|\widetilde{\boldsymbol{F}}_K \check{\boldsymbol{\beta}}_K^*\|_{p,0} = o_p\left(N_K^{-2/5}\right), \quad \|\widetilde{\boldsymbol{F}}_K \check{\boldsymbol{\beta}}_K^*\|_{\infty} = o_p\left(N_K^{-2/5}\right). \tag{3}$$

Finally, we can use the same techniques in proof of Theorem 1 by substituting $\check{\boldsymbol{\beta}}_K^*$ for $\widetilde{\boldsymbol{\beta}}_K^{\lfloor m \rfloor}$ to prove that the sufficient conditions of Lemma 1 hold when setting $\boldsymbol{\beta}_0 = \check{\boldsymbol{\beta}}_K^*$. This implies that $\check{\boldsymbol{\beta}}_K$ must be close to $\widetilde{\boldsymbol{\beta}}_K$ at the same rate as (3), which complete the proof. □

S.3.3 Proof of Proposition 1

Proof. Using the same technique in the proof of Theorem 2, we can derive that

$$E\{(\widetilde{\theta}_{Kj} - \theta_j)^2\} = \left(A_{1j}\rho_{Kj,2}^{\theta} + A_{2j}N_K^{-1}\rho_{Kj,-5}^{\theta}\right)^2 + A_{3j}N_K^{-2}\rho_{Kj,-9}^{\theta}$$

$$+ o_p\left(\left\{\rho^{\theta}_{Kj,2} + N_K^{-1}\rho^{\theta}_{Kj,-5}\right\}^2 + N_K^{-2}\rho^{\theta}_{Kj,-9}\right),$$

where A_{1j}, A_{2j}, A_{3j} are constants $\boldsymbol{\beta}^*$ and

$$\rho^{\theta}_{Kj,i} = \frac{1}{N_K}\sum_{k=1}^{K} n_k(\widetilde{\eta}^{\theta}_{k|K,j})^i,\ i = 2, -5, -9.$$

Using the same argument as S.3.4, when (21) and (22) hold, the online estimates $\widetilde{\theta}_{Kj}$ satisfies $\widetilde{\theta}_{Kj} - \theta_j = O_p\big(N_K^{-2/7}\big)$. Substituting the expression of $\widetilde{h}_{Kj}$ in (16) and h^*_{Kj} in (14), we obtain

$$\frac{\widetilde{h}_{Kj} - h^*_{Kj}}{h^*_{Kj}} = \frac{1}{5}\left(\frac{\theta_j}{\widetilde{\theta}_{Kj}}\right)^{\frac{1}{5}}\left(\frac{\widetilde{\sigma}^2_{K,j} - \sigma^2_j}{\sigma^2_j}\right) - \frac{1}{5}\frac{\widetilde{\theta}_{Kj} - \theta_j}{\theta_j}.$$

When $h^{\sigma}_{Kj} = O_p(N_K^{-1/5})$, $\widetilde{\sigma}^2_{Kj}(u) - \sigma^2_j(u) = O_p(N_K^{-2/5})$ and then $\widetilde{\sigma}^2_{K,j} - \sigma^2_j = O_p(N_K^{-2/5})$, and hence the convergence rate of $\widetilde{h}_{Kj}$ is dominated by $\widetilde{\theta}_{Kj} - \theta_j$. □

S.3.4 Proof of Theorem 3

Proof. This result follows directly from Theorem 4 in Yang and Yao (2022) based on the oracle properties of the proposed estimate in Theorem 2. The optimal bandwidth h^*_{Kj} as in (14) is strictly decreasing with K. Thus the candidate sequence shall be decreasing with $\eta_{kj1} = \widetilde{h}_{kj}$. There is no efficiency loss if $\widetilde{\eta}_{k|K,j} = \widehat{h}_{Kj}$ which implies that the optimal η_{lj} shall make $\widetilde{\eta}_{k|K,j}$ as close to $\widehat{h}_{Kj}$ as possible. With the convergence of $\widetilde{h}_{kj}$, we have

$$\frac{\widetilde{\eta}_{k|K,j}}{\widehat{h}_{Kj}} = g(l)\left(\frac{N_k}{N_K}\right)^{-\frac{1}{5}} + O_p\left(N_k^{-\frac{2}{7}}\right).$$

Write $g(l) = \{g(l)^{1/\lambda}\}^{\lambda}$ and note that N_k/N_K grows linearly, then λ shall be $1/5$ and the optimal $g(l)^{1/\lambda}$ shall be linear between $(0, 1)$. Hence, the optimal η is

$$\eta_{Kj\ell} = \left(\frac{L - \ell + 1}{L}\right)^{\frac{1}{5}}\widetilde{h}_{Kj}.$$

Next we derive the asymptotic lower bound for the relative efficiency. Writing $\widehat{h}_{Kj} = h^*_{Kj}\{1 + O_p(N_K^{-2/7})\}$ and using the expressions of h^*_{Kj} as in (14), one can derive

$$eff(\widetilde{\beta}_{Kj})^{-1} = \frac{1}{5}\left\{\sum_{k=1}^{K}\frac{n_k}{N_K}\left(\frac{\widetilde{\eta}_{k|K,j}}{\widetilde{h}_{Kj}}\right)^2\right\}^2 + \frac{4}{5}\left\{\sum_{k=1}^{K}\frac{n_k}{N_K}\left(\frac{\widetilde{\eta}_{k|K,j}}{\widetilde{h}_{Kj}}\right)^{-1}\right\} + O_p\left(N_K^{-\frac{2}{7}}\right). \quad (4)$$

As shown in Figure 3, when K tends large, there exists a breakpoint K_0: when $k \le K_0$, $\widetilde{\eta}_{k|K,j}$ equals the last candidate $(1/L)^{1/(d+4)}\widetilde{h}_{kj}$ for the limit number of candidates and for $k > K_0$, there are sufficient and close candidates to make a choice. From the property of the breakpoint we see that $\widetilde{\eta}_{K_0|K} \le \widetilde{h}_{Kj}$, thus $N_{K_0} \ge N_K/L$. When $k > K_0$, the combination rule guarantees

$$\frac{\widetilde{\eta}_{k|K,j}}{\widetilde{h}_{Kj}} = \left(\frac{L-l+1}{L}\frac{N_K}{N_k}\right)^{\frac{1}{5}} = 1 + O_p\left(\frac{N_{K_0}^{-\frac{1}{5}}}{L}\right) = 1 + O_p\left(\frac{N_K^{-\frac{1}{5}}}{L}\right),$$

which gives

$$\sum_{k=K_0+1}^{K}\frac{n_k}{N_K}\left(\frac{\widetilde{\eta}_{k|K,j}}{\widetilde{h}_{Kj}}\right)^2 = 1 - \frac{N_{K_0}}{N_K} + O_p\left(\frac{N_K^{-\frac{1}{5}}}{L}\right),$$
$$\sum_{k=K_0+1}^{K}\frac{n_k}{N_K}\left(\frac{\widetilde{\eta}_{k|K,j}}{\widetilde{h}_{Kj}}\right)^{-1} = 1 - \frac{N_{K_0}}{N_K} + O_p\left(\frac{N_K^{-\frac{1}{5}}}{L}\right).$$

When $k \le K_0$,

$$\frac{\widetilde{\eta}_{k|K,j}}{\widetilde{h}_{Kj}} = L^{-\frac{1}{5}} \cdot \left(\frac{N_k}{N_K}\right)^{-\frac{1}{5}}.$$

It can be derived that

$$\sum_{k=1}^{K_0}\frac{n_k}{N_K}\left(\frac{\widetilde{\eta}_{k|K,j}}{\widetilde{h}_{Kj}}\right)^2 = \frac{5}{3}\cdot L^{-\frac{2}{5}}\left(\frac{N_{K_0}}{N_K}\right)^{1-\frac{2}{5}},$$
$$\sum_{k=1}^{K_0}\frac{n_k}{N_K}\left(\frac{\widetilde{\eta}_{k|K,j}}{\widetilde{h}_{Kj}}\right)^{-1} = \frac{5}{6}\cdot L^{\frac{1}{5}}\left(\frac{N_{K_0}}{N_K}\right)^{1+\frac{1}{5}}.$$

Denote $\rho_0 = N_{K_0}/N_K$, one can derive from (4) that

$$eff(\widetilde{\beta}_{Kj})^{-1} = \frac{1}{5}\left\{\frac{5}{3}\cdot L^{-\frac{2}{5}}\rho_0^{1-\frac{2}{5}} + 1 - \rho_0\right\}^2$$

$$+\frac{4}{5}\left\{\frac{5}{6}\cdot L^{\frac{1}{5}}\rho_0^{1+\frac{1}{5}}+1-\rho_0\right\}+O_p\left(\frac{N_K^{-\frac{1}{5}}}{L}+N_K^{-\frac{2}{7}}\right). \tag{5}$$

The property of breakpoint K_0 also guarantee that

$$\left(\frac{1}{L}\right)^{\frac{1}{5}}\sum_{k=1}^{K_0}\frac{n_k}{N_{K_0}}\widetilde{h}_{kj}\geq\widetilde{h}_{Kj},$$

holds with probability 1 when K tends large, which is equivalent to

$$\left(\frac{1}{L}\right)^{\frac{1}{5}}\sum_{k=1}^{K_0}\left\{\frac{n_k}{N_{K_0}}\cdot\frac{\widetilde{C}_k N_k^{-\frac{1}{5}}}{\widetilde{C}_K N_K^{-\frac{1}{5}}}\right\}\geq 1.$$

Under Assumption (A6), we obtain

$$\frac{N_{K_0}}{N_K}\leq\left(\frac{5}{4}\right)^5\frac{1}{L}+O_p(N_K^{-\frac{2}{7}}),$$

i.e. $\rho_0\in[0,(5/4)^5/L]$ holds with probability 1 when K tends large. Note that (5) is strictly increasing with respect to ρ_0 on this domain. Hence we have

$$eff(\widetilde{\beta}_{Kj})^{-1}\leq 1+c_1L^{-1}+c_2L^{-2},$$

where

$$c_1=5^3/(3\times 2^5)+5^6/(6\times 4^5)-6/5\times(5/4)^5\approx 0.183$$

and

$$c_2=\{5^7/(6\times 4^6)-(5/4)^5\}^2/5\approx 0.0032.$$

This completes the proof of Theorem 3. □

References

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A. and Ritov, Y. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Vol. 4, Springer.

Blum, A. and Mansour, Y. (2007), 'From external to internal regret.', *Journal of Machine Learning Research* **8**(6).

Chatla, S. B. (2022), 'Nonparametric inference for additive models estimated via simplified smooth backfitting', *Annals of the Institute of Statistical Mathematics* pp. 1–27.

Chen, X., Linton, O. and Van Keilegom, I. (2003), 'Estimation of semiparametric models when the criterion function is not smooth', *Econometrica* **71**(5), 1591–1608.

Fan, J. and Jiang, J. (2005), 'Nonparametric inferences for additive models', *Journal of the American Statistical Association* **100**(471), 890–907.

Kong, E. and Xia, Y. (2019), 'On the efficiency of online approach to nonparametric smoothing of big data', *Statistica Sinica* **29**(1), 185–201.

Li, Y., Chen, X. and Li, N. (2019), 'Online optimal control with linear dynamics and predictions: Algorithms and regret analysis', *Advances in Neural Information Processing Systems* **32**.

Mammen, E. and Nielsen, J. (1999), 'The existence and asymptotic properties of a back-fitting projection algorithm under weak conditions', *The Annals of Statistics* **27**(5), 49.

Quan, M. and Lin, Z. (2022), 'Optimal one-pass nonparametric estimation under memory constraint', *Journal of the American Statistical Association* (just-accepted), 1–32.

Warmuth, M. K. and Kuzmin, D. (2008), 'Randomized online pca algorithms with regret bounds that are logarithmic in the dimension', *Journal of Machine Learning Research* **9**(Oct), 2287–2320.

Xue, D. and Yao, F. (2022), 'Dynamic penalized splines for streaming data', *Statistica Sinica* **32**, 1363–1380.

Yang, Y. and Yao, F. (2022), 'Online estimation for functional data', *Journal of the American Statistical Association* pp. published online, https://doi.org/10.1080/01621459.2021.2002158.

Yu, K., Park, B. U. and Mammen, E. (2008), 'Smooth backfitting in generalized additive models', *The Annals of Statistics* **36**(1), 228–260.