On-line supplementary material for 'Generalized additive and index models with shape constraints'

Yining Chen and Richard J. Samworth†

University of Cambridge, UK

1. Proofs: main results

PROOF OF PROPOSITION 1

Define the set

$$\Theta = \{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T \in \bar{\mathbb{R}}^n \mid \exists f \in \mathrm{cl}(\mathcal{F}) \text{ s.t. } \eta_i = f(\mathbf{X}_i), \forall i = 1, \ldots, n\}.$$

We can rewrite the optimization problem as finding $\hat{\boldsymbol{\eta}}_n$ such that

$$\hat{\boldsymbol{\eta}}_n \in \underset{\boldsymbol{\eta} \in \Theta}{\mathrm{argmax}}\, \bar{\ell}_n(\boldsymbol{\eta}),$$

where $\bar{\ell}_n(\boldsymbol{\eta}) = \frac{1}{n}\sum_{i=1}^n \ell_i(\eta_i)$, and where

$$\ell_i(\eta_i) = \begin{cases} Y_i\eta_i - B(\eta_i), & \text{if } \eta_i \in \mathrm{dom}(B); \\ \lim_{a \to -\infty} Y_i a - B(a), & \text{if } \eta_i = -\infty; \\ \lim_{a \to \infty} Y_i a - B(a), & \text{if the EF is Gaussian, Poisson or Binomial, and } \eta_i = \infty; \\ -\infty, & \text{if the EF is Gamma and } \eta_i \in [0, \infty]. \end{cases}$$

Note that $\bar{\ell}_n$ is continuous on the non-empty set Θ and $\sup_{\boldsymbol{\eta} \in \Theta} \bar{\ell}_n(\boldsymbol{\eta})$ is finite. Moreover, by Lemma 3 in Section 2, Θ is a closed subset of the compact set $\bar{\mathbb{R}}^n$, so is compact. It follows that $\bar{\ell}_n$ attains its maximum on Θ, so $\hat{S}_n \neq \emptyset$.

†*Address for correspondence:* Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, UK. CB3 0WB.

E-mail: {Y.Chen, R.Samworth}@statslab.cam.ac.uk

To show the uniqueness of $\hat{\boldsymbol{\eta}}_n$, we now suppose that both $\boldsymbol{\eta}_1 = (\eta_{11}, \ldots, \eta_{1n})^T$ and $\boldsymbol{\eta}_2 = (\eta_{21}, \ldots, \eta_{2n})^T$ maximize $\bar{\ell}_n$. The only way we can have $\eta_{1i} = \infty$ is if the family is Binomial and $Y_i = 1$. But then $\ell_i(-\infty) = -\infty$, so we cannot have $\eta_{2i} = -\infty$. It follows that $\boldsymbol{\eta}_* = (\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)/2$ is well-defined, and $\boldsymbol{\eta}_* \in \Theta$, since Θ is convex. Now we can use the strict concavity of $\bar{l}_n$ on its domain to conclude that $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2 = \boldsymbol{\eta}_*$.

To prove Theorem 1, we require the following lemma, which says (roughly) that if any of the additive components (or the intercept) of $f \in \mathcal{F}$ are large somewhere, then there is a non-trivial region on which either f is large, or $-f$ is large.

LEMMA 1. *Fix $a > 0$. There exists a finite collection $\mathcal{C}_a$ of disjoint compact subsets of $[-2a, 2a]^d$ each having Lebesgue measure at least $\left(\frac{a}{2d}\right)^d$, such that for any $f \overset{\mathcal{F}}{\sim} (f_1, \ldots, f_d, c)$,*

$$\max_{C \in \mathcal{C}_a} \max\Big\{\inf_{\mathbf{x} \in C} f(\mathbf{x}), \inf_{\mathbf{x} \in C} -f(\mathbf{x})\Big\} \geq \frac{1}{4} \max\Big\{\sup_{|x_1| \leq a} |f_1(x_1)|, \ldots, \sup_{|x_d| \leq a} |f_d(x_d)|, 2|c|\Big\}.$$

PROOF. Let $\max\{\sup_{|x_1| \leq a} |f_1(x_1)|, \ldots, \sup_{|x_d| \leq a} |f_d(x_d)|, 2|c|\} = M$ for some $M \geq 0$. Recalling that $f_1(0) = \ldots = f_d(0) = 0$, and owing to the shape restrictions, this is equivalent to

$$\max\{|f_1(-a)|, |f_1(a)|, \ldots, |f_d(-a)|, |f_d(a)|, 2|c|\} = M.$$

We will prove the lemma by construction. For $j = 1, \ldots, d$, consider the collection of intervals

$$\mathcal{D}_j = \begin{cases} \Big\{[-2a, -a], [a, 2a]\Big\}, & \text{if } l_j \in \{2, 3, 5, 6, 8, 9\}; \\ \Big\{[-2a, -a], [-a/(4d), a/(4d)], [a, 2a]\Big\}, & \text{if } l_j \in \{1, 4, 7\}. \end{cases}$$

Let $\mathcal{C}_a = \{\times_{j=1}^d D_j : D_j \in \mathcal{D}_j\}$, so that $|\mathcal{C}_a| \leq 3^d$. The two cases below validate our construction:

(a) $\max\{|f_1(-a)|, |f_1(a)|, \ldots, |f_d(-a)|, |f_d(a)|\} < M$. Then it must be the case that $|c| = M/2$ and, without loss of generality, we may assume $c = M/2$. For

$j = 1, \ldots, d$, if $l_j \in \{2, 3, 5, 6, 8, 9\}$, then due to the monotonicity and the fact that $f_j(0) = 0$, either

$$\inf_{x_j \in [-2a, -a]} f_j(x_j) \geq 0 \quad \text{or} \quad \inf_{x_j \in [a, 2a]} f_j(x_j) \geq 0.$$

For $l_j \in \{1, 4, 7\}$, by the convexity/concavity, $\sup_{x_j \in [-a/(4d), a/(4d)]} |f_j(x_j)| \leq M/(4d)$. Hence

$$\max_{C \in \mathcal{C}_a} \inf_{\mathbf{x} \in C} f(\mathbf{x}) \geq -dM/(4d) + M/2 = M/4.$$

(b) $\max\{|f_1(-a)|, |f_1(a)|, \ldots, |f_d(-a)|, |f_d(a)|\} = M$ and $|c| \leq M/2$. Without loss of generality, we may assume that $f_1(-a) = M$. Since $f_1(0) = 0$ and $|f_1(a)| \leq M$, we can assume $l_1 \in \{1, 3, 4, 6\}$. Therefore, $\inf_{x_1 \in [-2a, -a]} f_1(x_1) = M$. Let

$$D_j = \begin{cases} [-a/(4d), a/(4d)], & \text{if } l_j \in \{1, 4, 7\} \\ [a, 2a], & \text{if } l_j \in \{2, 5, 8\} \\ [-2a, -a], & \text{if } l_j \in \{3, 6, 9\} \end{cases}$$

for $j = 2, \ldots, d$. Now for $C = [-2a, -a] \times \times_{j=2}^{d} D_j$, we have

$$\inf_{\mathbf{x} \in C} f(\mathbf{x}) \geq M - (d-1)M/(4d) - M/2 \geq M/4.$$

Proof of Theorem 1

For convenience, we first present the proof of consistency in the case where the EF distribution is Binomial. Note that for the Binomial family the response Y is scaled to take values in $\{0, 1/T, 2/T, \ldots, 1\}$ for some known $T \in \mathbb{N}$, where T is the total number of trials. Consistency for the other EF distributions listed in Table 2 in the main text can be established using essentially the same proof structure with some minor modifications. We briefly outline these changes at the end of the proof.

Since the proof is rather long, we give here a brief, high-level description of the main ideas. In Step 1, we give a lower bound $\bar{L}_0$ for the scaled partial log-likelihood $\bar{\ell}_n(\cdot)$ in the limit as $n \to \infty$. This is helpful, because it allows us to conclude that any $f \in \mathrm{cl}(\mathcal{F})$ for which $\limsup_{n \to \infty} \bar{\ell}_n(f) \leq \bar{L}_0 - 1$, say, cannot belong to $\hat{S}_n$ for large n.

In particular, in Steps 2 and 3, we use this technique to show that when n is large, any element of $\hat{S}_n$ must be bounded on $[-a, a]^d$ for every $a > 0$ (independent of n) and the concave/convex components must be Lipschitz. In Step 4, we show that our reduced class of functions is a Glivenko–Cantelli class (i.e. satisfies a uniform law of large numbers). Since the population-level scaled partial log-likelihood is uniquely maximized at the true f_0, we can then conclude in Step 5 that when n is large, any maximizer of $\bar{\ell}_n(\cdot)$ over $\mathrm{cl}(\mathcal{F})$ must belong to a small ball around f_0, as required.

Step 1: Lower bound for the scaled partial log-likelihood. It follows from Assumption **1)** and the strong law of large numbers that

$$\liminf_{n\to\infty} \sup_{f\in\mathrm{cl}(\mathcal{F})} \bar{\ell}_n(f) \geq \lim_{n\to\infty} \bar{\ell}_n(f_0) = \mathbb{E}\{g^{-1}(f_0(\mathbf{X}))f_0(\mathbf{X}) - B(f_0(\mathbf{X}))\}$$

almost surely. We define $\bar{L}_0 = \mathbb{E}\{g^{-1}(f_0(\mathbf{X}))f_0(\mathbf{X}) - B(f_0(\mathbf{X}))\}$.

Step 2: Bounding $|\hat{f}_n|$ on $[-a, a]^d$ for any fixed $a > 0$. For $M > 0$, let

$$\mathcal{F}_{a,M} = \left\{ f \overset{\mathcal{F}}{\sim} (f_1, \ldots, f_d, c) : \max\{|f_1(-a)|, |f_1(a)|, \ldots, |f_d(-a)|, |f_d(a)|, 2|c|\} \leq M \right\}. \tag{1}$$

We will prove that there exists a deterministic constant $M = M(a) \in (0, \infty)$ such that, with probability one, we have $\hat{S}_n \subseteq \mathrm{cl}(\mathcal{F}_{a,M(a)})$ for sufficiently large n. To this end, let $\mathcal{C}_a = \{C_1, \ldots, C_N\}$ be the finite collection of compact subsets of $[-2a, 2a]^d$ constructed in the proof of Lemma 1, and set

$$M = 4B^{-1}\left(\frac{-\bar{L}_0 + 1}{\min_{1\leq k\leq N, t\in\{0,1\}} \mathbb{P}(\mathbf{X} \in C_k, Y = t)} \right).$$

Let

$$k_* \in \operatorname*{argmin}_{1\leq k\leq N} \limsup_{n\to\infty} \sup_{f\in\mathrm{cl}(\mathcal{F}\setminus\mathcal{F}_{a,M})} \frac{1}{n}\sum_{i=1}^{n}\{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\}1_{\{\mathbf{X}_i\in C_k \cap Y_i\in\{0,1\}\}}.$$

Then

$$
\begin{aligned}
&\limsup_{n\to\infty} \sup_{f\in \mathrm{cl}(\mathcal{F}\backslash\mathcal{F}_{a,M})} \bar{\ell}_n(f) \\
&\leq \limsup_{n\to\infty} \sup_{f\in \mathrm{cl}(\mathcal{F}\backslash\mathcal{F}_{a,M})} \frac{1}{n}\sum_{i=1}^{n}\{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\}\mathbb{1}_{\{\mathbf{X}_i\notin C_{k_*}\cup Y_i\notin\{0,1\}\}} \\
&\quad + \limsup_{n\to\infty} \sup_{f\in \mathrm{cl}(\mathcal{F}\backslash\mathcal{F}_{a,M})} \frac{1}{n}\sum_{i=1}^{n}\{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\}\mathbb{1}_{\{\mathbf{X}_i\in C_{k_*}\cap Y_i\in\{0,1\}\}} \\
&\leq \max_{1\leq k\leq N} \limsup_{n\to\infty} \sup_{f\in \mathrm{cl}(\mathcal{F}\backslash\mathcal{F}_{a,M})} \frac{1}{n}\sum_{i=1}^{n}\{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\}\mathbb{1}_{\{\mathbf{X}_i\notin C_k\cup Y_i\notin\{0,1\}\}} \qquad (2) \\
&\quad + \min_{1\leq k\leq N} \limsup_{n\to\infty} \sup_{f\in \mathrm{cl}(\mathcal{F}\backslash\mathcal{F}_{a,M})} \frac{1}{n}\sum_{i=1}^{n}\{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\}\mathbb{1}_{\{\mathbf{X}_i\in C_k\cap Y_i\in\{0,1\}\}}. \qquad (3)
\end{aligned}
$$

Now (2) is non-positive, since $Y_i\eta - B(\eta) = Y_i\eta - \log(1+e^\eta) \leq 0$ for all $\eta \in \bar{\mathbb{R}}$ and $Y_i \in \{0, 1/T, 2/T, \ldots, 1\}$. We now claim that the supremum over $f \in \mathrm{cl}(\mathcal{F}\backslash\mathcal{F}_{a,M})$ in (3) can be replaced with a supremum over $f \in \mathcal{F}\backslash\mathcal{F}_{a,M}$. To see this, let

$$\Theta_0 = \{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T \in \bar{\mathbb{R}}^n : \exists f \in \mathrm{cl}(\mathcal{F}\setminus\mathcal{F}_{a,M}) \text{ s.t. } \eta_i = f(\mathbf{X}_i), \forall i = 1, \ldots, n\}.$$

Suppose that $(\boldsymbol{\eta}^m) \in \Theta_0$ is such that the corresponding $(f^m) \in \mathrm{cl}(\mathcal{F}\setminus\mathcal{F}_{a,M})$ is a maximizing sequence in the sense that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{Y_i f^m(\mathbf{X}_i) - B(f^m(\mathbf{X}_i))\}\mathbb{1}_{\{\mathbf{X}_i\in C_k\cap Y_i\in\{0,1\}\}} \\
&\qquad\qquad \nearrow \sup_{f\in \mathrm{cl}(\mathcal{F}\backslash\mathcal{F}_{a,M})} \frac{1}{n}\sum_{i=1}^{n}\{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\}\mathbb{1}_{\{\mathbf{X}_i\in C_k\cap Y_i\in\{0,1\}\}}.
\end{aligned}
$$

By reducing to a subsequence if necessary, we may assume $\boldsymbol{\eta}^m \to \boldsymbol{\eta}^0$, say, as $m \to \infty$, where $\boldsymbol{\eta}^0 = (\eta_1^0, \ldots, \eta_n^0)^T \in \bar{\mathbb{R}}^n$. Since, for each $m \in \mathbb{N}$, we can find a sequence $(f^{m,k})_k \in \mathcal{F}\setminus\mathcal{F}_{a,M}$ such that $f^{m,k} \to f^m$ pointwise in $\bar{\mathbb{R}}$ as $k \to \infty$, it follows that we can pick $k_m \in \mathbb{N}$ such that $f^{m,k_m}(\mathbf{X}_i) \to \eta_i^0$ as $m \to \infty$, for all $i = 1, \ldots, n$. Moreover, $(\eta_1, \ldots, \eta_n) \mapsto \frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta_i - B(\eta_i)\}\mathbb{1}_{\{\mathbf{X}_i\in C_k\cap Y_i\in\{0,1\}\}}$ is continuous on $\bar{\mathbb{R}}^n$, and we deduce that $(f^{m,k_m}) \in \mathcal{F}\setminus\mathcal{F}_{a,M}$ is also a maximizing sequence, which establishes our claim.

Recall that by Lemma 1, for any $f \in \mathcal{F}\backslash\mathcal{F}_{a,M}$, we can always find $C_{k^*} \in \mathcal{C}_a$ such that

$$\max\Big\{\inf_{\mathbf{x}\in C_{k^*}} f(\mathbf{x}), \inf_{\mathbf{x}\in C_{k^*}} -f(\mathbf{x})\Big\} \geq M/4.$$

Combining the non-positivity of (2) and our argument above removing the closure in (3), we deduce by the strong law of large numbers that, almost surely,

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{f\in \mathrm{cl}(\mathcal{F}\backslash\mathcal{F}_{a,M})} \bar{\ell}_n(f) \\
&\quad \leq \max\Big\{-B(M/4)\mathbb{P}(\mathbf{X}\in C_{k^*}, Y=0), \quad \{-M/4 - B(-M/4)\}\mathbb{P}(\mathbf{X}\in C_{k^*}, Y=1)\Big\} \\
&\quad \leq -\min_{1\leq k\leq N, t\in\{0,1\}} \mathbb{P}\big(\mathbf{X}\in C_k, Y=t\big)B(M/4) = \bar{L}_0 - 1,
\end{aligned}$$

where, we have used the property that $B(t) = t + B(-t)$ for the penultimate inequality, and the definition of M for the final equality. Comparing this bound with the result of Step 1, we deduce that $\hat{S}_n \cap \mathrm{cl}(\mathcal{F}\backslash\mathcal{F}_{a,M}) = \emptyset$ for sufficiently large n, almost surely. But it is straightforward to check that $\mathrm{cl}(\mathcal{F}) = \mathrm{cl}(\mathcal{F}_{a,M}) \cup \mathrm{cl}(\mathcal{F}\setminus\mathcal{F}_{a,M})$, and the result follows.

Step 3: Lipschitz constant for the convex/concave components of $\hat{f}_n$ on $[-a,a]$. For $M_1, M_2 > 0$, let

$$\mathcal{F}_{a,M_1,M_2} = \Big\{ f \overset{\mathcal{F}}{\sim} (f_1,\ldots,f_d,c) \in \mathcal{F}_{a,M_1} : |f_j(z_1) - f_j(z_2)| \leq M_2|z_1 - z_2|, \forall z_1, z_2 \in [-a,a], \\ \forall j \text{ with } l_j \in \{1,4,5,6,7,8,9\}\Big\}.$$

For notational convenience, we define $W(a) = M(a) + M(a+1) + 1$. By Lemma 4 in Section 2,

$$\mathrm{cl}\big(\mathcal{F}_{a,M(a)}\big) \cap \mathrm{cl}\big(\mathcal{F}_{a+1,M(a+1)}\big) \subseteq \mathrm{cl}\big(\mathcal{F}_{a,M(a),W(a)}\big).$$

From this and the result of Step 2, we have that for any fixed $a > 0$, with probability one, $\hat{S}_n \subseteq \mathrm{cl}\big(\mathcal{F}_{a,M(a),W(a)}\big)$ for sufficiently large n.

Step 4: Glivenko–Cantelli Classes.

For, $a > 0$, $M_1 > 0$, $M_2 > 0$ and $j = 1,\ldots,d$, let

$$\check{\mathcal{F}}_{a,M_1,M_2} = \Big\{\check{f} : \mathbb{R}^d \to \mathbb{R} \,\Big|\, \check{f}(\mathbf{x}) = f(\mathbf{x})1_{\{\mathbf{x}\in[-a,a]^d\}}, f \in \mathcal{F}_{a,M_1,M_2}\Big\}$$

and

$$\left(\check{\mathcal{F}}_{a,M_1,M_2}\right)_j = \Big\{\check{f} : \mathbb{R}^d \to \mathbb{R} \Big| \check{f}(\mathbf{x}) = f_j(x_j)\mathbb{1}_{\{\mathbf{x}\in[-a,a]^d\}} \quad \text{for some } f \overset{\mathcal{F}}{\sim} (f_1,\ldots,f_d,c) \in \mathcal{F}_{a,M_1,M_2}\Big\}.$$

We first claim that each $(\check{\mathcal{F}}_{a,M_1,M_2})_j$ is a $P_{\mathbf{X}}$-Glivenko–Cantelli class, where $P_{\mathbf{X}}$ is the distribution of $\mathbf{X}$. To see this, note that by Theorem 2.7.5 of van der Vaart and Wellner (1996), there exists a universal constant $C > 0$ and functions $g_k^L, g_k^U : \mathbb{R} \to [0,1]$ for $k = 1,\ldots,N_1$ with $N_1 = e^{2M_1C/\epsilon}$ such that $\mathbb{E}|g_k^U(X_{1j}) - g_k^L(X_{1j})| \le \epsilon/(2M_1)$ and such that for every monotone function $g : \mathbb{R} \to [0,1]$, we can find $k^* \in \{1,\ldots,N_1\}$ with $g_{k^*}^L \le g \le g_{k^*}^U$. By Corollary 2.7.10, the same property holds for convex or concave functions from $[-a,a]$ to $[0,1]$, provided we use N_2 brackets, where $N_2 = \exp\big\{C\big(1+\frac{M_2}{2M_1}\big)^{1/2}(2M_1/\epsilon)^{1/2}\big\}$. It follows that if j corresponds to a monotone component, then the class of functions

$$\tilde{g}_k^L(x) = 2M_1(g_k^L(x_j) - 1/2)\mathbb{1}_{\{\mathbf{x}\in[-a,a]^d\}}, \quad \tilde{g}_k^U(x) = 2M_1(g_k^U(x_j) - 1/2)\mathbb{1}_{\{\mathbf{x}\in[-a,a]^d\}},$$

for $k = 1,\ldots,N_1$, forms an ϵ-bracketing set for $\left(\check{\mathcal{F}}_{a,M_1,M_2}\right)_j$ in the $L_1(P_{\mathbf{X}})$-norm. Similarly, if j corresponds to a convex or concave component, we can define in the same way an ϵ-bracketing set for $\left(\check{\mathcal{F}}_{a,M_1,M_2}\right)_j$ of cardinality N_2 for $\left(\check{\mathcal{F}}_{a,M_1,M_2}\right)_j$. We deduce by Theorem 2.4.1 of van der Vaart and Wellner (1996) that each $\left(\check{\mathcal{F}}_{a,M_1,M_2}\right)_j$ is a P_X-Glivenko–Cantelli class. But then

$$\sup_{\check{f}\in\check{\mathcal{F}}_{a,M_1,M_2}} \left|\frac{1}{n}\sum_{i=1}^n \check{f}(\mathbf{X}_i) - \mathbb{E}\check{f}(\mathbf{X})\right| \le \sum_{j=1}^d \sup_{\check{f}_j\in\left(\check{\mathcal{F}}_{a,M_1,M_2}\right)_j} \left|\frac{1}{n}\sum_{i=1}^n \check{f}_j(X_{ij}) - \mathbb{E}\check{f}_j(X_{1j})\right|,$$

so $\check{\mathcal{F}}_{a,M_1,M_2}$ is $P_{\mathbf{X}}$-Glivenko–Cantelli. We now use this fact to show that the class of functions

$$\mathcal{H}_{a,M_1,M_2} = \Big\{h_f : \mathbb{R}^d\times\mathbb{R} \to \mathbb{R} \mid h_f(\mathbf{x},y) = \big\{yf(\mathbf{x}) - B(f(\mathbf{x}))\big\}\mathbb{1}_{\{\mathbf{x}\in[-a,a]^d\}}, f \in \mathcal{F}_{a,M_1,M_2}\Big\}$$

is P-Glivenko–Cantelli, where P is the distribution of $(\mathbf{X}, Y)$. Define $f^*, f^{**} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ by $f^*(\mathbf{x},y) = y$ and $f^{**}(\mathbf{x},y) = \mathbb{1}_{\{\mathbf{x}\in[-a,a]^d\}}$. Let

$$\mathcal{F}_1 = \big\{f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \mid f(\mathbf{x},y) = \check{f}(\mathbf{x}), \check{f} \in \check{\mathcal{F}}_{a,M_1,M_2}\big\},$$

let $\mathcal{F}_2 = \{f^*\}$ and let $\mathcal{F}_3 = \{f^{**}\}$; finally define $\psi : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by $\psi(u, v, w) = \{vu - B(u)\}w$. Then $\mathcal{H} = \psi(\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3)$, where

$$\psi(\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3) = \{\psi(f_1(\mathbf{x}, y), f_2(\mathbf{x}, y), f_3(\mathbf{x}, y)) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2, f_3 \in \mathcal{F}_3\}.$$

Now $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$ are P-Glivenko–Cantelli, ψ is continuous and (recalling that $|Y| \leq 1$ in the Binomial setting),

$$\sup_{f_1 \in \mathcal{F}_1} \sup_{f_2 \in \mathcal{F}_2} \sup_{f_3 \in \mathcal{F}_3} |\psi(f_1(\mathbf{x}, y), f_2(\mathbf{x}, y), f_3(\mathbf{x}, y))| \leq M_1(d+1) + B(M_1(d+1)),$$

which is P-integrable. We deduce from Theorem 3 of van der Vaart and Wellner (2000) that $\mathcal{H}_{a,M_1,M_2}$ is P-Glivenko–Cantelli.

Step 5: Almost sure convergence of $\hat{f}_n$. For $\epsilon > 0$, let

$$B_\epsilon(f_0) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,\middle|\, \sup_{\mathbf{x} \in [-a_0, a_0]^d} |f(\mathbf{x}) - f_0(\mathbf{x})| \leq \epsilon \right\},$$

where we suppress the dependence of $B_\epsilon(f_0)$ on a_0 in the notation. Our aim to show that with probability 1, we have $\hat{S}_n \cap \mathrm{cl}\big(\mathcal{F} \setminus B_\epsilon(f_0)\big) = \emptyset$ for sufficiently large n. In Lemma 5 in Section 2, it is established that for any $\epsilon > 0$,

$$\begin{aligned}\zeta(a^*) := \mathbb{E}\big[\{Y f_0(\mathbf{X}) - B(f_0(\mathbf{X}))\}\mathbb{1}_{\{\mathbf{X} \in [-a^*, a^*]^d\}}\big] \\ - \sup_{f \in \mathcal{F}_{a_0, M(a_0), W(a_0)} \setminus B_\epsilon(f_0)} \mathbb{E}\big[\{Y f(\mathbf{X}) - B(f(\mathbf{X}))\}\mathbb{1}_{\{\mathbf{X} \in [-a^*, a^*]^d\}}\big]\end{aligned} \tag{4}$$

is positive and a non-decreasing function of $a^* > a_0 + 1$. Since we also have that (in the Binomial setting), $-\log 2 \leq g^{-1}(t)t - B(t) \leq 0$, we can therefore choose $a^* > a_0 + 1$ such that

$$\big|\mathbb{E}\big[\{Y f_0(\mathbf{X}) - B(f_0(\mathbf{X}))\}\mathbb{1}_{\{\mathbf{X} \notin [-a^*, a^*]^d\}}\big]\big| \leq \zeta(a^*)/3. \tag{5}$$

Let

$$\begin{aligned}\mathcal{F}^* = \mathrm{cl}\big(\mathcal{F}_{a_0, M(a_0)} \setminus B_\epsilon(f_0)\big) \,\cap\, \mathrm{cl}\big(\mathcal{F}_{a_0+1, M(a_0+1)} \setminus B_\epsilon(f_0)\big) \\ \cap\, \mathrm{cl}\big(\mathcal{F}_{a^*, M(a^*)} \setminus B_\epsilon(f_0)\big) \,\cap\, \mathrm{cl}\big(\mathcal{F}_{a^*+1, M(a^*+1)} \setminus B_\epsilon(f_0)\big).\end{aligned}$$

Observe that by the result of Step 2, we have that with probability one, $\hat{S}_n \subseteq \mathcal{F}^* \cup \mathrm{cl}(B_\epsilon(f_0))$ for sufficiently large n. By Lemma 6 in Section 2,

$$\left\{ \sup_{f \in \mathcal{F}^*} \frac{1}{n} \sum_{i=1}^n \{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\} \geq \bar{L}_0 - \zeta(a^*)/3 \right\}$$

$$\subseteq \left\{ \sup_{f \in (\mathcal{F}_{a_0, M(a_0), W(a_0)} \cap \mathcal{F}_{a^*, M(a^*)+1, W(a^*)+1}) \setminus B_\epsilon(f_0)} \frac{1}{n} \sum_{i=1}^n \{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\} \mathbb{1}_{\{\mathbf{X}_i \in [-a^*, a^*]^d\}} \right. \tag{6}$$

$$\left. + \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\} \mathbb{1}_{\{\mathbf{X}_i \notin [-a^*, a^*]^d\}} \geq \bar{L}_0 - \zeta(a^*)/3 \right\},$$

Here the closure operator in (6) can be dropped by the same argument as in Step 2. Now note that

$$\Big\{ h_f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \ \Big|\ h_f(\mathbf{x}, y) = \{y f(\mathbf{x}) - B(f(\mathbf{x}))\} \mathbb{1}_{\{\mathbf{x} \in [-a^*, a^*]^d\}},$$
$$f \in (\mathcal{F}_{a_0, M(a_0), W(a_0)} \cap \mathcal{F}_{a^*, M(a^*)+1, W(a^*)+1}) \setminus B_\epsilon(f_0) \Big\} \subseteq \mathcal{H}_{a^*, M(a^*)+1, W(a^*)+1},$$

so the class is P-Glivenko–Cantelli, by the result of Step 4. We therefore have that with probability one,

$$\limsup_{n \to \infty} \sup_{f \in (\mathcal{F}_{a_0, M(a_0), W(a_0)} \cap \mathcal{F}_{a^*, M(a^*)+1, W(a^*)+1}) \setminus B_\epsilon(f_0)} \frac{1}{n} \sum_{i=1}^n \{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\} \mathbb{1}_{\{\mathbf{X}_i \in [-a^*, a^*]^d\}}$$
$$= \sup_{f \in (\mathcal{F}_{a_0, M(a_0), W(a_0)} \cap \mathcal{F}_{a^*, M(a^*)+1, W(a^*)+1}) \setminus B_\epsilon(f_0)} \mathbb{E}\big[\{Y f(\mathbf{X}) - B(f(\mathbf{X}))\} \mathbb{1}_{\{\mathbf{X} \in [-a^*, a^*]^d\}}\big]$$
$$\leq \mathbb{E}\big[\{Y f_0(\mathbf{X}) - B(f_0(\mathbf{X}))\} \mathbb{1}_{\{\mathbf{X} \in [-a^*, a^*]^d\}}\big] - \zeta(a^*) \tag{7}$$
$$\leq \bar{L}_0 - 2\zeta(a^*)/3, \tag{8}$$

where (7) is due to (4), and where (8) is due to (5). In addition, under the Binomial setting, for every $n \in \mathbb{N}$,

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \{Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i))\} \mathbb{1}_{\{\mathbf{X}_i \notin [-a^*, a^*]^d\}}$$
$$\leq \frac{1}{n} \sum_{i=1}^n \sup_{t \in \mathbb{R}} \{Y_i t - B(t)\} \mathbb{1}_{\{\mathbf{X}_i \notin [-a^*, a^*]^d\}} \leq 0. \tag{9}$$

We deduce from (6), (8) and (9) that with probability one, $\hat{S}_n \subseteq \mathrm{cl}\big(B_\epsilon(f_0)\big)$ for sufficiently large n. Finally, since $\mathrm{cl}\big(B_\epsilon(f_0)\big)|_{[-a_0,a_0]^d} = B_\epsilon(f_0)|_{[-a_0,a_0]^d}$, the conclusion of Theorem 1 for Binomial models follows.

Consistency of other EF additive models. The proof for other EF models follows the same structure, but involves some changes in certain places. We list the modifications required for each step here:

- In Step 1, we add a term independent of f to the definition of the partial log-likelihood:

$$\tilde{\ell}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\Big[Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i)) - \sup_{t\in\mathrm{dom}(B)}\{Y_i t - B(t)\}\Big].$$

  Note that

$$\sup_{t\in\mathrm{dom}(B)}\{Y_i t - B(t)\} = \begin{cases} Y_i^2/2 & \text{if EF is Gaussian;} \\ Y_i \log Y_i - Y_i & \text{if EF is Poisson;} \\ -1 - \log Y_i & \text{if EF is Gamma.} \end{cases}$$

  This allows us to prove that $\mathbb{E}\{\tilde{\ell}_n(f_0)\} \in (-\infty, 0]$ in all cases: in particular, in the Gaussian case, $\mathbb{E}\{\tilde{\ell}_n(f_0)\} = -\phi_0/2$; for the Poisson, we can use Lemma 7 in Section 2 to see that $\mathbb{E}\{\tilde{\ell}_n(f_0)\} \in [-1, 0]$; for the Gamma, this claim follows from Lemma 8 in Section 2. It then follows from the strong law of large numbers that almost surely

$$\liminf_{n\to\infty} \sup_{f\in\mathrm{cl}(\mathcal{F})} \tilde{\ell}_n(f) \geq \mathbb{E}\{\tilde{\ell}_n(f_0)\} =: \tilde{L}_0.$$

- In Step 2, the deterministic constant $M = M(a) \in (0,\infty)$ needs to be chosen differently for different EF distributions. Let $\mathcal{C}_a = \{C_1, \ldots, C_N\}$ be the same finite collection of compact subsets defined previously. We then can pick

$$M = \begin{cases} 4\Big(\sqrt{\frac{2(-\tilde{L}_0+1)}{\min_{1\le k\le N}\mathbb{P}(\mathbf{X}\in C_k, |Y|\le 1)}} + 1\Big) & \text{if EF is Gaussian;} \\ 4\Big(\frac{-\tilde{L}_0+1}{\min_{1\le k\le N}\mathbb{P}(\mathbf{X}\in C_k, Y=1)} + 1\Big) & \text{if EF is Poisson;} \\ 4\Big(\frac{2(-\tilde{L}_0+1)}{\min_{1\le k\le N}\mathbb{P}(\mathbf{X}\in C_k, 1\le Y\le e)} + 4\Big) & \text{if EF is Gamma.} \end{cases}$$

- Step 3 is exactly the same for all the EF distributions listed in Table 2 in the main text.

- In Step 4, we define the class of functions

$$\tilde{\mathcal{H}}_{a,M_1,M_2} = \Big\{h_f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \mid$$
$$h_f(\mathbf{x}, y) = \Big[yf(\mathbf{x}) - B(f(\mathbf{x})) - \sup_{t\in\mathrm{dom}(B)} \{yt - B(t)\}\Big] 1_{\{\mathbf{x}\in[-a,a]^d\}}, f \in \mathcal{F}_{a,M_1,M_2}\Big\}.$$

  In the Gaussian case, we can rewrite $h_f(\mathbf{x}, y) = -\frac{1}{2}\{y - f(\mathbf{x})\}^2 1_{\{\mathbf{x}\in[-a,a]^d\}}$. By taking the P-integrable envelope function to be

$$F(\mathbf{x}, y) = \frac{1}{2}\{|y| + M_1(d + 1)\}^2 1_{\{\mathbf{x}\in[-a,a]^d\}} \geq \sup_{f\in\mathcal{F}_{a,M_1,M_2}} |h_f(\mathbf{x}, y)|,$$

  we can again deduce from Theorem 3 of van der Vaart and Wellner (2000) that $\tilde{\mathcal{H}}_{a,M_1,M_2}$ is P-Glivenko–Cantelli. Similarly, in the Poisson case, we can show that $\tilde{\mathcal{H}}_{a,M_1,M_2}$ is P-Glivenko–Cantelli by taking the envelope function to be $F(\mathbf{x}, y) = \big\{yM_1(d + 1) + e^{M_1(d+1)} + y + y \log y\big\} 1_{\{\mathbf{x}\in[-a,a]^d\}}$.

  The Gamma case is slightly more complex, mainly due to the fact that $\mathrm{dom}(B) \neq \mathbb{R}$. For $\delta > 0$, let

$$\tilde{\mathcal{H}}^{\delta}_{a,M_1,M_2} = \Big\{h_f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \mid$$
$$h_f(\mathbf{x}, y) = \big\{yf(\mathbf{x}) + \log\big(\max(-f(\mathbf{x}), \delta)\big) - 1 + \log y\big\} 1_{\{\mathbf{x}\in[-a,a]^d\}}, f \in \mathcal{F}_{a,M_1,M_2}\Big\}.$$

  Again, we can show that $\tilde{\mathcal{H}}^{\delta}_{a,M_1,M_2}$ is P-Glivenko–Cantelli by taking the envelope function for $\tilde{\mathcal{H}}^{\delta}_{a,M_1,M_2}$ to be $F(\mathbf{x}, y) = \big\{yM_1(d+1) + |\log\delta| + |\log(M_1(d+1))| + 1 + \log y\big\} 1_{\{\mathbf{x}\in[-a,a]^d\}}$.

- Step 5 for the Gaussian and Poisson settings are essentially a replication of that for the Binomial case. Only very minor changes are required:

  (a) where applicable, add the term $-\sup_{t\in\mathbb{R}}[Yt-B(t)]$ to $\{Yf_0(\mathbf{X})-B(f_0(\mathbf{X}))\}$ and $\{Yf(\mathbf{X}) - B(f(\mathbf{X}))\}$; make the respective change to $\{Y_if(\mathbf{X}_i) - B(f(\mathbf{X}_i))\}$ and $\{yf(\mathbf{x}) - B(f(\mathbf{x}))\}$;

(b) change $\bar{L}_0$ to $\tilde{L}_0$;

(c) change $\mathcal{H}_{a^*,M(a^*)+1,W(a^*)+1}$ to $\tilde{\mathcal{H}}_{a^*,M(a^*)+1,W(a^*)+1}$;

(d) rewrite (9) as

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \Big[Y_i f(\mathbf{X}_i) - B(f(\mathbf{X}_i)) - \sup_{t\in\mathbb{R}}\{Y_i t - B(t)\}\Big] \mathbb{1}_{\{\mathbf{X}_i \notin [-a^*,a^*]^d\}} \leq 0.$$

The analogue of Step 5 for the Gamma distribution is a little more involved. Set $\delta_0 = \inf_{\mathbf{x}\in[-a_0-1,a_0+1]^d} -f_0(\mathbf{x})/e^2 > 0$. Note that the above supremum is attained as f_0 is a continuous function. Then one can prove in a similar fashion to Lemma 5 that

$$\zeta = \mathbb{E}\big[\big\{Y f_0(\mathbf{X}) + \log(-f_0(\mathbf{X})) + 1 + \log Y\big\}\mathbb{1}_{\{\mathbf{X}\in[-a_0-1,a_0+1]^d\}}\big]$$
$$- \sup_{f\in\mathcal{F}_{a_0,M(a_0),W(a_0)}\setminus B_\epsilon(f_0)} \mathbb{E}\big[\big\{Y f(\mathbf{X}) + \log(\max(-f(\mathbf{X}),\delta_0)) + 1 + \log Y\big\}\mathbb{1}_{\{\mathbf{X}\in[-a_0-1,a_0+1]^d\}}\big] > 0.$$

Next we pick $a^* > a_0 + 1$ such that

$$\Big|\mathbb{E}\big[\big\{Y f_0(\mathbf{X}) + \log(-f_0(\mathbf{X})) + 1 + \log Y\big\}\mathbb{1}_{\{\mathbf{X}\notin[-a^*,a^*]^d\}}\big]\Big| \leq \zeta/3$$

and $\delta^* = \inf_{\mathbf{x}\in[-a^*-1,a^*+1]^d} -f_0(\mathbf{x})/e^2$. Write

$$\mathcal{F}^{**} = \Big(\mathcal{F}_{a_0,M(a_0),W(a_0)} \cap \mathcal{F}_{a^*,M(a^*)+1,W(a^*)+1}\Big)\setminus B_\epsilon(f_0).$$

With $\mathcal{F}^*$ defined as in Step 5, we have

$$\Big\{\sup_{f\in\mathcal{F}^*}\frac{1}{n}\sum_{i=1}^n \{Y_i f(\mathbf{X}_i) + \log(-f(\mathbf{X}_i)) + 1 + \log Y_i\} \geq \tilde{L}_0 - \zeta/3\Big\}$$
$$\subseteq \Bigg\{\sup_{f\in\mathcal{F}^{**}}\frac{1}{n}\sum_{i=1}^n \big\{Y_i f(\mathbf{X}_i) + \log(\max(-f(\mathbf{X}_i),\delta_0)) + 1 + \log Y_i\big\}\mathbb{1}_{\{\mathbf{X}_i\in[-a_0-1,a_0+1]^d\}}$$
$$+ \sup_{f\in\mathcal{F}^{**}}\frac{1}{n}\sum_{i=1}^n \big\{Y_i f(\mathbf{X}_i) + \log(\max(-f(\mathbf{X}_i),\delta^*)) + 1 + \log Y_i\big\}\mathbb{1}_{\{\mathbf{X}_i\in[-a^*,a^*]^d\setminus[-a_0-1,a_0+1]^d\}}$$
$$+ \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \big\{Y_i f(\mathbf{X}_i) + \log(-f(\mathbf{X}_i)) + 1 + \log Y_i\big\}\mathbb{1}_{\{\mathbf{X}_i\notin[-a^*,a^*]^d\}} \geq \tilde{L}_0 - \zeta/3\Bigg\}.$$

Again we apply Glivenko–Cantelli theorem to finish the proof, where we also use the fact that

$$\sup_{f\in\mathcal{F}^{**}} \mathbb{E}\big[\big\{Yf(\mathbf{X}) + \log(\max(-f(\mathbf{X}),\delta^*)) + 1 + \log Y\big\}1_{\{\mathbf{X}\in[-a^*,a^*]^d\setminus[-a_0-1,a_0+1]^d\}}\big]$$
$$\leq \mathbb{E}\big[\big\{Yf_0(\mathbf{X}) + \log(-f_0(\mathbf{X})) + 1 + \log Y\big\}1_{\{\mathbf{X}\in[-a^*,a^*]^d\setminus[-a_0-1,a_0+1]^d\}}\big].$$

We now indicate how to extend the proof of Theorem 1 to cover the SCLSE defined in Section 7. Here we assume assumptions **1**, **2** and **4**, but weaken **3** by only requiring that

$$Y_i = f_0(\mathbf{X}_i) + \epsilon_i,$$

for $i = 1,\ldots,n$, where $\mathbb{E}(\epsilon_i|\mathbf{X}_i) = 0$, $\mathrm{Var}(\epsilon_i|\mathbf{X}_i) = \phi_0$ and $f_0 \in \mathcal{F}$. The proof is almost identical to that for the Gaussian case, with the only minor complication being that we need to redefine M in Step 2. Specifically, we can set

$$M = 4\Big(\sqrt{\frac{2(-\tilde{L}_0+1)}{\min_{1\leq k\leq N}\mathbb{P}(\mathbf{X}\in C_k, |Y|\leq M^*)}} + M^*\Big),$$

where

$$M^* = 2^{1/2}\Big(\sup_{\mathbf{x}\in[-2a,2a]^d} f_0(\mathbf{x})^2 + \phi_0 + 1\Big)^{1/2}.$$

This choice of M ensures that $\mathbb{P}(\mathbf{X}\in C_k, |Y|\leq M^*) > 0$ for each k, because

$$(M^*)^2\mathbb{P}(\mathbf{X}\in C_k, |Y|\leq M^*) \geq (M^*)^2\mathbb{P}(\mathbf{X}\in C_k) - \mathbb{E}(Y^2 1_{\{\mathbf{X}\in C_k\}})$$
$$\geq \Big\{(M^*)^2 - 2\sup_{\mathbf{x}\in[-2a,2a]^d} f_0(\mathbf{x})^2 - 2\phi_0\Big\}\mathbb{P}(\mathbf{X}\in C_k) > 0.$$

Proof of Corollary 1

By Theorem 1, we have

$$\sup_{\hat{f}_n\in\hat{S}_n}|\hat{c}_n - c_0| = \sup_{\hat{f}_n\in\hat{S}_n}|\hat{f}_n(\mathbf{0}) - f_0(\mathbf{0})|\to 0$$

almost surely, as $n\to\infty$. Moreover, writing $I_j = \{0\}\times\ldots\times\{0\}\times[-a_0,a_0]\times\{0\}\times\ldots\times\{0\}$, we have

$$\sup_{\hat{f}_n\in\hat{S}_n}\sum_{j=1}^d \sup_{x_j\in[-a_0,a_0]}|\hat{f}_{n,j}(x_j) - f_{0,j}(x_j)| = \sup_{\hat{f}_n\in\hat{S}_n}\sum_{j=1}^d \sup_{\mathbf{x}\in I_j}|\hat{f}_n(\mathbf{x}) - f_0(\mathbf{x}) - \hat{c}_n + c_0|\to 0,$$

almost surely, using Theorem 1 again and the triangle inequality.

PROOF OF PROPOSITION 2

Fix an index matrix $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m) \in \mathbb{R}^{d\times m}$. For any sequence $\mathbf{A}^1, \mathbf{A}^2, \ldots \in \mathbb{R}^{d\times m}$ with $\lim_{k\to\infty} \|\mathbf{A}^k - \mathbf{A}\|_F = 0$, where $\|\cdot\|_F$ denotes the Frobenius norm, we claim that $\lim_{k\to\infty} \|(\mathbf{A}^k)^T\mathbf{X}_i - \mathbf{A}^T\mathbf{X}_i\|_1 = 0$ for every $i = 1, \ldots, n$. To see this, we write $\mathbf{A}^k = (\boldsymbol{\alpha}_1^k, \ldots, \boldsymbol{\alpha}_m^k)$. It then follows that

$$\begin{aligned}\|(\mathbf{A}^k)^T\mathbf{X}_i - \mathbf{A}^T\mathbf{X}_i\|_1 &= \sum_{h=1}^m \Big|((\mathbf{A}^k)^T\mathbf{X}_i)_h - (\mathbf{A}^T\mathbf{X}_i)_h\Big| = \sum_{h=1}^m \Big|\sum_{j=1}^d (A_{jh}^k - A_{jh})X_{ij}\Big| \\ &\leq \sum_{h=1}^m \bigg[\|\mathbf{X}_i\|_2 \bigg\{\sum_{j=1}^d (A_{jh}^k - A_{jh})^2\bigg\}^{1/2}\bigg] \leq \|\mathbf{X}_i\|_2\sqrt{m}\|\mathbf{A}^k - \mathbf{A}\|_F \to 0\end{aligned}$$

as $k \to \infty$, where we have applied the Cauchy–Schwarz inequality twice. Now write $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{im})^T = \mathbf{A}^T\mathbf{X}_i$ for every $i = 1, \ldots, n$ and take

$$a^* = \max_{1\leq i\leq n, 1\leq j\leq m} |Z_{ij}|.$$

Since $\bigcup_{M=1}^\infty \mathcal{F}_{a^*,M} = \mathcal{F}$ (where $\mathcal{F}_{a^*,M}$ is defined in (1), although d should be replaced there with m), it follows that

$$\lim_{M\to\infty} \sup_{f\in\mathcal{F}_{a^*,M}} \bar{\ell}_n\big(f; (\mathbf{Z}_1, Y_1), \ldots, (\mathbf{Z}_n, Y_n)\big) = \sup_{f\in\mathcal{F}} \bar{\ell}_n\big(f; (\mathbf{Z}_1, Y_1), \ldots, (\mathbf{Z}_n, Y_n)\big) = \Lambda_n(\mathbf{A}).$$

Therefore, for any $\epsilon > 0$, there exist $M_\epsilon > 0$ and $f^* \overset{\mathcal{F}}{\sim} (f_1^*, \ldots, f_m^*, c^*) \in \mathcal{F}_{a^*,M_\epsilon}$ such that

$$\bar{\ell}_n\Big(f^*; (\mathbf{Z}_1, Y_1), \ldots, (\mathbf{Z}_n, Y_n)\Big) \geq \Lambda_n(\mathbf{A}) - \epsilon.$$

We can then find piecewise linear and continuous functions $f_1^{**}, \ldots, f_m^{**}$ such that $f_j^{**}(Z_{ij}) = f_j^*(Z_{ij})$ for every $i = 1, \ldots, n$, $j = 1, \ldots, m$. Consequently, the additive function $f^{**}(\mathbf{z}) = \sum_{j=1}^m f_j^{**}(z_j) + c^*$ is continuous. It now follows that

$$\begin{aligned}\liminf_{k\to\infty} \Lambda_n(\mathbf{A}^k) &\geq \liminf_{k\to\infty} \bar{\ell}_n\big(f^{**}; ((\mathbf{A}^k)^T\mathbf{X}_1, Y_1), \ldots, ((\mathbf{A}^k)^T\mathbf{X}_n, Y_n)\big) \\ &= \bar{\ell}_n\big(f^{**}; (\mathbf{Z}_1, Y_1), \ldots, (\mathbf{Z}_n, Y_n)\big) \\ &= \bar{\ell}_n\big(f^*; (\mathbf{Z}_1, Y_1), \ldots, (\mathbf{Z}_n, Y_n)\big) \geq \Lambda_n(\mathbf{A}) - \epsilon.\end{aligned}$$

Since both $\epsilon > 0$ and the sequence $(\mathbf{A}^k)$ were arbitrary, the result follows.

PROOF OF THEOREM 3

The structure of the proof is essentially the same as that of Theorem 1. For the sake of brevity, we focus on the main changes and on the Gaussian setting. Following the strategy used in the proof of Theorem 1, we work with the logarithm of a normalised likelihood here:

$$\begin{aligned}\tilde{\ell}_n(f;\mathbf{A}) &\equiv \tilde{\ell}_n\big(f;(\mathbf{A}^T\mathbf{X}_1,Y_1),\ldots,(\mathbf{A}^T\mathbf{X}_n,Y_n)\big)\\ &= \bar{\ell}_n\big(f;(\mathbf{A}^T\mathbf{X}_1,Y_1),\ldots,(\mathbf{A}^T\mathbf{X}_n,Y_n)\big) - \frac{1}{n}\sum_{i=1}^n \sup_{t\in \mathrm{dom}(B)}\{Y_i t - B(t)\}\\ &= -\frac{1}{2n}\sum_{i=1}^n \big\{f(\mathbf{A}^T\mathbf{X}_i) - Y_i\big\}^2.\end{aligned}$$

So in Step 1, we can establish that $\mathbb{E}\tilde{\ell}_n(f_0;\mathbf{A}_0) = -\phi_0/2$.

In Step 2, we aim to bound $\tilde{f}_n^I$ on $[-a,a]^d$ for any fixed $a > 0$. Three cases are considered:

(a) If $m \geq 2$ and $\mathbf{L}_m \in \mathcal{L}_m$, then $\tilde{f}_n^I$ is either convex or concave. One can now use the convexity/concavity to show that $\limsup_{n\to\infty}\sup_{\mathbf{x}\in[-a,a]^d}|\tilde{f}_n^I(\mathbf{x})| < M(a)$ almost surely for some deterministic constant $M(a) < \infty$ that only depends on a. See, for instance, Proposition 4 of Lim and Glynn (2012) for a similar argument.

(b) Otherwise, if $\mathbf{L}_m \notin \mathcal{L}_m$, we will show that there exists deterministic $M(a) \in (0,\infty)$ such that with probability one,

$$\tilde{S}_n \subseteq \mathcal{G}^{\delta}_{a,M(a)} \tag{10}$$

for sufficiently large n, where we define

$$\mathcal{G}^{\delta}_{a,M} = \Big\{f^I : \mathbb{R}^d \to \mathbb{R} \;\Big|\; f^I(\mathbf{x}) = f(\mathbf{A}^T\mathbf{x}),\ \text{with } f \in \mathcal{F}_{a,M} \text{ and } \mathbf{A} \in \mathcal{A}^{\delta}\Big\}.$$

To see this, we first extend Lemma 1 to the following:

LEMMA 2. *Fix $a > 0$ and $\delta > 0$, and set $\tilde{\delta} = \min(\delta, d^{-1})$. For every $f^I(\mathbf{x}) = f(\mathbf{A}^T\mathbf{x}) = \sum_{j=1}^m f_j(\boldsymbol{\alpha}_j^T\mathbf{x}) + c$ with $f \overset{\mathcal{F}}{\sim} (f_1, \ldots, f_m, c)$ and $\mathbf{A} \in \mathcal{A}^\delta$, there exists a convex, compact subset D_{f^I} of $[-2\tilde{\delta}^{-1/2}ad, 2\tilde{\delta}^{-1/2}ad]^d$ having Lebesgue measure $\left(\frac{a}{2d}\right)^d$ such that*

$$\max\Big\{\inf_{\mathbf{x}\in D_{f^I}} f^I(\mathbf{x}), \inf_{\mathbf{x}\in D_{f^I}} -f^I(\mathbf{x})\Big\} \geq \frac{1}{4}\max\Big\{\sup_{|z_1|\leq a}|f_1(z_1)|, \ldots, \sup_{|z_m|\leq a}|f_m(z_m)|, 2|c|\Big\}. \tag{11}$$

PROOF. First consider the case $m = d$. Note that every $\mathbf{A} \in \mathcal{A}^\delta$ is invertible. In fact, if λ is an eigenvalue of $\mathbf{A}$, then $\delta^{1/2} \leq |\lambda| \leq 1$, where the upper bound follows from the Gerschgorin circle theorem (Gerschgorin, 1931; Gradshteyn and Ryzhik, 2007). Let $C_1, \ldots, C_N$ be the sets constructed for f in Lemma 1. Then, writing ν_d for Lebesgue measure on $\mathbb{R}^d$,

$$\min_{1\leq k\leq N} \nu_d\big((\mathbf{A}^T)^{-1}C_k\big) \geq \frac{1}{|\det(\mathbf{A}^T)|}\min_{1\leq k\leq N}\nu_d(C_k) \geq \left(\frac{a}{2d}\right)^d,$$

and

$$\bigcup_{1\leq k\leq N}(\mathbf{A}^T)^{-1}C_k \subseteq (\mathbf{A}^T)^{-1}[-2a, 2a]^d \subseteq [-2\tilde{\delta}^{-1/2}ad, 2\tilde{\delta}^{-1/2}ad]^d.$$

Thus (11) is satisfied. To complete the proof of this lemma, we note that for any $m < d$, we can always find a $d \times (d-m)$ matrix $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{d-m})$ such that

(i) $\|\boldsymbol{\beta}_j\|_1 = 1$ for every $j = 1, \ldots, d-m$.
(ii) $\boldsymbol{\beta}_j^T\boldsymbol{\beta}_k = 0$ for every $1 \leq j < k \leq d-m$.
(iii) $\mathbf{A}^T\mathbf{B} = 0$.

Let $\mathbf{A}_+ = (\mathbf{A}, \mathbf{B})$, so the modulus of every eigenvalue of $\mathbf{A}_+$ belongs to the interval $[\min(\delta^{1/2}, d^{-1/2}), 1]$. Since $f^I(\mathbf{x}) = f(\mathbf{A}^T\mathbf{x}) \equiv f'(\mathbf{A}_+^T\mathbf{x})$ with $f'(\mathbf{z}) = \sum_{j=1}^m f_j(z_j) + c$ for every $\mathbf{z} = (z_1, \ldots, z_d)^T \in \mathbb{R}^d$, the problem reduces to the case $m = d$.

Then, instead of using the strong law of large numbers to complete this step, we apply the Glivenko–Cantelli theorem for classes of convex sets (Bhattacharya and Rao, 1976, Theorem 1.11). This change is necessary to circumvent the fact that the set D_{f^I} depends on the function f^I (via its index matrix $\mathbf{A}$).

(c) Finally, if $m = 1$, then the Cauchy–Schwarz inequality gives that $\mathcal{A} \equiv \mathcal{A}^\delta$ with $\delta = d^{-1}$. Thus (10) still holds true.

Two different cases are considered in Step 4:

(a) If $m \geq 2$ and $\mathbf{L}_m \in \mathcal{L}_m$, then without loss of generality, we can assume $\mathbf{L}_m \in \{1,4,5,6\}^m$. It is enough to show that the set of functions

$$\mathcal{G}_{a,M_1,M_2} = \Big\{ h_f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \,\Big|\, h_f(\mathbf{x}, y) = -\frac{1}{2}\{f(\mathbf{x}) - y\}^2 1_{\{\mathbf{x}\in[-a,a]^d\}} \\ \text{with } f : \mathbb{R}^d \to \mathbb{R} \text{ convex}, \sup_{\mathbf{x}\in[-a,a]^d} |f(\mathbf{x})| \leq M_1, \\ \text{and } |f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq M_2 \|\mathbf{x}_1 - \mathbf{x}_2\| \text{ for any } \mathbf{x}_1, \mathbf{x}_2 \in [-a,a]^d \Big\}$$

is P-Glivenko–Cantelli, where P is the distribution of $(\mathbf{X}, Y)$. This follows from an application of Corollary 2.7.10 and Theorem 2.4.1 of van der Vaart and Wellner (1996), as well as Theorem 3 of van der Vaart and Wellner (2000).

(b) Otherwise, we need to show that the set of functions

$$\mathcal{G}^\delta_{a,M_1,M_2} = \Big\{ h_{f,\mathbf{A}} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \,\Big|\, h_{f,\mathbf{A}}(\mathbf{x}, y) = -\frac{1}{2}\{f(\mathbf{A}^T\mathbf{x}) - y\}^2 1_{\{\mathbf{x}\in[-a,a]^d\}} \\ \text{with } f \in \mathcal{F}_{a,M_1,M_2} \text{ and } \mathbf{A} \in \mathcal{A}^\delta \Big\}$$

is P-Glivenko–Cantelli. The proof is similar to that given in Step 4 of the proof of Theorem 1. The compactness of $\mathcal{A}^\delta$, together with a bracketing number argument is used here to establish the claim. See Lemma 9 in Section 2 for details.

Proof of Corollary 2

This result follows from Theorem 1 of Yuan (2011) and our Theorem 3. See also Theorem 5 of Samworth and Yuan (2012) for a similar type of argument.

2. Proofs: auxiliary results

Recall the definition of Θ from the proof of Proposition 1.

LEMMA 3. *The set Θ is a closed subset of $\bar{\mathbb{R}}^n$.*

PROOF. Suppose that, for each $m \in \mathbb{N}$, the vector $\boldsymbol{\eta}^m = (\eta_1^m, \ldots, \eta_n^m)^T$ belongs to Θ, and that $\boldsymbol{\eta}^m \to \boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ as $m \to \infty$. Then, for each $m \in \mathbb{N}$, there exists a sequence $(f^{m,k}) \in \mathcal{F}$ such that $f^{m,k}(\mathbf{X}_i) \to \eta_i^m$ as $k \to \infty$ for $i = 1, \ldots, n$. It follows that we can find $k_m \in \mathbb{N}$ such that $f^{m,k_m}(\mathbf{X}_i) \to \eta_i$ as $m \to \infty$, for each $i = 1, \ldots, n$.

For $j = 1, \ldots, d$, let $\{X_{(i),j}\}_{i=1}^{N_j}$ denote the distinct order statistics of $\{X_{ij}\}_{i=1}^n$ (thus $N_j < n$ if there are ties among $\{X_{ij}\}_{i=1}^n$). Moreover, let

$$\mathcal{V}_j = \{(-\infty, X_{(1),j}], [X_{(1),j}, X_{(2),j}], \ldots, [X_{(N_j-1),j}, X_{(N_j),j}], [X_{(N_j),j}, \infty)\},$$

and let $\mathcal{V} = \times_{j=1}^d \mathcal{V}_j$. Thus $|\mathcal{V}| = \prod_{j=1}^d (N_j + 1)$ and the union of all the sets in $\mathcal{V}$ is $\mathbb{R}^d$. Writing $f^{m,k_m} \overset{\mathcal{F}}{\sim} (f_1^{m,k_m}, \ldots, f_d^{m,k_m}, c^{m,k_m})$, we define a modified sequence $\tilde{f}^m \overset{\mathcal{F}}{\sim} (\tilde{f}_1^m, \ldots, \tilde{f}_d^m, \tilde{c}^m)$ at $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ by setting

$$\tilde{f}_j^m(x_j) = \begin{cases} \frac{(X_{(i+1),j} - x_j) f_j^{m,k_m}(X_{(i),j})}{X_{(i+1),j} - X_{(i),j}} + \frac{(x_j - X_{(i),j}) f_j^{m,k_m}(X_{(i+1),j})}{X_{(i+1),j} - X_{(i),j}} & \text{if } x_j \in [X_{(i),j}, X_{(i+1),j}] \\ \frac{(X_{(2),j} - x_j) f_j^{m,k_m}(X_{(1),j})}{X_{(2),j} - X_{(1),j}} + \frac{(x_j - X_{(1),j}) f_j^{m,k_m}(X_{(2),j})}{X_{(2),j} - X_{(1),j}} & \text{if } x_j \in (-\infty, X_{(1),j}] \\ \frac{(X_{(N_j),j} - x_j) f_j^{m,k_m}(X_{(N_j-1),j})}{X_{(N_j),j} - X_{(N_j-1),j}} + \frac{(x_j - X_{(N_j-1),j}) f_j^{m,k_m}(X_{(N_j),j})}{X_{(N_j),j} - X_{(N_j-1),j}} & \text{if } x_j \in [X_{(N_j),j}, \infty), \end{cases}$$

and $\tilde{c}^m = c^{m,k_m}$. Thus each component function $\tilde{f}_j^m$ is piecewise linear, continuous and satisfies the same shape constraint as f_j^{m,k_m}, and $\tilde{f}^m$ is piecewise affine and $\tilde{f}^m(\mathbf{X}_i) = f^{m,k_m}(\mathbf{X}_i) = \eta_i^m$ for $i = 1, \ldots, n$. The proof will therefore be concluded if we can show that a subsequence of $(\tilde{f}^m)$ converges pointwise in $\bar{\mathbb{R}}$. To do this, it suffices to show that, given an arbitrary $V \in \mathcal{V}$, we can find a subsequence of $(\tilde{f}^m|_V)$ (where $\tilde{f}^m|_V$ denotes the restriction of $\tilde{f}^m$ to V) converging pointwise in $\bar{\mathbb{R}}$. Note that we can write

$$\tilde{f}^m|_V(\mathbf{x}) = (\mathbf{a}^m)^T(\mathbf{x}^T, 1)^T$$

for some $\mathbf{a}^m = (a_1^m, \ldots, a_{d+1}^m)^T \in \mathbb{R}^{d+1}$. If the sequence $(\mathbf{a}^m)$ is bounded, then we can find a subsequence $(\mathbf{a}^{m_k})$, converging to $\mathbf{a} \in \mathbb{R}^{d+1}$, say. In that case, for all

$\mathbf{x} \in V$, we have $\tilde{f}^{m_k}|_V(\mathbf{x}) \to \mathbf{a}^T(\mathbf{x}^T, 1)^T$, and we are done. On the other hand, if $(\mathbf{a}^m)$ is unbounded, we can let $j^m = \text{argmax}_{j=1,\dots,d+1} |a_j^m|$, where we choose the largest index in the case of ties. Since j^m can only take $d+1$ values, we may assume without loss of generality that there is a subsequence (j^{m_k}) such that $j^{m_k} = d+1$ for all $k \in \mathbb{N}$ and such that $a_{d+1}^{m_k} \to \infty$ as $k \to \infty$. By choosing further subsequences if necessary, we may also assume that

$$\left(\frac{a_1^{m_k}}{a_{d+1}^{m_k}}, \dots, \frac{a_d^{m_k}}{a_{d+1}^{m_k}}\right)^T \to (\tilde{a}_1, \dots, \tilde{a}_d)^T =: \tilde{\mathbf{a}},$$

say, where $\tilde{\mathbf{a}} \in [-1, 1]^d$. Writing $V_1 = \{\mathbf{x} \in V : (\tilde{\mathbf{a}}^T, 1)(\mathbf{x}^T, 1)^T = 0\}$, $V_1^+ = \{\mathbf{x} \in V : (\tilde{\mathbf{a}}^T, 1)(\mathbf{x}^T, 1)^T > 0\}$ and $V_1^- = \{\mathbf{x} \in V : (\tilde{\mathbf{a}}^T, 1)(\mathbf{x}^T, 1)^T < 0\}$, we deduce that for large k,

$$\tilde{f}^{m_k}|_V(\mathbf{x}) = a_{d+1}^{m_k}\left(\frac{a_1^{m_k}}{a_{d+1}^{m_k}}, \dots, \frac{a_d^{m_k}}{a_{d+1}^{m_k}}, 1\right)^T (\mathbf{x}^T, 1)^T \to \begin{cases} \infty & \text{if } \mathbf{x} \in V_1^+ \\ -\infty & \text{if } \mathbf{x} \in V_1^-. \end{cases}$$

It therefore suffices to consider $\tilde{f}^{m_k}|_{V_1}$. We may assume that $\tilde{\mathbf{a}} \neq \mathbf{0}$ (otherwise $V_1 = \emptyset$ and we are done), so without loss of generality assume $\tilde{a}_d \neq 0$. But then, for $\mathbf{x} \in V_1$,

$$\tilde{f}^{m_k}|_{V_1}(\mathbf{x}) = (\mathbf{a}^{m_k})^T(\mathbf{x}^T, 1)^T = (\mathbf{b}^{m_k})^T(\mathbf{x}_{(-d)}^T, 1)^T,$$

where $\mathbf{x}_{(-d)} = (x_1, \dots, x_{d-1})^T$, and where $\mathbf{b}^{m_k} = (b_1^{m_k}, \dots, b_d^{m_k}) \in \mathbb{R}^d$, with $b_j^{m_k} = a_j^{m_k} - \frac{a_d^{m_k}}{\tilde{a}_d}\tilde{a}_j$ for $j = 1, \dots, d-1$ and $b_d^{m_k} = a_{d+1}^{m_k} - \frac{a_d^{m_k}}{\tilde{a}_d}$. Applying the same argument inductively, we find subsets $V_2, \dots, V_{d+1}$, where $V_1 \supseteq V_2 \supseteq \dots \supseteq V_{d+1}$, where V_j has dimension $d-j$ and $V_{d+1} = \emptyset$, such that a subsequence of $(\tilde{f}^{m_k})$ converges pointwise in $\bar{\mathbb{R}}$ for all $\mathbf{x} \in V \setminus V_j$.

Now recall the definitions of $\mathcal{F}_{a,M}$, $\mathcal{F}_{a,M_1,M_2}$, $M(a)$, $W(a)$ and $B_\epsilon(f_0)$ from the proof of Theorem 1.

LEMMA 4. *For any $a > 0$, we have* $\text{cl}(\mathcal{F}_{a,M(a)}) \cap \text{cl}(\mathcal{F}_{a+1,M(a+1)}) \subseteq \text{cl}(\mathcal{F}_{a,M(a),W(a)})$.

PROOF. We first consider the case $M(a) \leq M(a+1)$. Suppose $f \in \text{cl}(\mathcal{F}_{a,M(a)}) \cap \text{cl}(\mathcal{F}_{a+1,M(a+1)})$, so there exists a sequence (f^k) such that $f^k \in \mathcal{F}_{a,M(a)}$ and such that $f^k \overset{\mathcal{F}}{\sim} (f_1^k, \ldots, f_d^k, c^k)$ converges pointwise in $\bar{\mathbb{R}}$ to f. Our first claim is that there exists a subsequence (f^{k_m}) such that $f^{k_m} \in \mathcal{F}_{a+1,M(a+1)+1}$ for every $m \in \mathbb{N}$.

Indeed, suppose for a contradiction that there exists $K \in \mathbb{N}$ such that for every $k \geq K$, we have $f^k \notin \mathcal{F}_{a+1,M(a+1)+1}$. Let

$$\mathbf{b}^k = (b_1^k, \ldots, b_{2d+1}^k)^T = \Big(|f_1^k(-a-1)|, |f_1^k(a+1)|, \ldots, |f_d^k(-a-1)|, |f_d^k(a+1)|, 2|c^k|\Big)^T.$$

It follows from our hypothesis and the shape restrictions that $\max_{j=1,\ldots,2d+1} b_j^k > M(a+1)+1$ for $k \geq K$. Furthermore, we cannot have $\mathrm{argmax}_{j=1,\ldots,2d+1} b_j^k = 2d+1$ for any $k \geq K$, because $2|c^k| = 2|f^k(\mathbf{0})| \leq M(a) < M(a+1)+1$ for every $k \in \mathbb{N}$. We therefore let $j^k = \mathrm{argmax}_{j=1,\ldots,2d} b_j^k$, where we choose the largest index in the case of ties. Since j^k can only take $2d$ values, we may assume without loss of generality that there is a subsequence (j^{k_m}) such that $j^{k_m} = 2d$ for all $m \in \mathbb{N}$. But, writing $\mathbf{x}_0 = (0, \ldots, 0, a+1)^T \in \mathbb{R}^d$, this implies that

$$|f(\mathbf{x}_0) - f(\mathbf{0})| = \lim_{m\to\infty} |f^{k_m}(\mathbf{x}_0) - f^{k_m}(\mathbf{0})| = \lim_{m\to\infty} |f_d^{k_m}(a+1)| \geq M(a+1)+1.$$

On the other hand, since $f \in \mathrm{cl}\big(\mathcal{F}_{a+1,M(a+1)}\big)$, we can find $(\tilde{f}^m) \in \mathcal{F}_{a+1,M(a+1)}$ such that $\tilde{f}^m \overset{\mathcal{F}}{\sim} (\tilde{f}_1^m, \ldots, \tilde{f}_d^m, \tilde{c}^m)$ converges pointwise in $\bar{\mathbb{R}}$ to f. So

$$|f(\mathbf{x}_0) - f(\mathbf{0})| = \lim_{m\to\infty} |\tilde{f}^m(\mathbf{x}_0) - \tilde{f}^m(\mathbf{0})| = \lim_{m\to\infty} |\tilde{f}_d^m(a+1)| \leq M(a+1).$$

This contradiction establishes our first claim. Since $\mathcal{F}_{a,M(a)} \cap \mathcal{F}_{a+1,M(a+1)+1} \subseteq \mathcal{F}_{a,M(a),W(a)}$, we deduce that $f \in \mathrm{cl}\big(\mathcal{F}_{a,M(a),W(a)}\big)$ in the case where $M(a) \leq M(a+1)$.

Now if $M(a) > M(a+1)$, then for every $f \in \mathrm{cl}\big(\mathcal{F}_{a,M(a)}\big) \cap \mathrm{cl}\big(\mathcal{F}_{a+1,M(a+1)}\big)$, there exists a sequence (f^k) such that $f^k \in \mathcal{F}_{a+1,M(a+1)}$ and such that f^k converges pointwise in $\bar{\mathbb{R}}$ to f. By the shape restrictions, $\mathcal{F}_{a+1,M(a+1)} \subseteq \mathcal{F}_{a,M(a)}$, so $f^k \in \mathcal{F}_{a,M(a)}$. Consequently, $f^k \in \mathcal{F}_{a,M(a),W(a)}$ as above, so $f \in \mathrm{cl}\big(\mathcal{F}_{a,M(a),W(a)}\big)$.

LEMMA 5. *Under assumptions 1 - 4, for any $a, M_1, M_2, \epsilon > 0$,*

$$\mathbb{E}\big[\{Yf_0(\mathbf{X}) - B(f_0(\mathbf{X}))\}\mathbb{1}_{\{\mathbf{X}\in[-a-1,a+1]^d\}}\big] > \sup_{f\in\mathcal{F}_{a,M_1,M_2}\setminus B_\epsilon(f_0)} \mathbb{E}\big[\{Yf(\mathbf{X}) - B(f(\mathbf{X}))\}\mathbb{1}_{\{\mathbf{X}\in[-a-1,a+1]^d\}}\big].$$

PROOF. Since $B' = g^{-1}$, we have that for every $\mathbf{x} \in [-a-1, a+1]^d$, the expression

$$\mathbb{E}\{Yf(\mathbf{X}) - B(f(\mathbf{X}))|\mathbf{X} = \mathbf{x}\} = g^{-1}(f_0(\mathbf{x}))f(\mathbf{x}) - B(f(\mathbf{x}))$$

is uniquely maximized by taking $f(\mathbf{x}) = f_0(\mathbf{x})$. Moreover, since f_0 is continuous by assumption **4**, it is uniformly continuous on $[-a-1, a+1]^d$. We may therefore assume that for any $\epsilon' > 0$, there exists $\gamma(\epsilon') > 0$ such that $|f_{0,j}(z_1) - f_{0,j}(z_2)| < \epsilon'$ for every $j = 1, \ldots, d$ and every $z_1, z_2 \in [-a-1, a+1]$ with $|z_1 - z_2| < \gamma(\epsilon')$. For any $f \in \mathcal{F}_{a,M_1,M_2} \backslash B_\epsilon(f_0)$, there exists $\mathbf{x}^* = (x_1^*, \ldots, x_d^*)^T \in [-a, a]^d$ such that $|f(\mathbf{x}^*) - f_0(\mathbf{x}^*)| > \epsilon$. Let $C_{\mathbf{x}^*,1} = \times_{j=1}^d D_j \subseteq [-a-1, a+1]^d$ where

$$D_j = \begin{cases} [x_j^*, x_j^* + \min\{\gamma(\frac{\epsilon}{2d}), 1\}] & \text{if } l_j = 2 \\ [x_j^* - \min\{\gamma(\frac{\epsilon}{2d}), 1\}, x_j^*] & \text{if } l_j = 3 \\ \left[x_j^* - \min\left\{\frac{1}{M_2}\frac{\epsilon}{4d}, \gamma(\frac{\epsilon}{4d}), 1\right\}, x_j^* + \min\left\{\frac{1}{M_2}\frac{\epsilon}{4d}, \gamma(\frac{\epsilon}{4d}), 1\right\}\right] & \text{if } l_j \in \{1, 4, 5, 6, 7, 8, 9\}. \end{cases}$$

Define $C_{\mathbf{x}^*,2}$ similarly, but with the intervals in the cases $l_j = 2$ and $l_j = 3$ exchanged. Then the shape constraints ensure that $\max\{\inf_{\mathbf{x} \in C_{\mathbf{x}^*,1}} |f(\mathbf{x}) - f_0(\mathbf{x})|, \inf_{\mathbf{x} \in C_{\mathbf{x}^*,2}} |f(\mathbf{x}) - f_0(\mathbf{x})|\} > \epsilon/2$. But the d-dimensional Lebesgue measures of $C_{\mathbf{x}^*,1}$ and $C_{\mathbf{x}^*,2}$ do not depend on $\mathbf{x}^*$, and $\min\{\mathbb{P}(\mathbf{X} \in C_{\mathbf{x}^*,1}), \mathbb{P}(\mathbf{X} \in C_{\mathbf{x}^*,2})\}$ is a continuous function of $\mathbf{x}^*$, so by assumption **2**, we have

$$\xi = \inf_{\mathbf{x}^* \in [-a,a]^d} \min\{\mathbb{P}(\mathbf{X} \in C_{\mathbf{x}^*,1}), \mathbb{P}(\mathbf{X} \in C_{\mathbf{x}^*,2})\} > 0.$$

Moreover, writing $\underline{f}_0 = \inf_{\mathbf{x} \in [-a-1,a+1]^d} f_0(\mathbf{x})$ and $\bar{f}_0 = \sup_{\mathbf{x} \in [-a-1,a+1]^d} f_0(\mathbf{x})$, and using the fact that $s \mapsto \left[\{g^{-1}(f_0(\mathbf{x}))f_0(\mathbf{x}) - B(f_0(\mathbf{x}))\} - \{g^{-1}(f_0(\mathbf{x}))s - B(s)\}\right]$ is convex, we deduce that

$$\begin{aligned} &\mathbb{E}\left[\{Yf_0(\mathbf{X}) - B(f_0(\mathbf{X}))\}\mathbb{1}_{\{\mathbf{X} \in [-a-1,a+1]^d\}}\right] \\ &\quad - \sup_{f \in \mathcal{F}_{a,M_1,M_2} \backslash B_\epsilon(f_0)} \mathbb{E}\left[\{Yf(\mathbf{X}) - B(f(\mathbf{X}))\}\mathbb{1}_{\{\mathbf{X} \in [-a-1,a+1]^d\}}\right] \\ &\geq \xi \inf_{\mathbf{x} \in [-a-1,a+1]^d} \inf_{|t - f_0(\mathbf{x})| > \epsilon/2} \left[\{g^{-1}(f_0(\mathbf{x}))f_0(\mathbf{x}) - B(f_0(\mathbf{x}))\} - \{g^{-1}(f_0(\mathbf{x}))t - B(t)\}\right] \\ &\geq \frac{1}{16}\xi\epsilon^2 \inf_{s \in [\underline{f}_0 - \epsilon/2, \bar{f}_0 + \epsilon/2]} (g^{-1})'(s) > 0. \end{aligned}$$

LEMMA 6. *For any $a^* > a_0 + 1$, we have*

$$\mathrm{cl}\Big(\mathcal{F}_{a_0,M(a_0)}\backslash B_\epsilon(f_0)\Big) \cap \mathrm{cl}\Big(\mathcal{F}_{a_0+1,M(a_0+1)}\backslash B_\epsilon(f_0)\Big) \cap \mathrm{cl}\Big(\mathcal{F}_{a^*,M(a^*)}\backslash B_\epsilon(f_0)\Big)$$
$$\cap \mathrm{cl}\Big(\mathcal{F}_{a^*+1,M(a^*+1)}\backslash B_\epsilon(f_0)\Big) \subseteq \mathrm{cl}\bigg(\big(\mathcal{F}_{a_0,M(a_0),W(a_0)} \cap \mathcal{F}_{a^*,M(a^*)+1,W(a^*)+1}\big)\backslash B_\epsilon(f_0)\bigg).$$

PROOF. The proof is very similar indeed to the proof of Lemma 4, so we omit the details.

Recall the definition of $\tilde{l}_n(f_0)$ from the proof of Theorem 1 in Section 1.

LEMMA 7. *Suppose that Z has a Poisson distribution with mean $\mu \in (0, \infty)$. Then*

$$\mu \log \mu \leq \mathbb{E}(Z \log Z) \leq \mu \log \mu + 1.$$

It follows that, under the Poisson setting, $\mathbb{E}\{\tilde{l}_n(f_0)\} \in [-1, 0]$.

PROOF. The lower bound is immediate from Jensen's inequality. For the upper bound, let $Z_0 = (Z - \mu)/\sqrt{\mu}$, so $\mathbb{E}(Z_0) = 0$ and $\mathbb{E}(Z_0^2) = 1$. It follows from the inequality $\log(1 + z) \leq z$ for any $z > -1$ that

$$\begin{aligned}
\mathbb{E}(Z \log Z) &= \mathbb{E}\big[(\mu + \sqrt{\mu}Z_0)\{\log \mu + \log(1 + Z_0/\sqrt{\mu})\}1_{\{Z_0 > -\sqrt{\mu}\}}\big] \\
&\leq \mathbb{E}\big[(\mu + \sqrt{\mu}Z_0)(\log \mu + Z_0/\sqrt{\mu})\big] \\
&= \mu \log \mu + (\log \mu + 1)\sqrt{\mu}\mathbb{E}(Z_0) + \mathbb{E}(Z_0^2) = \mu \log \mu + 1.
\end{aligned}$$

Finally, we note that

$$\begin{aligned}
\mathbb{E}\tilde{l}_n(f_0) &= \mathbb{E}\big[\mathbb{E}\{Y f_0(\mathbf{X}) - B(f_0(\mathbf{X})) - Y \log Y + Y|\mathbf{X}\}\big] \\
&= \mathbb{E}\big\{e^{f_0(\mathbf{X})} f_0(\mathbf{X}) - \mathbb{E}(Y \log Y|\mathbf{X})\big\} \in [-1, 0].
\end{aligned}$$

LEMMA 8. *In the Gamma setting, under assumptions 1 and 3, $\mathbb{E}\{\tilde{l}_n(f_0)\} \in (-\infty, 0)$.*

PROOF. Since $-Yf_0(\mathbf{X})|\mathbf{X} \sim \Gamma(1/\phi_0, 1/\phi_0)$, we have

$$\mathbb{E}\tilde{l}_n(f_0) = \mathbb{E}\big[\mathbb{E}\{Yf_0(\mathbf{X}) - B(f_0(\mathbf{X})) - \log Y + 1|\mathbf{X}\}\big] = \mathbb{E}\big[\mathbb{E}\{\log(-Yf_0(\mathbf{X}))|\mathbf{X}\}\big]$$
$$= \log\phi_0 + \psi_D(1/\phi_0) \in (-\infty, 0),$$

where $\psi_D(\cdot)$ denotes the digamma function.

LEMMA 9. *In the Gaussian setting, under assumptions 1 and 2 and conditions 2 and 3,*

$$\mathcal{G}^{\delta}_{a,M_1,M_2} = \Big\{h_{f,\mathbf{A}} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \;\Big|\; h_{f,\mathbf{A}}(\mathbf{x}, y) = -\frac{1}{2}\{f(\mathbf{A}^T\mathbf{x}) - y\}^2 1_{\{\mathbf{x}\in[-a,a]^d\}}$$
$$\text{with } f \in \mathcal{F}_{a,M_1,M_2} \text{ and } \mathbf{A} \in \mathcal{A}^{\delta}\Big\}$$

is P-Glivenko–Cantelli.

PROOF. Following the argument in Step 4 of the proof of Theorem 1, it suffices to show that

$$(\mathring{\mathcal{F}}_{a,M_1,M_2})_j = \Big\{\mathring{f} : \mathbb{R}^d \to \mathbb{R}\Big|\mathring{f}(\mathbf{x}) = f_j(\boldsymbol{\alpha}_j^T\mathbf{x})1_{\{\mathbf{x}\in[-a,a]^d\}}$$
$$\text{for some } f \overset{\mathcal{F}}{\sim} (f_1, \ldots, f_m, c) \in \mathcal{F}_{a,M_1,M_2} \text{and } \boldsymbol{\alpha}_j \in \mathbb{R}^d \text{ with } \|\boldsymbol{\alpha}_j\|_1 = 1\Big\}$$

is P-Glivenko–Cantelli for every $j = 1, \ldots, m$. In the following, we present the proof in case $l_j = 2$. Other cases can be shown in a similar manner.

By Theorem 2.7.5 of van der Vaart and Wellner (1996), there exists a universal constant $C > 0$ such that for any $\epsilon > 0$ and any $\boldsymbol{\alpha}_0 \in \mathbb{R}^d$, there exist functions $g_k^L, g_k^U : \mathbb{R} \to [0,1]$ for $k = 1, \ldots, N_3$ with $N_3 = e^{4M_1C/\epsilon}$ such that $\mathbb{E}|g_k^U(\boldsymbol{\alpha}_0^T\mathbf{X}) - g_k^L(\boldsymbol{\alpha}_0^T\mathbf{X})| \le \epsilon/(4M_1)$ and such that for every monotone function $g : \mathbb{R} \to [0,1]$, we can find $k^* \in \{1, \ldots, N_3\}$ with $g_{k^*}^L \le g \le g_{k^*}^U$. Since $\mathbf{X}$ has a Lebesgue density, for every k we can find $\tau_k^L, \tau_k^U > 0$ such that

$$\mathbb{E}\big|g_k^L(\boldsymbol{\alpha}_0^T\mathbf{X}) - g_k^L(\boldsymbol{\alpha}_0^T\mathbf{X} - \tau_k^L)\big| \le \frac{\epsilon}{8M_1} \quad \text{and} \quad \mathbb{E}\big|g_k^U(\boldsymbol{\alpha}_0^T\mathbf{X} + \tau_k^U) - g_k^L(\boldsymbol{\alpha}_0^T\mathbf{X})\big| \le \frac{\epsilon}{8M_1}.$$

By picking $\tau = \min\{\tau_1^L, \ldots, \tau_N^L, \tau_1^U, \ldots, \tau_N^U\}/a$ (which implicitly depends on $\boldsymbol{\alpha}_0$), we claim that the class of functions

$$\tilde{g}_k^L(\mathbf{x}) = 2M_1(g_k^L(\boldsymbol{\alpha}_0^T\mathbf{x} - \tau a) - 1/2)1_{\{\mathbf{x}\in[-a,a]^d\}}, \tilde{g}_k^U(\mathbf{x}) = 2M_1(g_k^U(\boldsymbol{\alpha}_0^T\mathbf{x} + \tau a) - 1/2)1_{\{\mathbf{x}\in[-a,a]^d\}}$$

for $k = 1, \ldots, N_3$, form an ϵ-bracketing set in the $L_1(P_{\mathbf{X}})$-norm for the set of functions

$$\mathring{\mathcal{F}}^{\boldsymbol{\alpha}_0,\tau}_{a,M_1} = \Big\{ \mathring{f} : \mathbb{R}^d \to \mathbb{R} \Big| \mathring{f}(\mathbf{x}) = f(\boldsymbol{\alpha}^T\mathbf{x}) 1_{\{\mathbf{x} \in [-a,a]^d\}}, \text{ with } f : \mathbb{R} \to \mathbb{R} \text{ increasing}, \sup_{x\in\mathbb{R}} |f(x)| \leq M_1 \text{ and } \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|_1 \leq \tau \Big\}.$$

To see this, we note that

$$\sup_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|_1 \leq \tau, \mathbf{x}\in[-a,a]^d} |\boldsymbol{\alpha}^T\mathbf{x} - \boldsymbol{\alpha}_0^T\mathbf{x}| \leq \tau a.$$

It follows by monotonicity that for $k = 1, \ldots, N_3$,

$$\mathbb{E}|\tilde{g}_k^U(\mathbf{X}) - \tilde{g}_k^L(\mathbf{X})| \leq 2M_1\mathbb{E}\big|g_k^U(\boldsymbol{\alpha}_0^T\mathbf{X} + \tau a) - g_k^U(\boldsymbol{\alpha}_0^T\mathbf{X})\big| + 2M_1\mathbb{E}\big|g_k^U(\boldsymbol{\alpha}_0^T\mathbf{X}) - g_k^L(\boldsymbol{\alpha}_0^T\mathbf{X})\big| + 2M_1\mathbb{E}\big|g_k^L(\boldsymbol{\alpha}_0^T\mathbf{X}) - g_k^L(\boldsymbol{\alpha}_0^T\mathbf{X} - \tau a)\big| \leq \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon.$$

Therefore, $\{\tilde{g}_k^L, \tilde{g}_k^U\}_{k=1}^{N_3}$ is indeed an ϵ-bracketing set.

Now for every $\boldsymbol{\alpha}_0 \in \mathbb{R}^d$ with $\|\boldsymbol{\alpha}_0\|_1 = 1$, we can pick $\tau(\boldsymbol{\alpha}_0) > 0$ such that a finite ϵ-bracketing set can be found for $\mathring{\mathcal{F}}^{\boldsymbol{\alpha}_0,\tau(\boldsymbol{\alpha}_0)}_{a,M_1}$. Since $\{\boldsymbol{\alpha}_0 \in \mathbb{R}^d : \|\boldsymbol{\alpha}_0\|_1 = 1\}$ is compact, we can pick $\boldsymbol{\alpha}_0^1, \ldots, \boldsymbol{\alpha}_0^{N^*}$ such that

$$\{\boldsymbol{\alpha}_0 \in \mathbb{R}^d : \|\boldsymbol{\alpha}_0\|_1 = 1\} \subseteq \bigcup_{k=1,\ldots,N^*} \{\boldsymbol{\alpha} \in \mathbb{R}^d : \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0^k\|_1 \leq \tau(\boldsymbol{\alpha}_0^k)\}.$$

Consequently, for every $\epsilon > 0$, a finite ϵ-bracketing set can be found for $\big(\mathring{\mathcal{F}}_{a,M_1,M_2}\big)_j$. Finally, we complete the proof by applying Theorem 2.4.1 of van der Vaart and Wellner (1996).

3. Running time

Table 1. Average running time (in seconds) of SCMLE, SCAM, GAMIS, MARS, Tree, CAP and MCR on problems 1, 2, 3 with sample sizes $n = 200, 500, 1000, 2000, 5000$ in the Gaussian setting.

Problem 1

Method	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
SCMLE	0.13	0.34	0.90	1.86	7.35
SCAM_{10}	0.91	1.72	4.17	7.43	18.59
SCAM_{20}	5.69	9.91	20.72	39.21	100.45
GAMIS	0.11	0.20	0.46	1.46	3.93
MARS	0.01	0.01	0.02	0.05	0.12
Tree	0.01	0.01	0.02	0.03	0.08
CAP	0.61	1.75	2.47	3.86	8.60
MCR	30.17	411.80	-	-	-

Problem 2

Method	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
SCMLE	0.07	0.18	0.37	0.88	3.03
SCAM_{10}	2.85	3.27	6.26	12.22	29.78
SCAM_{20}	3.62	9.90	20.72	39.21	100.45
GAMIS	0.11	0.20	0.44	1.39	3.92
MARS	0.01	0.01	0.02	0.04	0.09
Tree	0.01	0.01	0.02	0.03	0.07
CAP	0.11	0.32	0.55	0.97	1.93
MCR	33.31	427.98	-	-	-

Problem 3

Method	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
SCMLE	0.35	0.95	2.37	5.41	20.21
SCAM_{10}	23.08	25.77	38.60	70.67	143.91
SCAM_{20}	91.65	101.50	121.41	154.12	249.44
GAMIS	0.45	0.60	1.10	3.19	8.09
MARS	0.01	0.02	0.04	0.08	0.22
Tree	0.01	0.02	0.03	0.05	0.12
CAP	0.10	0.37	0.99	1.83	4.20
MCR	26.61	303.40	-	-	-

Table 2. Average running time (in seconds) of SCMLE, SCAM and GAMIS on problems 1, 2, 3 with sample sizes $n = 200, 500, 1000, 2000, 5000$ in the Poisson and Binomial settings.

Problem 1

Model	Method	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Poisson	SCMLE	0.33	0.78	1.76	3.98	13.08
	SCAM_{10}	1.24	2.40	4.92	9.99	30.54
	SCAM_{20}	6.11	13.11	24.31	51.38	111.52
	GAMIS	0.25	0.50	1.00	2.43	7.08
Binomial	SCMLE	0.24	0.53	1.23	3.22	9.51
	SCAM_{10}	0.80	1.09	1.92	5.24	9.06
	SCAM_{20}	2.47	4.13	6.53	11.52	22.65
	GAMIS	0.25	0.47	0.93	2.49	6.66

Problem 2

Model	Method	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Poisson	SCMLE	0.20	0.41	0.84	1.80	5.10
	SCAM_{10}	1.97	2.67	5.17	11.34	25.35
	SCAM_{20}	7.55	12.28	18.81	29.24	64.14
	GAMIS	0.24	0.42	0.94	2.43	6.62
Binomial	SCMLE	0.16	0.35	0.72	1.49	4.63
	SCAM_{10}	1.82	3.06	6.38	9.60	25.87
	SCAM_{20}	7.37	11.28	19.01	33.47	76.68
	GAMIS	0.24	0.47	0.94	2.34	6.59

Problem 3

Model	Method	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Poisson	SCMLE	0.90	2.29	5.59	12.68	42.58
	SCAM_{10}	8.85	16.93	22.77	39.69	77.08
	SCAM_{20}	25.01	42.43	56.53	69.89	146.92
	GAMIS	0.91	1.62	2.99	7.02	19.01
Binomial	SCMLE	0.46	1.10	2.50	5.37	18.54
	SCAM_{10}	5.80	6.29	8.73	14.10	30.07
	SCAM_{20}	16.98	24.76	38.20	56.47	115.68
	GAMIS	1.18	1.53	2.83	6.93	16.41

Table 3. Average running times (in seconds) of different methods for the shape-constrained additive index models (Problems 4 and 5).

Problem 4

Method	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
SCAIE	4.36	6.61	12.20	23.50	69.52
SSI	26.10	112.44	411.16	1855.37	-
PPR	0.01	0.01	0.01	0.02	0.05
MARS	0.01	0.03	0.05	0.10	0.25
Tree	0.01	0.01	0.01	0.03	0.03
CAP	0.48	1.24	1.90	3.02	6.69
MCR	38.21	496.54	-	-	-

Problem 5

Method	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
SCAIE	3.78	8.76	20.32	62.68	203.20
PPR	0.01	0.02	0.03	0.05	0.12
MARS	0.01	0.01	0.02	0.03	0.04
Tree	0.01	0.01	0.01	0.02	0.03

References

Bhattacharya, R. N. and Rao R. R. (1976) *Normal Approximation and Asymptotic Expansions.* Wiley, New York.

Gerschgorin, S. (1931) Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk*, **6**, 749–754.

Gradshteyn, I. S. and Ryzhik, I. M. (2007) *Table of Integrals, Series, and Products.* Academic Press, San Diego, California.

Lim, E., and Glynn, P. W. (2012) Consistency of multidimensional convex regression. *Oper. Res.*, **60**, 196–208.

Samworth, R. J. and Yuan, M. (2012) Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.*, **40**, 2973–3002.

van der Vaart, A., and Wellner, J. A. (1996) *Weak convergence and empirical processes: with applications to statistics* Springer, New York.

van der Vaart, A., and Wellner, J. A. (2000) Preservation theorems for Glivenko–Cantelli and uniform Glivenko–Cantelli classes. In *High Dimensional Probability II*, 115–133, Birkhauser, Boston.

Yuan, M. (2011) On the identifiability of additive index models. *Statist. Sinica*, **21**, 1901–1911.