where

$$D = \begin{bmatrix} \int_{\mathbb{R}} M(x)dx & 0 \\ 0 & \operatorname{diag}\left(\begin{bmatrix} \int_{\mathbb{R}} M_j M_j^{\prime} dx \\ \vdots \\ M_j M_{jj} \end{bmatrix}\right)_{j,\ldots,d} \end{bmatrix}; \quad A = \begin{bmatrix} 1 & 0^{\prime} \\ 0 & \Pi \end{bmatrix}.$$

Since $g^{0}(\mathbf{e}_K(\ ))\ V(\mathbf{e}_K(\ ))$ is bounded and $\mathbf{p}_K(\ )$ is bounded away from 0, one can obtain that D is bounded. Using the same technique, we can prove that Π is one-to-one and onto, and $k\Pi\ k\ \ dk\ k$. Hence Π has a bounded inverse, which implies that A also has, i.e. there exists a constant c such that $k\mathbf{F}_K(\mathbf{e}_K^{b\ c})k < c$.

The other conditions are veri ed as follows. To prove that $\mathbf{F}_K$ satis es the Lipschitz condition $k\mathbf{F}_K\ \ \mathbf{F}_K\ \ {}^{0}k\ \ c\ k\ \ \ {}^{0}k$, by the decomposition (2), it is su cient to prove the boundedness of D and A, which is guaranteed by the smoothness and boundedness of $g^{0}(\mathbf{e}_K(\ ))\ V(\mathbf{e}_K(\ ))$ and $\mathbf{p}_K(\ )$. Finally, by the uniform continuity of $\mathbf{F}_K$ and the fact $\mathbf{F}_K\mathbf{e}_K = 0$, there exists a positive constant r such that $\sup_{\beta 2 B_r\ \tilde{\beta}_K} k\mathbf{F}_K\ \ k < (2c\ c\ )$. This proves that if $\mathbf{e}_K^{b\ c}\ 2\ B_r(\mathbf{e}_K)$, then $kf\mathbf{F}_K(\mathbf{e}_K^{b\ c})g\ \ \mathbf{F}_K(\mathbf{e}_K^{b\ c})k < (2c\ c\ )$. Note that M is close to $\overline{Q}_K$, the two norms $k\ \ k_M$ and $k\ \ k_{\overline{Q}_K}$ are equivalent. Conclusion (1) now follows from Lemma 1. Theorem 1 (2) follows from the property of Hilbert-Schmidt operators Π_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 rst prove Lemma 2. Recall that is de ned in (23). The proof is based on the decomposition of .

Proof. We further decompose $_j(x_j) = {}^A_j(x_j) + {}^B_j(x_j)$; $_j(x_j) = {}^A_j(x_j) + {}^B_j(x_j)$ that

$$\begin{bmatrix} {}^A_j(x_j) \\ {}^A_j(x_j) \end{bmatrix} = \begin{bmatrix} {}_j(x_j) \\ {}_j(x_j) \end{bmatrix} \begin{bmatrix} E(\,_j(x_j)\, j\, X\,;\,:\,:\,:\,;\, X_K) \\ E(\,_j(x_j)\, j\, X\,;\,:\,:\,:\,;\, X_K) \end{bmatrix};$$

$$\begin{bmatrix} {}^B_j(x_j) \\ {}^B_j(x_j) \end{bmatrix} = \begin{bmatrix} E(\,_j(x_j)\, j\, X\,;\,:\,:\,:\,;\, X_K) \\ E(\,_j(x_j)\, j\, X\,;\,:\,:\,:\,;\, X_K) \end{bmatrix}:$$

We omit the constants in the functions and de ne the spaces

$$\overline{H}\,(Q) = \quad f = (f\,(x\,);\,:\,:\,:\,;\, f_d(x_d);\, f\,(x\,);\, f\,_d(x_d))^{>}:$$

$$\int (Q\quad f_j + Q\,_j f\,_j)dx = 0 \text{ for } j = 1;\,:\,:\,:\,;\, d\quad;$$

$$\overline{H}_j(Q) = f f\ 2\,\overline{H}\,(Q) : f \text{ depends only on } x_j g:$$

In this proof, we consider $2d$-dimensional functions. Let $_j$ be the projection operator from $\overline{H}\,(Q_K)$ onto $\overline{H}_j(Q_K)$. Then for any $2\,\overline{H}\,(Q_K)$,

$$_j\ _\ell(x_\ell) = f_\ell(x_\ell); \qquad _j\ _\ell(x_\ell) = f\,_\ell(x_\ell);$$

where for ` $6 = j$, $f_\ell(x_\ell) = \,_\ell(x_\ell)$ and $f\,_\ell(x_\ell) = \,_\ell(x_\ell)$, and

$$f_j(x_j) = g_j(x_j) \int g_j(x_j)\overline{Q}_{K,}\,(x)dx;$$

$$\begin{bmatrix} g_j(x_j) \\ f\,_j(x_j) \end{bmatrix} = \begin{bmatrix} \overline{Q}_{K,} & \overline{Q}_{K,\,j} \\ \overline{Q}_{K,j} & \overline{Q}_{K,jj} \end{bmatrix} dx\,_j \int \begin{bmatrix} \overline{Q}_{K,} & \overline{Q}_{K,\,\ell} \\ \overline{Q}_{K,j} & \overline{Q}_{K,j\ell} \end{bmatrix} \begin{bmatrix} {}_\ell(x_\ell) \\ {}_\ell(x_\ell) \end{bmatrix} dx\,_j:$$

Then de ne $T = \quad _d$. Recall that $!_{ki}(x;\ )$ plays the same role of $K_h(X_{ki}\quad x)$ 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 < 1 w.p.1. Let $_j(x_j) = \int P^K_k\ P_{k,}(\quad ; e_{kjK})dx\,_j$, $_j(x_j) = \int P^K_k\ P_{k,j}(\quad ; e_{kjK})dx\,_j$, $_{,j}(x_j) = \int \overline{Q}_{K,}\,(x)dx\,_j$ and $_{,\,j} = \int \overline{Q}_{K,\,j}(x)dx\,_j$. Then we have the following expression for $,\quad = T\quad +\quad$, where $\quad = \quad +\quad$. By iterative applications, we have

$^{s} = \sum_{i}^{1} T^{i}\ ^{s}$ 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 2 (0; 1),

$$\sup\ {}_{j}^{A}(x_j) \quad {}_{j}(x_j) \quad = o_p(N_K{}^{/}\ );$$

$$\sup\ {}_{j}^{B}(x_j) \quad b_j(x_j) \ = o_p(N_K{}^{/}\ );$$

and the asymptotic distribution of follows from the standard theory of kernel smoothing.

□

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

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

$$kF_K \quad {}_K k_{p,} \ = o_p \quad N_K{}^{/} \quad ; \quad kF_K \quad {}_K k_1 \ = o_p \quad N_K{}^{/} \quad :$$

We note that $\hat{F}_K$ is the linear approximation of F_K, then $kF_K \quad {}_K k \ ' \ k\hat{F}_K \quad {}_K k$ for norms $k \quad k_{p,}$ and $k \quad k_1$, which gives

$$k\hat{F}_K \quad {}_K k_{p,} \ = o_p \quad N_K{}^{/} \quad ; \quad k\hat{F}_K \quad {}_K k_1 \ = o_p \quad N_K{}^{/} \quad : \tag{3}$$

Finally, we can use the same techniques in proof of Theorem 1 by substituting $_K$ for e_K^{bmc} to prove that the su cient conditions of Lemma 1 hold when setting $= \ _K$. This implies that $_K$ must be close to e_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

$$E f(e_{Kj} \quad {}_j)\ g = \ A \ {}_j \ {}^{\theta}_{Kj,} \ + A \ {}_j N_K \quad {}^{\theta}_{Kj,} \quad + A \ {}_j N_K \quad {}^{\theta}_{Kj,}$$

$$+ o_p \quad {}^{\theta}_{Kj,} + N_K \quad {}^{\theta}_{Kj,} \quad + N_K \quad {}^{\theta}_{Kj,} \quad ;$$

where $A \ _j; A \ _j; A \ _j$ are constants and

$$ {}^{\theta}_{Kj,i} = \frac{1}{N_K} \sum_k n_k (\hat{e}^{\theta}_{kjK,j})^i; \ i = 2; \ 5; \ 9:$$

Using the same argument as S.3.4, when (21) and (22) hold, the online estimates $\hat{e}_{Kj}$ satis es $\hat{e}_{Kj} \quad _j = O_p \ N_K \ ^{/} \ $. Substituting the expression of $\hat{h}_{Kj}$ in (16) and h_{Kj} in (14), we obtain

$$\frac{\hat{h}_{Kj} \quad h_{Kj}}{h_{Kj}} = \frac{1}{5} \left(\frac{ \ _j}{\hat{e}_{Kj}} \right)^{\frac{1}{5}} \frac{\hat{e}_{K,j} \quad _j}{ \ _j} \quad \frac{1}{5} \frac{\hat{e}_{Kj} \quad _j}{ \ _j}:$$

When $h^{\sigma}_{Kj} = O_p(N_K \ ^{/} \ )$, $\hat{e}_{Kj}(u) \quad _j(u) = O_p(N_K \ ^{/} \ )$ and then $\hat{e}_{K,j} \quad _j = O_p(N_K \ ^{/} \ )$, and hence the convergence rate of $\hat{h}_{Kj}$ is dominated by $\hat{e}_{Kj} \quad _j$. □

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 $ \ _{kj} = \hat{h}_{kj}$. There is no e ciency loss if $\hat{e}_{kjK,j} = \hat{h}_{Kj}$ which implies that the optimal $ \ _{lj}$ shall make $\hat{e}_{kjK,j}$ as close to $\hat{h}_{Kj}$ as possible. With the convergence of $\hat{h}_{kj}$, we have

$$\frac{\hat{e}_{kjK,j}}{\hat{h}_{Kj}} = g(l) \left(\frac{N_k}{N_K} \right)^{\frac{1}{5}} + O_p \left(N_k \ ^{\frac{2}{7}} \right) :$$

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

$$ \ _{Kj\ell} = \left(\frac{L \quad \ell + 1}{L} \right)^{\frac{1}{5}} \hat{h}_{Kj}:$$

Next we derive the asymptotic lower bound for the relative e ciency. Writing $\hat{h}_{Kj} = h_{Kj}\{1 + O_p(N_K^{\ /})\}$ and using the expressions of h_{Kj} as in (14), one can derive

62 0 7.141Td [107]TJ/F10 11.9593Tf 4.234 0 Td15Td [107]TJ/F1+- 11.9552 Tf 147.886 0
K5]TJ/F29 7.9701 Tf 10.623 6.11 T2Tf 243]TJ/F2/F23 7.97014 Td [(62 0 7.782T [(59]TJ/F2

$$eff(e_{Kj})$$

$$+\frac{4}{5}\quad\frac{5}{6}\quad L^{\frac{1}{5}}\quad{}^{\frac{1}{5}}+1\qquad+O_p\quad\frac{N_K{}^{\frac{1}{5}}}{L}+N_K{}^{\frac{2}{7}}\quad\Big!\ : \tag{5}$$

The property of breakpoint K also guarantee that

$$\frac{1}{L}\quad{}^{\frac{1}{5}}\ \overset{K_0}{\underset{k}{X}}\ \frac{n_k}{N_{K_0}}\ \hat{e}_{kj}\quad\hat{e}_{Kj};$$

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

$$\frac{1}{L}\quad{}^{\frac{1}{5}}\ \overset{K_0}{\underset{k}{X}}\ \frac{n_k}{N_{K_0}}\ \frac{\hat{e}_k N_k{}^{\frac{1}{5}}}{\hat{e}_K N_K{}^{\frac{1}{5}}}\quad 1:$$

Under Assumption (A6), we obtain

$$\frac{N_{K_0}}{N_K}\qquad\frac{5}{4}\quad\frac{1}{L}+O_p(N_K{}^{\frac{2}{7}});$$

i.e. $2\ [0;(5=4)\ =L]$ holds with probability 1 when K tends large. Note that (5) is strictly increasing with respect to on this domain. Hence we have

$$eff(\hat{e}_{Kj})\qquad 1+c\ L\quad+c\ L\quad;$$

where

$$c\ =5\ =(3\quad 2\ )+5\ =(6\quad 4\ )\quad 6=5\quad(5=4)\qquad 0{:}183$$

and

$$c\ =f5\ =(6\quad 4\ )\quad(5=4)\ g\ =5\quad 0{:}0032:$$

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 simpli ed smooth back tting', *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 e ciency 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- tting 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 back tting in generalized additive models', *The Annals of Statistics* **36**(1), 228{260.