ISSN: 1935-7524

Supplementary Material for On Mixing Rates for Bayesian CART

Abstract: This supplementary material contains the description of the Bayesian CART Algorithm (in Section S1) as well as proofs of all the theorems in the main text. In particular, the proof of posterior consistency in Theorem 2.2 is presented in Section S2. The proofs for the mixing rates of Bayesian CART in Theorem 5.1 and Theorem 5.2 are presented in Section S3 and Section S4 and S4.2. The proof of mixing rate upper bound in Theorem 5.3 is presented in Section S5, and the proof of the improved mixing rate of locally informed versions in Theorem 5.4 is presented in Section S6 and S6.2 together with the proof of Remark 10. In Section S7, we provide the mixing rate bound by applying the result of [60] to our settings in order to compare the bounds and Section S8 contains some additional

Contents

S1 Bayesian CART Algorithm 2

S2 Proof of Theorem 2.2 (Establishing Consistency) 2
- S2.0.1 Trees do not overfit. . . . 3
- S2.0.2 Trees do not underfit. . . . 5

S3 Proof of Theorem 5.1 (Bayesian CART Mixing Lower Bound) 7

S4 Proof of Theorem 5.2 (Bayesian CART Mixing Upper Bound) 10
- S4.1 Canonical Path Ensemble for Bayesian CART 10
- S4.2 Proof of Theorem 5.2 12
- S4.3 Proof of Lemma S2 14
- S4.4 Proof of Lemma S3 14
  - S4.4.1 When $\mathcal{T}^* \subset \mathcal{T}$ (The Overfitted Case) 15
  - S4.4.2 When $\mathcal{T}^* \not\subset \mathcal{T}$ (The Underfitted Case) 17

S5 Proof of Theorem 5.3 (Twiggy Bayesian CART Mixing Upper Bound) 19
- S5.1 Canonical Path Ensemble for Twiggy Bayesian CART 19
- S5.2 Version of Lemma S2 for Twiggy Bayesian CART 20
- S5.3 Version of Lemma S3 for Twiggy Bayesian CART 21
  - S5.3.1 When $\mathcal{T}^* \subset \mathcal{T}$ (The Overfitted Case) 21
  - S5.3.2 When $\mathcal{T}^* \not\subset \mathcal{T}$ (The Underfitted Case) 23

S6 Proof of Theorem 5.4 (Mixing Upper Bound for Locally Informed Versions) 25
- S6.1 Two Drift Conditions 25
- S6.2 Proof of Theorem 5.4 26
  - S6.2.1 Application of general two-stage drift condition 27
- S6.3 Proof of Proposition S2 29
  - S6.3.1 Drift condition for overfitted models (R_2) 31
  - S6.3.2 Drift condition for underfitted models (R_1) 34
- S6.4 Proof of Remark 10 37

S7 Comparison to [60] 42

S8 Additional Visualizations 44

S1. Bayesian CART Algorithm

The algorithmic description of the original Bayesian CART (dyadic version) is in Algorithm 1.

Algorithm 1 *Original Bayesian CART (Dyadic Version).*

Input

Input: The maximum iteration number T_{max}, the initial tree $\mathcal{T}^0$, the posterior $\Pi(\mathcal{T} \mid Y)$

Sampling

For $i = 1, ..., T_{max}$

Sample $u_i \sim Unif(0, 1)$

If $u_i > 0.5$ or $\mathcal{T}^i = \mathcal{T}_{null}$, propose a new candidate tree by GROW

Else, propose a new candidate tree $\widetilde{\mathcal{T}}$ by PRUNE

GROW

Randomly pick a terminal node $(l^*, k^*) \in \mathcal{T}^i_{ext}$.

Split (l^*, k^*) into two daughter nodes by splitting the interval I_{lk} at a dyadic rational midpoint[a] by

$\widetilde{\mathcal{T}}_{int} \leftarrow \mathcal{T}^i_{int} \cup \{(l^*, k^*)\}$

$\widetilde{\mathcal{T}}_{ext} \leftarrow \mathcal{T}^i_{ext} \backslash \{(l^*, k^*)\} \cup \{(l^* + 1, 2k^*), (l^* + 1, 2k^* + 1)\}$

Set $\mathcal{T}^{i+t} = \tilde{\mathcal{T}}$ with probability $\alpha(\mathcal{T}^i, \widetilde{\mathcal{T}}) = \min\left\{1, \frac{\Pi(\widetilde{\mathcal{T}} \mid Y)|\mathcal{T}^i_{ext}|}{\Pi(\mathcal{T}^i \mid Y)|\mathcal{P}(\widetilde{\mathcal{T}})|}\right\}$

PRUNE

Randomly pick a parent of two terminal nodes $(l^*, k^*) \in \mathcal{P}(\mathcal{T}^i)$.

Collapse the nodes below it and turn it into a terminal node by

$\widetilde{\mathcal{T}}_{int} \leftarrow \mathcal{T}^i_{int} \backslash \{(l^*, k^*)\}$.

$\widetilde{\mathcal{T}}_{ext} \leftarrow \mathcal{T}^i_{ext} \backslash \{(l^* + 1, 2k^*), (l^* + 1, 2k^* + 1)\} \cup \{(l^*, k^*)\}$.

Set $\mathcal{T}^{i+t} = \tilde{\mathcal{T}}$ with probability $\alpha(\mathcal{T}^i, \widetilde{\mathcal{T}}) = \min\left\{1, \frac{\Pi(\widetilde{\mathcal{T}} \mid Y)|\mathcal{P}(\mathcal{T}^i)|}{\Pi(\mathcal{T}^i \mid Y)|\widetilde{\mathcal{T}}_{ext}|}\right\}$

[a] This can be extended to the fullblown original version by first choosing a direction and a split point uniformly.

S2. Proof of Theorem 2.2 (Establishing Consistency)

We assume that the truth f_0 is a step function as in Assumption 1 (a) or (b) with signals $\mathcal{B}(A) \equiv \{(l, k) : C_{f_0} > |\beta^*_{lk}| > A \log n/\sqrt{n}\} \subseteq \{(l, k) : l < L\}$. Recall that $\mathcal{T}^*$ is the smallest tree that includes $\mathcal{B}(A)$ as internal nodes and $\mathcal{T}^L_{full} = \{(l, k) : l < L\}$ is the full tree up to depth L. Recall the $(n \times p)$ Haar wavelet regression matrix $\boldsymbol{X}$ with wavelets up to the maximal

resolution L_{max} (i.e. $p = n/2$). We will work conditionally on the event space $\mathcal{A}_n$ defined as

$$\mathcal{A}_n \equiv \{\boldsymbol{\varepsilon} : \|\boldsymbol{X}'\boldsymbol{\varepsilon}\|_\infty \leq 2\|\boldsymbol{X}\|\sqrt{\log p}\}, \tag{S1}$$

where $\|\boldsymbol{X}\| = \max_{1\leq j\leq p} \|X_j\|_2$. It is known that $\mathbb{P}(\mathcal{A}_n^c) \leq 2/p = 4/n \to 0$.

We split the set of eligible trees $\mathbb{T} = \mathbb{T}_L$ into

$$\mathbb{T} = \mathcal{T}^* \cup \mathbb{T}_U \cup \mathbb{T}_O,$$

where $\mathbb{T}_U = \{\mathcal{T} \in \mathbb{T} : \mathcal{T}^* \not\subseteq \mathcal{T}\}$ are all under-fitted trees that miss at least one internal node inside $\mathcal{T}^*_{int}$ and $\mathbb{T}_O = \{\mathcal{T} \in \mathbb{T} : \mathcal{T}^* \subset \mathcal{T}\}$ are all over-fitted trees that include inside at least one redundant internal node in $\mathcal{T}_{full}\backslash\mathcal{T}^*_{int}$. We show below that on the event $\mathcal{A}_n$ we have $\Pi[\mathbb{T}_O \mid Y] = o(1)$ and $\Pi[\mathbb{T}_U \mid Y] = o(1)$ for $c > 5/2$.

S2.0.1. Trees do not overfit.

We decompose the overfitted set $\mathbb{T}_O = \bigcup_{K=1}^{2^L} \Lambda(\mathcal{T}^*, K)$ into shells depending on how many extra internal nodes the overfitted tree $\mathcal{T} \in \mathbb{T}_O$ has relative to $\mathcal{T}^*$, where

$$\Lambda(\mathcal{T}^*, K) = \{\mathcal{T} \in \mathbb{T}_O : |\mathcal{T}_{int}| - |\mathcal{T}^*_{int}| = K\}.$$

We can write

$$\frac{\Pi[\Lambda(\mathcal{T}^*, K) \mid Y]}{\Pi(\mathcal{T}^* \mid Y)} = \sum_{\mathcal{T}\in\Lambda(\mathcal{T}^*,K)} \frac{\Pi(\mathcal{T})N_{\mathcal{T}}(Y)}{\Pi(\mathcal{T}^*)N_{\mathcal{T}^*}(Y)} \tag{S2}$$

where the marginal likelihood ratio can be written as (using the expression in (8))

$$\frac{N_{\mathcal{T}}(Y)}{N_{\mathcal{T}^*}(Y)} = (1+n)^{-K/2} \exp\left\{\frac{1}{2(n+1)} Y'[\boldsymbol{X}_{\mathcal{T}}\boldsymbol{X}'_{\mathcal{T}} - \boldsymbol{X}_{\mathcal{T}^*}\boldsymbol{X}'_{\mathcal{T}^*}]Y\right\}.$$

For $\mathcal{T} \in \Lambda(\mathcal{T}^*, K)$ we denote with $\mathcal{T}^0 \equiv \mathcal{T}^* \to \mathcal{T}^1 \to \cdots \to \mathcal{T}^K \equiv \mathcal{T}$ the sequence of nested trees obtained from $\mathcal{T}^*$ by growing one internal node (at a depth l_j) at a time towards reaching $\mathcal{T}$. We will use a shorthand notation $p_l = p_{lk}$ for the split probability. The prior ratio of two consecutive trees in this sequence satisfies

$$\frac{\Pi(\mathcal{T}^j)}{\Pi(\mathcal{T}^{j-1})} = \frac{p_{l_j}}{1-p_{l_j}} \times (1-p_{l_j+1})^2$$

Then we find

$$\frac{\Pi(\mathcal{T})N_Y(\mathcal{T})}{\Pi(\mathcal{T}^*)N_Y(\mathcal{T}^*)} = (1+n)^{-K/2}\prod_{j=1}^{K}\frac{p_{l_j}}{1-p_{l_j}}\times(1-p_{l_j+1})^2\times\exp\left\{-\frac{Y'(P_{j-1}-P_j)Y}{2(n+1)}\right\}$$
$$= (1+n)^{-K/2}\prod_{j=1}^{K}\frac{p_{l_j}}{1-p_{l_j}}\times(1-p_{l_j+1})^2\times\exp\left\{\frac{|X'_{[j]}Y|^2}{2(n+1)}\right\}, \tag{S3}$$

where

$$P_j = \boldsymbol{X}_{\mathcal{T}^j}\boldsymbol{X}'_{\mathcal{T}^j} = P_{j-1} + X_{[j]}X'_{[j]}$$

and where $X_{[j]}$ is the column added at the j^{th} step of branch growing. Since $\mathcal{T}^*_{int}$ contains *all* signals, we have $\boldsymbol{\beta}^*_{\setminus\mathcal{T}^*} = \mathbf{0}$. Then for any $j = 1,\dots,K$ we have on the event $\mathcal{A}_n$ (since $\boldsymbol{\nu} = \boldsymbol{\epsilon}$ under Assumption 1 and due to orthogonality of X)

$$|X'_{[j]}Y| = |X'_{[j]}(\boldsymbol{X}_{\mathcal{T}^*}\boldsymbol{\beta}^*_{\mathcal{T}^*} + \boldsymbol{X}_{\setminus\mathcal{T}^*}\boldsymbol{\beta}^*_{\setminus\mathcal{T}^*} + \boldsymbol{\nu})| \le 2\sqrt{n\log n}$$

Using $p_l = p_{lk} = n^{-c} < 1/2$ we obtain

$$\frac{\Pi(\mathcal{T})N_Y(\mathcal{T})}{\Pi(\mathcal{T}^*)N_Y(\mathcal{T}^*)} \le \exp\left(-\frac{K}{2}\log(1+1/n) - K\,(c-3/2)\log n\right) \le \mathrm{e}^{-K(c-3/2)\log n}. \tag{S4}$$

Noting that the cardinality of $\Lambda(\mathcal{T}^*, K)$ can be for each K bounded by

$$\mathrm{card}[\Lambda(\mathcal{T}^*, K)] \le \prod_{j=1}^{K}(|\mathcal{T}^*_{ext}| + j - 1)$$

we find an upper bound for (S2)

$$\frac{\Pi[\Lambda(\mathcal{T}^*, K)\,|\,Y]}{\Pi(\mathcal{T}^*\,|\,Y)} \le (|\mathcal{T}^*_{ext}| + K - 1)^K \mathrm{e}^{-K(c-3/2)\log n}. \tag{S5}$$

Since

$$\frac{\Pi(\mathbb{T}_O\,|\,Y)}{\Pi(\mathcal{T}^*\,|\,Y)} \le \sum_{K=1}^{2^L}\frac{\Pi(\Lambda(\mathcal{T}^*, K)\,|\,Y)}{\Pi(\mathcal{T}^*\,|\,Y)} < \sum_{K=1}^{2^L}\mathrm{e}^{K\log[(|\mathcal{T}^*_{ext}|+K-1)]}\mathrm{e}^{-K\,(c-3/2)\,\log n}$$
$$< \sum_{k=1}^{2^L}\mathrm{e}^{-K(c-5/2)\log n} \le n^{5/2-c}\frac{1-[n^{(5/2-c)}]^{n/2}}{1-n^{5/2-c}} < \frac{1}{n^{c-5/2}-1} \tag{S6}$$

we obtain that $\Pi[\mathbb{T}_O\,|\,Y] < \frac{1}{n^{c-5/2}-1}$.

S2.0.2. Trees do not underfit.

We now show that the probability of trees that miss at least one signal goes to zero. In particular, we show that (on the event $\mathcal{A}_n$) we have

$$\Pi\left[\mathcal{T} \in \mathbb{T} : \mathcal{B}(A) \nsubseteq \mathcal{T}_{int} \mid Y\right] \to 0 \quad \text{as } n \to \infty \tag{S7}$$

for

$$\mathcal{B}(A) \equiv \{(l,k) : C_{f_0} > |\beta^*_{lk}| > A \log n/\sqrt{n}\}. \tag{S8}$$

The proof of (S7) follows the route of Lemma 3 in [9] and Section 9.0.2 in [51]. For simplicity, we have focused in this work on the regular design case where the regression matrix is orthogonal and thereby $\Sigma_{\mathcal{T}} = c_n(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}})^{-1} = \frac{1}{n+1} I_{|\mathcal{T}_{ext}|}$. Suppose that $(l_S, k_S) \in \mathcal{B}(A)$ is a signal node for some $A > 0$ and let $\mathcal{T}$ be such that $(l_S, k_S) \notin \mathcal{T}$. We grow a branch from $\mathcal{T}$ that extends towards (l_S, k_S) to obtain an enlarged tree $\mathcal{T}^+ \supset \mathcal{T}$. In other words $\mathcal{T}^+$ is the smallest tree that contains $\mathcal{T}$ and (l_S, k_S) as an internal node. For details, we refer to Lemma 3 in [9]. We define $K = |\mathcal{T}^+_{int} \backslash \mathcal{T}_{int}|$ and write (using the expression in (8))

$$\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)} = (1+n)^{K/2} \exp\left\{\frac{1}{2(n+1)} Y'[\boldsymbol{X}_{\mathcal{T}}\boldsymbol{X}'_{\mathcal{T}} - \boldsymbol{X}_{\mathcal{T}^+}\boldsymbol{X}'_{\mathcal{T}^+}]Y\right\}. \tag{S9}$$

We denote with $\mathcal{T}^0 = \mathcal{T} \to \mathcal{T}^1 \to \cdots \to \mathcal{T}^K = \mathcal{T}^+$ the sequence of nested trees obtained by adding one additional internal node (l_j, k_j) towards (l_S, k_S). Then we find

$$\begin{aligned}\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)} &= (1+n)^{K/2} \prod_{j=1}^{K} \exp\left\{\frac{Y'(P_{j-1} - P_j)Y}{2(n+1)}\right\} \\ &= (1+n)^{K/2} \prod_{j=1}^{K} \exp\left\{-\frac{|X'_{[j]}Y|^2}{2(n+1)}\right\},\end{aligned} \tag{S10}$$

where $P_j = \boldsymbol{X}_{\mathcal{T}^j}\boldsymbol{X}'_{\mathcal{T}^j} = P_{j-1} + X_{[j]}X'_{[j]}$ and where $X_{[j]}$ is the column added at the j^{th} step of branch growing. Let $X_{[K]}$ be the *last* column to be added to $\boldsymbol{X}_{\mathcal{T}^+}$, i.e. the *signal* column associated with (l_S, k_S). We will be denoting simply $\beta^*_{[K]} \equiv \beta^*_{(l_S,k_S)}$ the coefficient associated with $X_{[K]}$. Then (from the orthogonality of X)

$$|X'_{[K]}Y|^2 = |X'_{[K]}X_{[K]}\beta^*_{[K]} + X'_{[K]}\boldsymbol{\nu}|^2$$

Using the inequality $(a+b)^2 \geq a^2/2 - b^2$ we find that

$$|X'_{[K]}Y|^2 \geq n^2|\beta^*_{[K]}|^2/2 - |X'_{[K]}\boldsymbol{\nu}|^2.$$

On the event $\mathcal{A}_n$ and using the fact that $F_0 - \boldsymbol{X}\boldsymbol{\beta}^* = 0$ under the step function Assumption 1 we find that

$$|X'_{[K]}\boldsymbol{\nu}| = |X'_{[K]}\boldsymbol{\varepsilon}| \leq 2\sqrt{n\log n}$$

which yields

$$\frac{|X'_{[K]}Y|^2}{2(n+1)} \geq \frac{n^2|\beta^*_{[K]}|^2}{4(n+1)} - \frac{4n\log n}{2(n+1)}.$$

From the signal assumption $(l_S, k_S) \in \mathcal{B}(A)$, we have $|\beta^*_{[K]}| > A\log n/\sqrt{n}$ for some $A > 0$ and thereby

$$\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)} \leq \exp\left\{\frac{K}{2}\log(1+n) - \frac{nA^2\log^2 n}{4(n+1)} + \frac{4n\log n}{2(n+1)}\right\} \tag{S11}$$

The prior ratio satisfies (using again the notation $p_l = p_{lk}$)

$$\frac{\Pi(\mathcal{T})}{\Pi(\mathcal{T}^+)} = \frac{1-p_{l_0}}{p_{l_0}} \times \left(\prod_{j=1}^{K-1}\frac{1}{p_{l_j}(1-p_{l_j})}\right) \times \frac{1}{(1-p_{l_K})^2}. \tag{S12}$$

Defining

$$b(n) \coloneqq \frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}^+ \mid Y)}$$

with $p_l = p_{lk} = n^{-c} < 1/2$, we have $\Pi(\mathcal{T})/\Pi(\mathcal{T}^+) \leq 2^K \mathrm{e}^{c\,K\log n}$, and thereby (since $K \leq L \leq L_{max} = \log_2[n/2]$)

$$b(n) \leq 2^K \exp\left\{c\,K\log n + \frac{K}{2}\log(1+n) - \frac{nA^2\log^2 n}{4(n+1)} + \frac{4n\log n}{2(n+1)}\right\}. \tag{S13}$$

Following the proof technique in Lemma 2 in [9] we conclude that for some sufficiently large $A > 0$

$$\Pi[(l_S, k_S) \notin \mathcal{T}_{int} \mid Y] \leq l_S \times b(n) \leq \mathrm{e}^{-(A^2/4)\log^2 n}.$$

Thereby,

$$\Pi[\mathcal{B}(A) \nsubseteq \mathcal{T}_{int} \,|Y] \leq \sum_{(l_S,k_S)\in\mathcal{B}(A)} \Pi[(l_S, k_S) \notin \mathcal{T}_{int} \,|Y] \leq \mathrm{e}^{-(A^2/4)\log^2 n} 2^L \leq \mathrm{e}^{-(A^2/8)\log^2 n} \to 0.$$

This concludes the proof of (S7). Because $\mathcal{T}^*$ is the minimal tree that contains $\mathcal{B}(A)$ as its internal nodes, this implies $\Pi[\mathcal{T} \in \mathbb{T} : \mathcal{T}^* \not\subseteq \mathcal{T} \mid Y] = \Pi[\mathbb{T}_U \mid Y] = o(1)$.

S3. Proof of Theorem 5.1 (Bayesian CART Mixing Lower Bound)

We assume that the true signal $f_0(x) = \psi_{l*k*}(x)$ consists of just one deepest leftmost wavelet coefficient with $0 < l^* < L$ and $k^* = 0$, where $|\beta^*_{l^*k^*}| > A \log n/\sqrt{n}$ according to Assumption 1 (b). Figure 1 (a) illustrates a special case when $l^* = 3$. During the proof, we take advantage of the bottleneck ratio bound [54]

$$Gap(P) \leq 2\Phi, \quad \text{where} \quad \Phi = \min_{\substack{A \subset \mathbb{T} \\ 0 < \Pi[A \mid Y] \leq 1/2}} \frac{\sum_{\mathcal{T} \in A, \mathcal{T}' \in \mathbb{T} \setminus A} \Pi(\mathcal{T} \mid Y) P(\mathcal{T}, \mathcal{T}')}{\Pi[A \mid Y]} \tag{S14}$$

is the conductance which measures the ability of the chain to escape from any small region of the state space (and make a rapid progress to the equilibrium).

We now choose $A \subset \mathbb{T}$ that gives a small value of the ratio inside the minimum in (S14), thereby providing a small upper bound of the conductance. Intuitively, among trees without the signal, the posterior is smaller for deeper trees. Recall that in Bayesian CART, the transition probability is non-zero only between trees that differ by one internal node. The signal node (l^*, k^*) is only reachable from trees that include $(l^* - 1, 0)$. The set of trees that include $(l^* - 1, 0)$ thus comprises a bottleneck between trees that capture the signal node $(l^*, 0)$ and those that do not. Using this intuition, we will calculate the bottleneck ratio w.r.t.

$$A_{\setminus (l^*-1,0)} \equiv \{\mathcal{T} \in \mathbb{T} \mid (l^* - 1, 0) \notin \mathcal{T}_{int}\} \tag{S15}$$

to bound the conductance. Note that $\Pi[A_{\setminus (l^*-1,0)} \mid Y] < 1/2$ since the posterior is concentrated to the true tree with $(l^*, 0)$ (See, Theorem 2.2)[1]. A tree $\mathcal{T} \in A_{\setminus (l^*-1,0)}$ must contain $(l^* - 2, 0)$ to have a non-zero transition probability $P(\mathcal{T}, \mathcal{T}')$ for $\mathcal{T}' \in A^c_{\setminus (l^*-1,0)}$. Therefore, denoting

$$B_{l^*-1} \equiv A_{\setminus (l^*-1,0)} \cap \{\mathcal{T} \in \mathbb{T} \mid (l^* - 2, 0) \in \mathcal{T}_{int}\},$$

[1] By consistency established in Theorem 2.2, on the event space $\mathcal{A}_n$ with $c > 5/2$, we have $\Pi(\mathcal{T}^* \mid Y) > 1/2$ with probability at least $1 - 4/n$ when the signal is large enough. Since $\mathcal{T}^* \notin A_{\setminus (l^*-1,0)}$, we have $\Pi[A_{\setminus (l^*-1,0)} \mid Y] \leq 1/2$.

the bottleneck ratio w.r.t. $A_{\setminus(l^*-1,0)}$ bounds the conductance from above simply by

$$\Phi \leq \frac{\sum_{\mathcal{T}\in B_{l^*-1}} \Pi(\mathcal{T} \mid Y) P(\mathcal{T}, A^c_{\setminus(l^*-1,0)})}{\Pi[A_{\setminus(l^*-1,0)} \mid Y]} \leq \frac{\Pi[B_{l^*-1} \mid Y]}{\Pi[A_{\setminus(l^*-1,0)} \mid Y]}. \tag{S16}$$

We now show that the tightest upper bound in (S16) is obtained when $l^* = L-1$. Namely, we first derive a bound w.r.t. a general l^*, and show that the bound in (S16) becomes smaller as l^* increases. We will work conditionally on the set $\mathcal{A}_n$ defined in (S1). Recall the definition of $\mathcal{T}^*$ from Assumption 1 (b) as the minimal tree that contains the signal $\mathcal{B} = \{(l^*, 0)\}$.

To bound the ratio in (S16), we decompose $A_{\setminus(l^*-1,0)}$ into l^* disjoint subsets that contain the leftmost node at a certain level and exclude the leftmost node at the next level:

$$B_i \equiv \{\mathcal{T} \in \mathbb{T} \mid (i-1,0) \in \mathcal{T}_{int}, (i,0) \notin \mathcal{T}_{int}\} \quad \text{for} \quad i = 0, 1, \ldots, l^*-1.$$

It is easy to see that $A_{\setminus(l^*-1,0)} = \bigcup_{i=0}^{l^*-1} B_i$, so that

$$\Pi[A_{\setminus(l^*-1,0)} \mid Y] = \sum_{i=0}^{l^*-1} \Pi[B_i \mid Y]. \tag{S17}$$

Therefore, the bound in (S16) can be rewritten as

$$\Phi \leq \frac{\Pi[B_{l^*-1} \mid Y]}{\Pi[A_{\setminus(l^*-1,0)} \mid Y]} = \frac{\Pi[B_{l^*-1} \mid Y]}{\Pi[B_{l^*-1} \mid Y] + \sum_{i=0}^{l^*-2} \Pi[B_i \mid Y]}.$$

To see how small $\Pi[B_{l^*-1} \mid Y]$ is compared to $\sum_{i=0}^{l^*-2} \Pi[B_i \mid Y]$, we will scrutinize the posterior ratio $\Pi[B_{i-1} \mid Y]/\Pi[B_i \mid Y]$ for each B_i. We first characterize a simple relationship between B_i and B_{i-1} where each tree in B_i can be obtained by attaching a "mini-tree" to some tree inside B_{i-1}. For each $\mathcal{T} \in B_i$, we denote with $M(\mathcal{T})$ the operator that removes all descendants $D_{i-1,0}(\mathcal{T})$ of the node $(i-1,0)$, i.e. for $\mathcal{T}' = M(\mathcal{T})$

$$\mathcal{T}'_{int} = \mathcal{T}_{int} \setminus D_{i-1,0}(\mathcal{T}).$$

The mapping $M(\cdot)$ is many to one and for each $\mathcal{T}' \in B_{i-1}$ we denote

$$\mathcal{N}(\mathcal{T}') = \{\mathcal{T} \in B_i : \mathcal{T}' = M(\mathcal{T})\}$$

the nonempty set of those $\mathcal{T} \in B_i$ that map onto the same $\mathcal{T}'$. Then we can write

$$\Pi(B_i \mid Y) = \sum_{\mathcal{T}\in B_i} \frac{\Pi(\mathcal{T} \mid Y)}{\Pi(M(\mathcal{T}) \mid Y)} \Pi(M(\mathcal{T}) \mid Y) = \sum_{\mathcal{T}'\in B_{i-1}} \Pi(\mathcal{T}' \mid Y) \sum_{\mathcal{T}\in\mathcal{N}(\mathcal{T}')} \frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}' \mid Y)}.$$

Each tree $\mathcal{T} \in \mathcal{N}(\mathcal{T}')$ differs from $\mathcal{T}'$ by addition of at least one node without signal. We decompose $\mathcal{N}(\mathcal{T}') = \cup_{K=1}^{2^{Lmax}} \mathcal{N}(\mathcal{T}', K)$ into shells according to how many extra noise nodes the trees have, where $\mathcal{N}(\mathcal{T}', K) = \{\mathcal{T} \in \mathcal{N}(\mathcal{T}') : |\mathcal{T}_{int} \backslash \mathcal{T}'_{int}| = K\}$. We can use the posterior ratio expression (S4) for nested models to conclude that for any $\mathcal{T} \in \mathcal{N}(\mathcal{T}', K)$ we have

$$\frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}' \mid Y)} \leq \mathrm{e}^{-K(c-3/2)\log n}$$

The cardinality of the set $\mathcal{N}(\mathcal{T}', K)$ is at most the number of all binary trees with K nodes. This corresponds to the Catalan number $\mathbb{C}_K$, which according to Lemma S-3 in [9], satisfies $\mathbb{C}_K \asymp 4^K / K^{3/2}$. Then

$$\Pi(B_i \mid Y) \lesssim \sum_{\mathcal{T}' \in B_{i-1}} \Pi(\mathcal{T}' \mid Y) \times \sum_{K=1}^{n/2} 4^K \mathrm{e}^{-K(c-3/2)\log n} \leq \frac{1}{n^{(c-3/2)}/4 - 1} \times \Pi(B_{i-1} \mid Y).$$

Denoting with $\gamma_n = \frac{C}{n^{(c-3/2)}/4-1}$ the "shrinkage factor" for some $C > 1$, the posterior of $A_{\backslash(l^*-1,0)}$ satisfies

$$\begin{aligned}\Pi[A_{\backslash(l^*-1,0)} \mid Y] &= \sum_{i=0}^{l^*-1} \Pi[B_i \mid Y] \geq \Pi[B_{l^*-1} \mid Y] \sum_{i=0}^{l^*-1} \left(\frac{1}{\gamma_n}\right)^i \\ &= \Pi[B_{l^*-1} \mid Y] \frac{\gamma_n}{1-\gamma_n} \left[\left(\frac{1}{\gamma_n}\right)^{l^*} - 1\right]. \end{aligned} \tag{S18}$$

Therefore, it follows from (S18) and (S16) that

$$\begin{aligned}\Phi^{-1} &\geq \frac{\Pi[A_{\backslash(l^*-1,0)} \mid Y]}{\Pi[B_{l^*-1} \mid Y]} \geq \frac{\gamma_n}{1-\gamma_n} \left[\left(\frac{1}{\gamma_n}\right)^{l^*} - 1\right] \\ &\geq \frac{C}{n^{(c-3/2)}/4 - C} \left[\left(\frac{n^{(c-3/2)}/4-1}{C}\right)^{l^*} - 1\right] > \left(\frac{n^{(c-3/2)}/4-1}{C}\right)^{l^*-1} - 1\end{aligned}$$

where we used the fact that $\gamma_n > 4C/n^{(c-3/2)}$ and that $C/(n^{(c-3/2)}/4-1) < 1$ for large enough n. Therefore, we have

$$\frac{1}{Gap(P)} \geq \frac{1}{2\Phi} \geq \frac{1}{2}\left(\left(\frac{n^{(c-3/2)}/4-1}{C}\right)^{l^*-1} - 1\right).$$

As this quantity increases with l^*, the maximum is reached for $l^* = L - 1$. By applying the mixing time lower bound in (19) using the spectral gap, we obtain

$$\tau_\epsilon \geq \log\left(\frac{1}{2\epsilon}\right)\frac{1}{2}\left[\frac{1}{Gap(P)} - 1\right] > \log\left(\frac{1}{2\epsilon}\right)\frac{1}{4}\left[\left(\frac{n^{(c-3/2)}/4-1}{C}\right)^{L-2} - 3\right].$$

S4. Proof of Theorem 5.2 (Bayesian CART Mixing Upper Bound)

The proof of Theorem 5.2 rests on the canonical path argument and the sandwich relation (19). Together with (20), this yields $\tau_\epsilon \leq l(\mathcal{E})\rho(\mathcal{E})\big(\log[1/\min_{\mathcal{T}\in\mathbb{T}_L}\Pi(\mathcal{T}\mid Y)] + \log 1/\epsilon\big)$. In the next section, we present Lemma S2 and Lemma S3 which provide an upper bound for the first two terms on the right side. The logarithmic term is handled by the posterior consistency result in Theorem 2.2. In the next section, we provide details of the canonical path construction and describe basic properties of our canonical path ensemble. Similarly as in [60], whose canonical path architecture was inspired by stepwise variable selection, our construction was inspired by the CART algorithm [4].

S4.1. Canonical Path Ensemble for Bayesian CART

We denote with $\mathcal{T}^*$ the signal-spanning tree from Assumption 1. First, we construct a canonical path $T_{\mathcal{T},\mathcal{T}^*}$ from any tree $\mathcal{T} \in \mathbb{T}_L\backslash\{\mathcal{T}^*\}$ towards $\mathcal{T}^*$ *along edges* in the graph with a transition matrix P. To this end, we introduce the *transition function* $\mathcal{G} : \mathbb{T}_L\backslash\mathcal{T}^* \to \mathbb{T}_L$ that maps the current state $\mathcal{T} \in \mathbb{T}_L$ onto the next state $\mathcal{G}(\mathcal{T}) \in \mathbb{T}_L$ that is *"closer"* to $\mathcal{T}^*$, where closeness is determined by the Hamming distance $h(\mathcal{T},\mathcal{T}^*)$ between binary tree encodings[2]. The canonical path $T_{\mathcal{T},\mathcal{T}^*} = \{\mathcal{T}^0,\mathcal{T}^1,\ldots,\mathcal{T}^k\}$ is constructed by composing the transition function so that

$$\mathcal{T}^0 \equiv \mathcal{T} \to \mathcal{T}^1 \equiv \mathcal{G}(\mathcal{T}) \to \cdots \to \mathcal{T}^k \equiv \mathcal{G}^k(\mathcal{T}) \equiv \mathcal{T}^*,$$

where $\mathcal{G}^k(\cdot) = \mathcal{G}\circ\cdots\circ\mathcal{G}(\cdot)$ is a composition of $\mathcal{G}$. Below, we describe one particular transition function $\mathcal{G}(\mathcal{T})$ which reduces the (Hamming) distance after each step, i.e. $h[\mathcal{G}(\mathcal{T}),\mathcal{T}^*] < h(\mathcal{T},\mathcal{T}^*)\ \forall\mathcal{T} \in \mathbb{T}_L\backslash\mathcal{T}^*$. The mapping corresponds to a deterministic version of the PRUNE and GROW steps of the Bayesian CART algorithm from Section 2.1.2.

[2] A binary tree encoding consists of a $(2^L \times 1)$ ordered (according to $2^l + k$) binary vector indicating whether or not $(l,k) \in \mathcal{T}_{int}$.

(1) Assume $\mathcal{T} \supset \mathcal{T}^*$ is **overfitted**, i.e. $\mathcal{T}$ forms an envelope around $\mathcal{T}^*$ and contains at least one signal-less node. The mapping $\mathcal{G}(\cdot)$ finds the deepest rightmost redundant node, say $(l,k) \in \mathcal{T}_{int} \backslash \mathcal{T}^*_{int}$, and turns it into a bottom node. More formally $\mathcal{G}(\mathcal{T}) = \mathcal{T}^-$ where

$$\mathcal{T}^-_{int} = \mathcal{T}_{int} \backslash \{(l,k)\} \quad \text{and} \quad \mathcal{T}^-_{ext} = \mathcal{T}_{ext} \backslash \{(l+1,2k),(l+1,2k+1)\} \cup \{(l,k)\} \tag{S19}$$

where $(l,k) = \arg \max_{(l',k') \in \mathcal{T}_{int} \backslash \mathcal{T}^*_{int}} (2^{l'} + k')$.

(2) Assume $\mathcal{T} \not\supseteq \mathcal{T}^*$ is **underfitted**, i.e. $\mathcal{T}$ misses at least one influential node in $\mathcal{T}^*$.

(i) If $\mathcal{T} \subset \mathcal{T}^*$, the mapping $\mathcal{G}(\cdot)$ finds the deepest rightmost external node in $\mathcal{T}_{ext} \backslash \mathcal{T}^*_{int}$, say (l,k), and turns it into an internal node. More formally $\mathcal{G}(\mathcal{T}) = \mathcal{T}^+$ where

$$\mathcal{T}^+_{int} = \mathcal{T}_{int} \cup \{(l,k)\} \quad \text{and} \quad \mathcal{T}^+_{ext} = \mathcal{T}_{ext} \cup \{(l+1,2k),(l+1,2k+1)\} \backslash \{(l,k)\} \tag{S20}$$

where $(l,k) = \arg \max_{(l',k') \in \mathcal{T}_{ext} \backslash \mathcal{T}^*_{int}} (2^{l'} + k')$.

(ii) If $\mathcal{T} \not\subset \mathcal{T}^*$, the tree $\mathcal{T}$ contains redundant internal nodes. The mapping $\mathcal{G}(\cdot)$ again finds the deepest rightmost redundant node, say (l,k), and turns it into a bottom node. We have the same expression for $\mathcal{T}^- = \mathcal{G}(\mathcal{T})$ as in (S19).

Definition S1. For $\mathcal{T}' \in \mathbb{T}_L$ let $\bar{T}_{\mathcal{T},\mathcal{T}^*}$ denote the reverse of a path $T_{\mathcal{T},\mathcal{T}^*}$. The Bayesian CART canonical path ensemble is defined as $\mathcal{E} = \{T_{\mathcal{T},\mathcal{T}'} : (\mathcal{T},\mathcal{T}') \in \mathbb{T}_L \times \mathbb{T}_L\}$, where for each canonical path $T_{\mathcal{T},\mathcal{T}'}$ is obtained by collapsing the paths $T_{\mathcal{T},\mathcal{T}^*}$ and $\bar{T}_{\mathcal{T}',\mathcal{T}^*}$, i.e. $T_{\mathcal{T},\mathcal{T}'} = T_{\mathcal{T}\backslash\mathcal{T}'} \cup \bar{T}_{\mathcal{T}'\backslash\mathcal{T}}$, where $T_{\mathcal{T}\backslash\mathcal{T}'} := T_{\mathcal{T},\mathcal{T}^*} \backslash (T_{\mathcal{T},\mathcal{T}^*} \cap T_{\mathcal{T}',\mathcal{T}^*})$[3].

Below, we characterize important properties of $\mathcal{E}$ which are instrumental in the sandwich relation (19) and in the proof of Theorem 5.2.

Lemma S2. *Let $\mathcal{E}$ be the canonical path ensemble for Bayesian CART and let $|T_{\mathcal{T},\mathcal{T}'}|$ denote the length of the path $T_{\mathcal{T},\mathcal{T}'} \in \mathcal{E}$ between $\mathcal{T},\mathcal{T}' \in \mathbb{T}_L$. For $\mathcal{T}^*$ defined in Assumption 1, we have $\ell(\mathcal{E}) \equiv \max_{\mathcal{T},\mathcal{T}' \in \mathbb{T}_L} |T_{\mathcal{T},\mathcal{T}'}| \leq 2^{L+1}$.*

[3] By construction each step in $T_{\mathcal{T},\mathcal{T}^*}$ reduces the Hamming distance, and thus we can show similarly to [60] that $\mathcal{E}$ is an ensemble of *simple* paths.

Proof. See Section S4.3.

The following lemma characterizes the behavior of the congestion parameter $\rho(\mathcal{E})$ for the canonical ensemble $\mathcal{E}$ constructed above.

Lemma S3. *Assume the model (1) with the Bayesian CART prior with $p_{lk} = n^{-c}$ with $c > 1$. Under Assumption 1 (a), the canonical path ensemble $\mathcal{E}$ for the Bayesian CART algorithm from Section 2.1.2 satisfies $\rho(\mathcal{E}) \leq 2^{L+1}[1 + o(1)]$ with probability at least $1 - 4/n$.*

Proof. See Section S4.4.

S4.2. Proof of Theorem 5.2

We start with the sandwich relation (19) and find a lower bound to $\min_{\mathcal{T}\in\mathbb{T}} \Pi(\mathcal{T} \mid Y)$. By consistency established in Theorem 2.2, for any $\mathcal{T} \in \mathbb{T}$ we have (with probability at least $1 - 4/n$)

$$\Pi(\mathcal{T} \mid Y) = \Pi(\mathcal{T}^* \mid Y)\frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}^* \mid Y)} \geq \frac{1}{2}\frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}^* \mid Y)}.$$

We will again split the eligible trees $\mathbb{T}$ into overfitted $\mathbb{T}_O$ and underfitted $\mathbb{T}_U$. For any tree $\mathcal{T} \in \mathbb{T}_O$ with K extra internal nodes, we know from Section S2.0.1 that (using the shorthand notation $p_l = p_{lk}$)

$$\frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}^* \mid Y)} = (1+n)^{-K/2} \prod_{j=1}^{K} \left(\frac{p_{l_j}}{1-p_{l_j}} \times (1-p_{l_j+1})^2 \right) \times \exp\left\{ \frac{|X'_{[j]}Y|^2}{2(n+1)} \right\}. \tag{S21}$$

With $p_l = n^{-c} < 1/2$ we obtain

$$\min_{\mathcal{T}\in\mathbb{T}_O} \Pi(\mathcal{T} \mid Y) > \frac{1}{2} \min_{\mathcal{T}\in\mathbb{T}_O} \frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}^* \mid Y)} > \frac{1}{2}\left(\frac{1}{2n^c\sqrt{1+n}} \right)^K > \frac{e^{-\frac{n}{2}[\log 2 + (c+1/2)\log(1+n)]}}{2}. \tag{S22}$$

Similarly as in Section S4.4.2, we consider two under-fitted cases $\mathcal{T} \in \mathbb{T}_U$. First, assume that $\mathcal{T} \in \mathbb{T}_U$ and at the same time $\mathcal{T} \subset \mathcal{T}^*$. This means that $\mathcal{T}$ misses at least one signal node, e.g. $(l_S, k_S) \in \mathcal{B}(A)$. We denote with $\mathcal{T}^0 = \mathcal{T} \to \mathcal{T}^1 \to \cdots \to \mathcal{T}^K = \mathcal{T}^+$ the sequence of nested trees obtained by adding one additional internal node (l_j, k_j) towards (l_S, k_S). As in Section S2.0.2, we find that

$$\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)} = (1+n)^{K/2} \prod_{j=1}^{K} \exp\left\{ -\frac{|X'_{[j]}Y|^2}{2(n+1)} \right\}, \tag{S23}$$

We have

$$|X'_{[j]}Y|^2 = |X'_{|j|}(\boldsymbol{X}_{\mathcal{T}^*}\boldsymbol{\beta}^*_{\mathcal{T}^*} + \boldsymbol{\varepsilon})|^2 \leq 2n^2|\beta^*_{l_j k_j}|^2 + 8n\,\log n$$

and thereby

$$\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)} > (1+n)^{K/2}\exp\left\{-\frac{n^2\,\|\boldsymbol{\beta}^*_{\mathcal{T}^+\setminus\mathcal{T}}\|^2}{n+1} - \frac{4n\,K\log n}{n+1}\right\}$$

If $\mathcal{T}^+ = \mathcal{T}^*$ we stop tree growing, otherwise we repeat the same process with $\mathcal{T}^+$, extending a branch towards missing signal to create $\mathcal{T}^{++}$. We stop after M steps where $\mathcal{T}^{+\cdots+} = \mathcal{T}^*$ We then bound

$$\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^*)} = \frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)} \times \frac{N_Y(\mathcal{T}^+)}{N_Y(\mathcal{T}^{++})} \times \cdots \times \frac{N_Y(\mathcal{T}^{+\cdots+})}{N_Y(\mathcal{T}^*)} \tag{S24}$$

$$> (1+n)^{1/2}\exp\left\{-\frac{n^2\,\|\boldsymbol{\beta}^*_{\mathcal{T}^*}\|^2}{n+1} - \frac{4n\,|\mathcal{T}^*_{int}|\log n}{n+1}\right\}. \tag{S25}$$

The prior ratio satisfies (with $p_l = n^{-c} < 1/2$)

$$\frac{\Pi(\mathcal{T})}{\Pi(\mathcal{T}^+)} = \frac{1-p_{l_1}}{p_{l_1}} \times \left(\prod_{j=2}^{K-1}\frac{1}{p_{l_j}(1-p_{l_j})}\right) \times \frac{1}{(1-p_{l_K})^2} > n^{c(K-1)} \tag{S26}$$

This yields

$$\min_{\mathcal{T}\in\mathbb{T}_U:\mathcal{T}\subset\mathcal{T}^*} \Pi(\mathcal{T}\mid Y) > (1+n)^{1/2}\exp\left\{-\frac{n^2\,\|\boldsymbol{\beta}^*_{\mathcal{T}^*}\|^2}{n+1} - \frac{4n\,|\mathcal{T}^*_{int}|\log n}{n+1}\right\}. \tag{S27}$$

Now we focus on the under-fitted trees that are not necessarily contained inside $\mathcal{T}^*$. Consider $\mathcal{T} \in \mathbb{T}_U$ such that $\mathcal{T} \not\subset \mathcal{T}^*$. Then we combine growing and pruning operations from the previous steps. First, we prune the tree $\mathcal{T}$ into the largest tree $\mathcal{T}_U$ that underfits, i.e. $\mathcal{T}_U$ is the largest tree such that $\mathcal{T}_U \in \mathbb{T}_U$ and $\mathcal{T}_U \subset \mathcal{T}$, and write

$$\frac{\Pi(\mathcal{T}\mid Y)}{\Pi(\mathcal{T}^*\mid Y)} = \frac{\Pi(\mathcal{T}\mid Y)}{\Pi(\mathcal{T}^U\mid Y)}\frac{\Pi(\mathcal{T}^U\mid Y)}{\Pi(\mathcal{T}^*\mid Y)}$$

and combining the expression (S22) with (S27) we find

$$\min_{\mathcal{T}\in\mathbb{T}_U:\mathcal{T}\not\subset\mathcal{T}^*}\frac{\Pi(\mathcal{T}\mid Y)}{\Pi(\mathcal{T}^*\mid Y)} > \min_{\mathcal{T}\in\mathbb{T}_U:\mathcal{T}\not\subset\mathcal{T}^*}\frac{\Pi(\mathcal{T}\mid Y)}{\Pi(\mathcal{T}_U\mid Y)} \times \min_{\mathcal{T}\in\mathbb{T}_U:\mathcal{T}\subset\mathcal{T}^*}\frac{\Pi(\mathcal{T}\mid Y)}{\Pi(\mathcal{T}^*\mid Y)}$$

$$> \frac{1}{2}\exp\left\{-n\left[1+\left(c+\frac{1}{2}\right)\log(1+n)\right] - \frac{n^2\,|\mathcal{T}^*_{int}|C^2_{f_0}}{n+1} - \frac{4n\,|\mathcal{T}^*_{int}|\log n}{n+1}\right\}$$

Now, by Lemma S2 we have $l(\mathcal{E}) \leq 2^{L+1}$ and by Lemma S3 we have $\rho(\mathcal{E}) \leq 2^{L+1}[1 + o(1)]$ for $c > 1$ on the event space $\mathcal{A}_n$. Plugging these into (19), we obtain

$$\begin{aligned}\tau_\epsilon &\leq l(\mathcal{E})\rho(\mathcal{E})\Big(\log\left[\frac{1}{\min_{\mathcal{T}\in\mathbb{T}}\Pi(\mathcal{T}\mid Y)}\right] + \log(1/\varepsilon)\Big)\\ &\leq 2^{2(L+1)+1}\left\{n\left[\left(c+\frac{1}{2}\right)\log(1+n) + |\mathcal{T}^*_{int}|C^2_{f_0}+1\right] + 4\,|\mathcal{T}^*_{int}|\log n + \log\left(\frac{2}{\epsilon}\right)\right\}.\end{aligned}$$

S4.3. Proof of Lemma S2

We want to upper bound the length of the longest canonical path constructed in Section S4.1. Let us first bound $|T_{\mathcal{T},\mathcal{T}^*}|$ when $\mathcal{T} \supset \mathcal{T}^*$. In order to reach $\mathcal{T}^*$ from $\mathcal{T}$ on a canonical path, we remove one redundant node at a time. There are at most 2^L nodes of which $(2^L - |\mathcal{T}^*_{int}|)$ are redundant. Thereby, we have $\max_{\mathcal{T}:\mathcal{T}\supset\mathcal{T}^*}\{|T_{\mathcal{T},\mathcal{T}^*}|\} \leq (2^L - |\mathcal{T}^*_{int}|)$. Conversely, for any $\mathcal{T} \subset \mathcal{T}^*$, the canonical path from $\mathcal{T}$ towards $\mathcal{T}^*$ adds one node in $\mathcal{T}^*_{int}\backslash\mathcal{T}_{int}$ at a time. This means $\max_{\mathcal{T}:\mathcal{T}\subset\mathcal{T}^*}|T_{\mathcal{T},\mathcal{T}^*}| \leq |\mathcal{T}^*_{int}|$. When $\mathcal{T} \not\subset \mathcal{T}^*$ and $\mathcal{T} \not\supset \mathcal{T}^*$, the path from $\mathcal{T}$ towards $\mathcal{T}^*$ follows by first deleting redundant nodes and then adding nodes towards reaching $\mathcal{T}^*$. This can be achieved in at most $(2^L - |\mathcal{T}^*_{int}| + |\mathcal{T}^*_{int}|)$ steps. Finally, for any two trees $\mathcal{T}, \mathcal{T}' \in \mathbb{T}$ the canonical path $T_{\mathcal{T},\mathcal{T}'}$ is obtained by collapsing $T_{\mathcal{T},\mathcal{T}*}$ and $\bar{T}_{\mathcal{T}',\mathcal{T}^*}$. Thereby, we have $\max_{\mathcal{T},\mathcal{T}'\in\mathbb{T}}|T_{\mathcal{T},\mathcal{T}'}| \leq 2^{L+1}$.

S4.4. Proof of Lemma S3

We will work conditionally on the set $\mathcal{A}_n$ defined in (S1), where $p = 2^{L_{max}} = n/2$. We know that the complement of this set has a vanishing probability $\mathbb{P}(\mathcal{A}^c_n) \leq 2/p \to 0$. We denote by $T_{\mathcal{T},\mathcal{T}'} \in \mathcal{E}$ a canonical path between two nodes $\mathcal{T}, \mathcal{T}' \in \mathbb{T}$. We will find an upper bound for the congestion parameter $\rho(\mathcal{E})$ defined in (21) as

$$\rho(\mathcal{E}) = \max_{e\in\mathcal{E}}\frac{1}{Q(e)}\sum_{(\bar{\mathcal{T}},\bar{\mathcal{T}}'):e\in T_{\bar{\mathcal{T}},\bar{\mathcal{T}}'}}\Pi(\bar{\mathcal{T}}\mid Y)\Pi(\bar{\mathcal{T}}'\mid Y),$$

where for an edge e between $(\mathcal{T}, \mathcal{T}')$ we have

$$Q(e) \equiv Q(\mathcal{T},\mathcal{T}') = \Pi(\mathcal{T}\mid Y)P(\mathcal{T},\mathcal{T}').$$

First, we denote with

$$\Delta(\mathcal{T}') = \{\mathcal{T} : \mathcal{T}' \in T_{\mathcal{T},\mathcal{T}^*}\} \tag{S28}$$

a set of precedents of a tree $\mathcal{T}'$ that lie on a canonical path towards $\mathcal{T}^*$. Note that $\mathcal{T}' \in \Delta(\mathcal{T}')$. For any given edge $e_{\mathcal{T},\mathcal{T}'} = T_{\mathcal{T},\mathcal{T}'}$ between two *adjacent* trees $\mathcal{T}$ and $\mathcal{T}'$ where $\mathcal{T} \in \Delta(\mathcal{T}')$, we have

$$N(e) \equiv \{(\bar{\mathcal{T}},\bar{\mathcal{T}}') \mid e \in T_{\bar{\mathcal{T}},\bar{\mathcal{T}}'}\} \subset \Delta(\mathcal{T}) \times \mathbb{T}.$$

Then we can find an upper bound for the congestion parameter in (21) as

$$\rho(\mathcal{E}) \leq \max_{e_{\mathcal{T},\mathcal{T}'} \in \mathcal{E}} \frac{\Pi[\Delta(\mathcal{T})]}{Q(e_{\mathcal{T},\mathcal{T}'})} \leq \max_{(\mathcal{T},\mathcal{T}') \in \Gamma^*} \frac{\Pi[\Delta(\mathcal{T})]}{Q(\mathcal{T},\mathcal{T}')}, \tag{S29}$$

where

$$\Gamma^* = \{(\mathcal{T},\mathcal{T}') \in \mathbb{T} \times \mathbb{T} \mid e_{\mathcal{T},\mathcal{T}'} = T_{\mathcal{T},\mathcal{T}'} \quad \text{and} \quad \mathcal{T} \in \Delta(\mathcal{T}')\}$$

Now we find a lower bound for $Q(\mathcal{T},\mathcal{T}')$ for an *adjacent* pair $(\mathcal{T},\mathcal{T}')$ such that $\mathcal{T} \in \Delta(\mathcal{T}')$ or, equivalently, for $\mathcal{T}' = \mathcal{G}(\mathcal{T})$, where $\mathcal{G}(\cdot)$ is the mapping introduced in Section S4.1. For the "lazy" walk explained in Section 4 with a transition matrix $P = \widetilde{P}/2 + I/2$ where $\widetilde{P}$ is the original transition matrix, we have

$$Q(\mathcal{T},\mathcal{T}') = \frac{1}{2}\Pi(\mathcal{T} \mid Y)\widetilde{P}(\mathcal{T},\mathcal{T}') = \frac{1}{2}\Pi(\mathcal{T} \mid Y)S(\mathcal{T} \to \mathcal{T}') \min\left\{1, \frac{\Pi(\mathcal{T}' \mid Y)S(\mathcal{T}' \to \mathcal{T})}{\Pi(\mathcal{T} \mid Y)S(\mathcal{T} \to \mathcal{T}')}\right\}.$$

Plugging this into (S29) we obtain

$$\rho(\mathcal{E}) \leq 2 \max_{(\mathcal{T},\mathcal{T}') \in \Gamma^*} \left\{ \frac{\Pi[\Delta(\mathcal{T})]}{\Pi(\mathcal{T} \mid Y)S(\mathcal{T} \to \mathcal{T}')} \times \max\left[1, \frac{\Pi(\mathcal{T} \mid Y)S(\mathcal{T} \to \mathcal{T}')}{\Pi(\mathcal{T}' \mid Y)S(\mathcal{T}' \to \mathcal{T})}\right] \right\}. \tag{S30}$$

We now bound the ratio $\rho(\mathcal{E})$ assuming that $\mathcal{T}$ is either *overfitting* or *underfitting*. We continue using the notation $\mathbb{T}_O = \{\mathcal{T} : \mathcal{T} \supset \mathcal{T}^*\}$ and $\mathbb{T}_U = \{\mathcal{T} : \mathcal{T} \not\supset \mathcal{T}^*\}$.

S4.4.1. When $\mathcal{T}^* \subset \mathcal{T}$ (The Overfitted Case)

When $\mathcal{T} \in \mathbb{T}_O$ subsumes the tree $\mathcal{T}^*$, the mapping $\mathcal{G}(\cdot)$ picks the deepest rightmost internal node, say $(l_S, k_S) \in \mathcal{T}_{int} \backslash \mathcal{T}^*_{int}$, and turns it into a bottom node. We denote with $\mathcal{T}^- = \mathcal{G}(\mathcal{T})$

such a pruned tree. We also define the collection of *pre-terminal nodes* of a tree $\mathcal{T} \in \mathbb{T}$ as

$$\mathcal{P}(\mathcal{T}) = \{(l,k) \in \mathcal{T}_{int} : \{(l+1,2k),(l+1,2k+1)\} \in \mathcal{T}_{ext}\},$$

i.e. these are internal nodes whose children are the bottom nodes. Using the posterior ratio expression in (S3) and (S4) for overfitted trees with $K = 1$ we obtain (for $j = 2^{l_S} + k_S + 1$)

$$\frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}^- \mid Y)} = (1+n)^{-1/2} \frac{p_{l_S}}{1-p_{l_S}} \times (1-p_{l_S+1})^2 \times \exp\left\{\frac{|X'_{[j]}Y|^2}{2(n+1)}\right\} \leq \mathrm{e}^{-(c-3/2)\log n}$$

Since we cannot preclude that $\mathcal{T} = \mathcal{T}^L_{full}$, the proposal ratio satisfies

$$\frac{S(\mathcal{T} \to \mathcal{T}^-)}{S(\mathcal{T}^- \to \mathcal{T})} \leq \frac{2|\mathcal{T}^-_{ext}|}{|\mathcal{P}(\mathcal{T})|} < 4.$$

Then the ratio inside the Metropolis-Hastings acceptance probability in (S30) satisfies

$$\frac{\Pi(\mathcal{T} \mid Y) S(\mathcal{T} \to \mathcal{T}^-)}{\Pi(\mathcal{T}^- \mid Y) S(\mathcal{T}^- \to \mathcal{T})} \leq 4\mathrm{e}^{-(c-3/2)\log n} = o(1) \quad \text{for } c > 3/2. \tag{S31}$$

We now focus on the second ratio in the product in (S30). When $\mathcal{T}^* \subset \mathcal{T}$, all precedents $\mathcal{T}' \in \Delta(\mathcal{T})$ (recall the definition of $\Delta(\mathcal{T})$ in (S28)) are also *overfitted* models, i.e. $\mathcal{T}^* \subset \mathcal{T}'$ and $\mathcal{T} \subset \mathcal{T}'$ for all $\mathcal{T}' \in \Delta(\mathcal{T}) \backslash \{\mathcal{T}\}$. We decompose $\Delta(\mathcal{T}) = \cup_{K=1}^{2^L} \Delta(\mathcal{T}, K)$ into shells depending how many steps away each tree $\mathcal{T}' \in \Delta(\mathcal{T})$ is on a canonical path towards $\mathcal{T}$. Namely, for $K \in \mathbb{N}$, we denote with

$$\Delta(\mathcal{T}, K) = \{\mathcal{T}' \in \Delta(\mathcal{T}) : |T_{\mathcal{T}',\mathcal{T}}| = K\} \tag{S32}$$

the set of precedents that are K steps away from $\mathcal{T}$ on some canonical path. Using again the posterior ratio for overfitted models in (S3) and (S4), we obtain for $\mathcal{T}' \in \Delta(\mathcal{T}, K)$

$$\frac{\Pi(\mathcal{T}' \mid Y)}{\Pi(\mathcal{T} \mid Y)} \leq \mathrm{e}^{-K(c-3/2)\log n}.$$

Moreover, the cardinality of $\Delta(\mathcal{T}, K)$ for $K \geq 1$ satisfies $\mathrm{card}[\Delta(\mathcal{T}, K))] \leq \prod_{j=1}^{K} (|\mathcal{T}_{ext}| + j - 1)$ and $\mathrm{card}[\Delta(\mathcal{T}, 0))] = 1$. Thereby

$$\begin{aligned} \frac{\Pi[\Delta(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)} &= 1 + \sum_{K=1}^{2^L} \sum_{\mathcal{T}' \in \Delta(\mathcal{T},K)} \frac{\Pi(\mathcal{T}' \mid Y)}{\Pi(\mathcal{T} \mid Y)} \leq 1 + \sum_{K=1}^{2^L} \mathrm{e}^{K \log[(|\mathcal{T}_{ext}|+K-1)]} \mathrm{e}^{-(c-3/2)K\log n} \\ &< 1 + \frac{1}{n^{(c-5/2)} - 1}. \end{aligned} \tag{S33}$$

Finally, because $\mathcal{T} \neq \mathcal{T}_{null}$, we have from (12)

$$S(\mathcal{T} \to \mathcal{T}^{-}) = \frac{1}{2}\frac{1}{|\mathcal{P}(\mathcal{T})|}.$$

Then we obtain

$$\frac{\Pi[\Delta(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)S(\mathcal{T} \to \mathcal{T}')} \times \max\left[1, \frac{\Pi(\mathcal{T} \mid Y)S(\mathcal{T} \to \mathcal{T}')}{\Pi(\mathcal{T}' \mid Y)S(\mathcal{T}' \to \mathcal{T})}\right] \leq 2|\mathcal{P}(\mathcal{T})|\left(1 + \frac{1}{n^{(c-5/2)} - 1}\right).$$

S4.4.2. When $\mathcal{T}^* \not\subset \mathcal{T}$ (The Underfitted Case)

We consider two cases of underfitting: (1) when $\mathcal{T} \not\subset \mathcal{T}^*$ and, at the same time, $\mathcal{T} \not\supset \mathcal{T}^*$ and (2) when $\mathcal{T} \subset \mathcal{T}^*$. First, if the tree underfits and contains extra nodes, those are deleted first which coincides with the previous case.

We now focus on the second case when $\mathcal{T} \subset \mathcal{T}^*$. Then $\mathcal{G}(\mathcal{T})$ proceeds by adding an additional node towards completing $\mathcal{T}^*$. We denote the resulting enlarged tree by $\mathcal{T}^+ = \mathcal{G}(\mathcal{T})$. Using the expression of posterior ratio in (S10) and (S13) with $K = 1$ we find that

$$\frac{S(\mathcal{T} \to \mathcal{T}^+)}{S(\mathcal{T}^+ \to \mathcal{T})}\frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}^+ \mid Y)} \leq \frac{2|\mathcal{P}(\mathcal{T}^+)|}{|\mathcal{T}_{ext}|}\mathrm{e}^{-(A^2/8)\log^2 n} = o(1).$$

Next, note that precedents $\Delta(\mathcal{T})$ in (S28) of an underfitted model $\mathcal{T}$ are also underfitted models and can be divided into two (besides the singleton set $\{\mathcal{T}\}$) mutually exclusive categories, i.e. $\Delta(\mathcal{T}) = \{\mathcal{T}\} \cup \mathcal{U}_1(\mathcal{T}) \cup \mathcal{U}_2(\mathcal{T})$. The first set, denoted with $\mathcal{U}_1(\mathcal{T})$, consists of all precedents $\Delta(\mathcal{T})$ that are also subsets of $\mathcal{T}^*$, i.e. $\mathcal{U}_1(\mathcal{T}) \equiv \{\mathcal{T}' \in \Delta(\mathcal{T}) : \mathcal{T}' \subset \mathcal{T}^*\}$. The second set, denoted with $\mathcal{U}_2(\mathcal{T})$, are all the precedents that have some redundant nodes and are *not* included in $\mathcal{T}^*$, i.e. $\mathcal{U}_2(\mathcal{T}) = \{\mathcal{T}' \in \Delta(\mathcal{T}) : \mathcal{T}' \not\subseteq \mathcal{T}^*\}$. We denote with $\Delta(\mathcal{T}, K) \subset \Delta(\mathcal{T})$ those precedents that are K steps away from $\mathcal{T}$ on a canonical path (i.e. all trees inside $\mathcal{U}_1(\mathcal{T})$ that have K fewer internal nodes compared to $\mathcal{T}$ and all trees inside $\mathcal{U}_2(\mathcal{T})$ that have K extra internal nodes compared to $\mathcal{T}$), where the cardinality satisfies $\mathrm{card}[\Delta(\mathcal{T}, K)] \leq 2^{LK}$. Because under the Assumption 1 (a), *all* internal nodes in $\mathcal{T}^*$ are signals, we can modify the

expressions in (S10) to include all K signals (not just one) to obtain for large enough A

$$\frac{\Pi[\mathcal{U}_1(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)} \leq \sum_{K=1}^{|\mathcal{T}^*_{int}|} \sum_{\mathcal{T}' \in \Delta(\mathcal{T},K) \cap \mathcal{U}_1(\mathcal{T})} \frac{\Pi(\mathcal{T}' \mid Y)}{\Pi(\mathcal{T} \mid Y)} < \sum_{K=1}^{|\mathcal{T}^*_{int}|} \mathrm{e}^{-K(A^2/8)\log^2 n} < \frac{1}{n^{(A^2/8)\log n} - 1}. \tag{S34}$$

We now consider the second type of underfitting precedents $\mathcal{U}_2(\mathcal{T})$. For each such $\mathcal{T}' \in \mathcal{U}_2(\mathcal{T})$, the canonical path $T_{\mathcal{T}',\mathcal{T}}$ first proceeds by removing redundant nodes and at some point reaches a tree $U(\mathcal{T}')$ which already underfits. In other words, $U(\mathcal{T}') \in \mathcal{U}_1$ is defined as the largest subtree obtained from $\mathcal{T}'$ by removing all redundant branches (without signal). This means that $U(\mathcal{T}')$ is the largest tree that satisfies $U(\mathcal{T}') \subset \mathcal{T}'$ and, at the same time, $U(\mathcal{T}') \subset \mathcal{T}^*$. The mapping $\mathcal{T}' \to U(\mathcal{T}')$ is many-to-one and for any $\widetilde{\mathcal{T}} \in \mathcal{U}_1(\mathcal{T})$ such that there exists $\mathcal{T}' \in \mathcal{U}_2(\mathcal{T})$ so that $U(\mathcal{T}') = \widetilde{\mathcal{T}}$ we have

$$\mathcal{N}(\mathcal{T}, \widetilde{\mathcal{T}}) \equiv \{\mathcal{T}' \in \mathcal{U}_2(\mathcal{T}) : U(\mathcal{T}') = \widetilde{\mathcal{T}}\} \subseteq \mathbb{T}_O(\widetilde{\mathcal{T}}),$$

where $\mathbb{T}_O(\widetilde{\mathcal{T}}) = \{\mathcal{T} : \widetilde{\mathcal{T}} \subset \mathcal{T}\}$ are all trees that contain $\widetilde{\mathcal{T}}$. Using the same logic as in (S5) and (S6) we find that

$$\frac{\Pi[\mathcal{N}(\mathcal{T}, \widetilde{\mathcal{T}}) \mid Y]}{\Pi(\widetilde{\mathcal{T}} \mid Y)} \leq \frac{1}{n^{c-5/2} - 1}$$

and thereby using (S34)

$$\begin{aligned}
\frac{\Pi[\mathcal{U}_2(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)} &= \sum_{\mathcal{T}' \in \mathcal{U}_2(\mathcal{T})} \frac{\Pi[U(\mathcal{T}') \mid Y]}{\Pi(\mathcal{T} \mid Y)} \frac{\Pi(\mathcal{T}' \mid Y)}{\Pi[U(\mathcal{T}') \mid Y]} \\
&= \sum_{\substack{\widetilde{\mathcal{T}} \in U_1(\mathcal{T}): \\ \mathcal{N}(\mathcal{T}, \widetilde{\mathcal{T}}) \neq \emptyset}} \sum_{\mathcal{T}' \in \mathcal{N}(\mathcal{T}, \widetilde{\mathcal{T}})} \frac{\Pi(\widetilde{\mathcal{T}} \mid Y)}{\Pi(\mathcal{T} \mid Y)} \frac{\Pi(\mathcal{T}' \mid Y)}{\Pi(\widetilde{\mathcal{T}} \mid Y)} \\
&\leq \sum_{\substack{\widetilde{\mathcal{T}} \in U_1(\mathcal{T}): \\ \mathcal{N}(\mathcal{T}, \widetilde{\mathcal{T}}) \neq \emptyset}} \frac{\Pi(\widetilde{\mathcal{T}} \mid Y)}{\Pi(\mathcal{T} \mid Y)} \sum_{\mathcal{T}' \in \mathbb{T}_O(\widetilde{\mathcal{T}})} \frac{\Pi(\mathcal{T}' \mid Y)}{\Pi(\widetilde{\mathcal{T}} \mid Y)} \\
&\leq \frac{1}{n^{c-5/2} - 1} \sum_{\substack{\widetilde{\mathcal{T}} \in U_1(\mathcal{T}): \\ \mathcal{N}(\mathcal{T}, \widetilde{\mathcal{T}}) \neq \emptyset}} \frac{\Pi(\widetilde{\mathcal{T}} \mid Y)}{\Pi(\mathcal{T} \mid Y)} \\
&\leq \frac{1}{n^{c-5/2} - 1} \frac{\Pi[\mathcal{U}_1(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)} < \frac{1}{n^{c-5/2} - 1} \times \frac{1}{n^{(A^2/8)\log n} - 1}
\end{aligned}$$

Putting it all together, we have

$$\frac{\Pi[\Delta(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y) S(\mathcal{T} \to \mathcal{T}^+)} \leq 2|\mathcal{T}_{ext}| \, (1 + o(1)) \, .$$

The bound for second underfitting case (b) when $\mathcal{T} \not\subset \mathcal{T}^*$ and, at the same time, $\mathcal{T} \not\supset \mathcal{T}^*$ proceeds analogously, only without the set $\mathcal{U}_1(\mathcal{T})$ that is empty.

Putting it all together, and noting that $|\mathcal{P}(\mathcal{T})| \leq 2^L$ and $|\mathcal{T}_{ext}| \leq 2^L$, the bound in (S30) yields $\rho \leq 2^{L+1}(1 + o(1))$ for $c > 5/2$.

S5. Proof of Theorem 5.3 (Twiggy Bayesian CART Mixing Upper Bound)

We follow the same recipe as in the proof of Theorem 5.2. We first need to show Lemma S2 and Lemma S3 for the canonical path ensemble for Twiggy Bayesian CART constructed in Section S5.1 below.

S5.1. Canonical Path Ensemble for Twiggy Bayesian CART

We again construct a canonical path $T_{\mathcal{T},\mathcal{T}^*}$ between any $\mathcal{T} \in \mathbb{T} \backslash \mathcal{T}^*$ and the spanning tree $\mathcal{T}^*$ from Assumption 1. Recall the definition of signals $\mathcal{B}(A) = \{(l,k) : C_{f_0} > |\beta^*_{lk}| > A \log n/\sqrt{n}\}$, where $\mathcal{T}^* = \mathcal{B}(A)$ under Assumption 1 (a) and $\mathcal{B}(A) \subseteq \mathcal{T}^*$ Assumption 1 (b). The transition function $\mathcal{G}(\mathcal{T})$ for Twiggy Bayesian CART is defined as follows:

(1) Assume $\mathcal{T} \supset \mathcal{T}^*$ is **overfitted**, i.e. $\mathcal{T}$ forms an envelope around $\mathcal{T}^*$ and contains at least one redundant node. Denote the set of all redundant internal nodes whose descendants form a twig as

$$S(\mathcal{T}) \equiv \left\{(l,k) \in \mathcal{T}_{int} \backslash \mathcal{T}^*_{int} : \exists (l^*,k^*) \in \mathcal{T}_{int} \; s.t. \; D_{lk}(\mathcal{T}) = [(l,k) \leftrightarrow (l^*,k^*)]\right\} .$$

Note that this set contains all pre-terminal nodes. The mapping $\mathcal{G}(\cdot)$ finds the most shallow leftmost node inside $S(\mathcal{T})$, say $(\widetilde{l},\widetilde{k})$, and turns it into a bottom node with the twig below removed. More formally, we define $\mathcal{G}(\mathcal{T}) = \mathcal{T}^-$ as

$$\mathcal{T}^-_{int} = \mathcal{T}_{int} \setminus [(\widetilde{l},\widetilde{k}) \leftrightarrow (l^*,k^*)] \quad \text{where} \quad (\widetilde{l},\widetilde{k}) = \arg \min_{(l,k) \in S(\mathcal{T})} (2^l + k). \tag{S35}$$

Picking the shallowest (as opposed to deepest) node for removal gives us an opportunity to remove more than one node at a time, thereby shortening the path towards $\mathcal{T}^*$.

(2) Assume $\mathcal{T} \not\supseteq \mathcal{T}^*$ is **underfitted**, i.e. $\mathcal{T}$ misses at least one node in $\mathcal{T}^*$.

(i) If $\mathcal{T} \subset \mathcal{T}^*$, the mapping $\mathcal{G}(\cdot)$ finds the deepest rightmost node inside $\mathcal{T}^*$ missed by $\mathcal{T}$, say (l^*, k^*), and grows a twig towards it. More formally, we define $\mathcal{T}^+ = \mathcal{G}(\mathcal{T})$ where

$$\mathcal{T}^+_{int} = \mathcal{T}_{int} \cup [(\widetilde{l}, \widetilde{k}) \leftrightarrow (l^*, k^*)] \quad \text{where} \quad (l^*, k^*) = \arg \max_{(l,k) \in \mathcal{T}^*_{int} \setminus \mathcal{T}_{int}} (2^l + k)$$

and where $(\widetilde{l}, \widetilde{k})$ is the closest node to (l^*, k^*) inside $\mathcal{T}_{ext}$. Note that taking the deepest rightmost node gives us an opportunity to add more than one signal node at a time.

(ii) If $\mathcal{T} \not\subset \mathcal{T}^*$, i.e., $\mathcal{T}$ contains redundant nodes and the mapping $\mathcal{G}(\cdot)$ is the same as in the overfitting the case (1).

It is easy to see that this transition function reduces the Hamming distance after each step. Compared to Bayesian CART, however, it may take larger leaps. It can be shown that the canonical path ensemble for Twiggy Bayesian CART satisfies the statements of Lemma S2 and Lemma S3 for the unstructured signal Assumption 1 (b) (see proof of Theorem 5.3 in Section S5).

S5.2. Version of Lemma S2 for Twiggy Bayesian CART

The proof is similar to the Bayesian CART version. Let us first bound $|T_{\mathcal{T},\mathcal{T}^*}|$ when $\mathcal{T} \supset \mathcal{T}^*$. In order to reach $\mathcal{T}^*$ from $\mathcal{T}$ on a canonical path, we remove at least one redundant node at a time. There are at most 2^L nodes of which $(2^L - |\mathcal{T}^*_{int}|)$ are redundant. Thereby, we have $\max_{\mathcal{T}:\mathcal{T}\supset\mathcal{T}^*} \{|T_{\mathcal{T},\mathcal{T}^*}|\} \leq (2^L - |\mathcal{T}^*_{int}|)$. Using a more complicated argument, one could take advantage of removals of entire twigs to show that the removal can be achieved in up to $|\mathcal{P}(\mathcal{T}) \setminus \mathcal{T}^*_{int}|$ steps. Conversely, for any $\mathcal{T} \subset \mathcal{T}^*$, the canonical path from $\mathcal{T}$ towards $\mathcal{T}^*$ adds

a twig towards a node in $\mathcal{P}(\mathcal{T}^*)\backslash\mathcal{T}_{int}$ at a time. This means $\max_{\mathcal{T}:\mathcal{T}\subset\mathcal{T}^*} |T_{\mathcal{T},\mathcal{T}^*}| \leq |\mathcal{P}(\mathcal{T}^*)|$. When $\mathcal{T} \not\subset \mathcal{T}^*$ and $\mathcal{T} \not\supset \mathcal{T}^*$, the path from $\mathcal{T}$ towards $\mathcal{T}^*$ follows by first deleting redundant nodes and then adding nodes towards reaching $\mathcal{T}^*$. This can be achieved in at most $(2^L - |\mathcal{T}^*_{int}| + |\mathcal{P}(\mathcal{T})|)$ steps. Finally, for any two trees $\mathcal{T}, \mathcal{T}' \in \mathbb{T}$ the canonical path $T_{\mathcal{T},\mathcal{T}'}$ is obtained by collapsing $T_{\mathcal{T},\mathcal{T}_*}$ and $\bar{T}_{\mathcal{T}',\mathcal{T}^*}$. Thereby, we have $\max_{\mathcal{T},\mathcal{T}'\in\mathbb{T}} |T_{\mathcal{T},\mathcal{T}'}| \leq 2^{L+1}$. The bound can be sharpened to 2^L using a more complicated argument.

S5.3. Version of Lemma S3 for Twiggy Bayesian CART

We will again work on the event space $\mathcal{A}_n$ in (S1), which has probability at least $1 - 4/n$. The strategy is the same as in the proof of Lemma S3. We again split the considerations into overfitted and underfitted trees.

S5.3.1. When $\mathcal{T}^* \subset \mathcal{T}$ (The Overfitted Case)

When $\mathcal{T}$ subsumes the tree $\mathcal{T}^*$, the mapping $\mathcal{G}(\cdot)$ finds the shallowest leftmost node inside $\mathcal{T}_{int}\backslash\mathcal{T}^*_{int}$, say (l,k), such that the entire branch below (l,k) is a twig, and removes the twig, turning $(l,k) \in \mathcal{T}_{int}$ into an external node. In other words, (l,k) has been converted to a bottom node and its descendants erased. More formally, recall the definition of ancestors of (l,k) inside $\mathcal{T}$ as

$$A_{lk}(\mathcal{T}) = \{(l',k') \in \mathcal{T}_{int} : \exists j \in \{0,1,\ldots,L-1\}\ s.t.\ (l',k') = (l-j, \lfloor k/2^j \rfloor)\}$$

and descendants of (l,k) as

$$D_{lk}(\mathcal{T}) = \{(l',k') \in \mathcal{T}_{int} : (l,k) \in A_{l'k'}(\mathcal{T})\}.$$

Moreover, (l,k) is such that $\exists(\widetilde{l},\widetilde{k}) \in \mathcal{T}_{int}$ such that $D_{lk} = [(l,k) \leftrightarrow (\widetilde{l},\widetilde{k})]$. Writing $\mathcal{T}^- = \mathcal{G}(\mathcal{T})$, we have

$$\mathcal{T}^-_{int} = \mathcal{T}_{int} \setminus [(l,k) \leftrightarrow (\widetilde{l},\widetilde{k})].$$

We now provide bounds for the two terms in (S30). We denote the length (number of nodes) of the twig $[(l,k) \leftrightarrow (\widetilde{l},\widetilde{k})]$ by k. This means that $\mathcal{T}^{-}_{int}$ has k fewer internal nodes compared to $\mathcal{T}$ and from the construction all of them are signal-less nodes. With the proposal distribution described in Section 3.1, and since we cannot preclude that $\mathcal{T} = \mathcal{T}^{L}_{full}$, we have

$$\frac{S(\mathcal{T} \to \mathcal{T}^{-})}{S(\mathcal{T}^{-} \to \mathcal{T})} \leq \frac{2^{L}\frac{D^{L}-1}{D-1}}{|\mathcal{P}(\mathcal{T})|} \leq n\,\frac{D^{L}-1}{D-1}.$$

Using the posterior ratio expression in (S3) and (S4) for overfitted trees we obtain

$$\frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}^{-} \mid Y)}\frac{S(\mathcal{T} \to \mathcal{T}^{-})}{S(\mathcal{T}^{-} \to \mathcal{T})} \leq n\,\frac{D^{L}-1}{D-1}\mathrm{e}^{-k(c-3/2)\log n} \leq \frac{D^{L}-1}{D-1}\mathrm{e}^{-k(c-5/2)\log n}.$$

This means that the second maximum quantity in (S30) is no greater than a constant multiple of $\mathrm{e}^{-(c-5/2-\log D)\log n}$ which is $o(1)$ for $c > 5/2 + \log D$. We now focus on the first ratio in the product in (S30). When $\mathcal{T}^{*} \subset \mathcal{T}$, all precedents $\mathcal{T}' \in \Delta(\mathcal{T})$ (recall the definition of $\Delta(\mathcal{T})$ in (S28)) are also *overfitted* models, i.e. $\mathcal{T}^{*} \subset \mathcal{T}'$ and $\mathcal{T} \subset \mathcal{T}'$ for all $\mathcal{T}' \in \Delta(\mathcal{T})\backslash\{\mathcal{T}'\}$. Similarly as in the proof of Lemma S3 in Section S4.4, we decompose $\Delta(\mathcal{T}) = \cup_{K=1}^{2L}\Delta(\mathcal{T},K)$ into shells $\Delta(\mathcal{T},K) = \{\mathcal{T}' \in \Delta(\mathcal{T}) : |T_{\mathcal{T}',\mathcal{T}}| = K\}$ defined as in (S32). The difference now is that each tree $\mathcal{T}' \in \Delta(\mathcal{T},K)$ can have *more than K redundant nodes*. We denote with $\boldsymbol{\kappa} = (k(1),\ldots,k(K))' \in (\mathbb{N}\backslash\{0\})^{K}$ the vector of numbers of redundant nodes deleted at each of the K steps on the canonical path from $\mathcal{T}' \in \Delta(\mathcal{T},K)$ towards $\mathcal{T}^{*}$. Using again the posterior ratio for overfitted models in (S3) and (S4), we now obtain for $\mathcal{T}' \in \Delta(\mathcal{T},K)$

$$\frac{\Pi(\mathcal{T}' \mid Y)}{\Pi(\mathcal{T} \mid Y)} \leq \mathrm{e}^{(3/2-c)\sum_{j=1}^{K}k(j)\log n}.$$

Moreover, we define

$$\Delta(\mathcal{T},K,\boldsymbol{\kappa}) = \{\mathcal{T}' \in \Delta(\mathcal{T},K) : |\mathcal{T}^{j}_{int}\backslash\mathcal{T}^{j-1}_{int}| = k(j),\ \forall j = 1,\ldots,K\}$$

all the trees that are K steps away from $\mathcal{T}$ and that differ from $\mathcal{T}$ by adding exactly $k(j)$ nodes at each step. When $K = 1$, the number of such precedents is at most the number of binary trees with $k(1)$ internal nodes. This corresponds to the Catalan number $\mathbb{C}_{K}$, which according to Lemma S-3 in [9], satisfies $\mathbb{C}_{k(1)} \asymp 4^{k(1)}/k(1)^{3/2}$. Then it is easy to see that for $K \geq 1$ we

have

$$\text{card}[\Delta(\mathcal{T}, K, \boldsymbol{\kappa})] \leq \prod_{j=1}^{K} \mathrm{e}^{k(j)\log 4 - 3/2 \log k(j)}.$$

This yields (since $K \leq \sum_j k(j) \leq 2^L$ and $c > 5/2$) for $n \geq 8$

$$\begin{aligned}
\frac{\Pi[\Delta(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)} &= 1 + \sum_{K=1}^{2^L} \sum_{\boldsymbol{\kappa}: \sum_j k(j) \leq 2^L} \sum_{\mathcal{T}' \in \Delta(\mathcal{T}, K, \boldsymbol{\kappa})} \frac{\Pi(\mathcal{T}' \mid Y)}{\Pi(\mathcal{T} \mid Y)} \\
&\lesssim 1 + \sum_{K=1}^{2^L} \sum_{\boldsymbol{\kappa}: \sum_j k(j) \leq 2^L} \frac{\mathrm{e}^{\sum_j k(j)[\log 8 - c \log n]}}{\prod_j k(j)^{3/2}} \\
&\leq 1 + \sum_{K=1}^{2^L} \left(\frac{n}{2}\right)^K \mathrm{e}^{-K[(c-3/2)\log n - \log 8]} \leq 1 + \sum_{K=1}^{2^L} \mathrm{e}^{-K[(c-5/2)\log n - \log 8]} \\
&= 1 + \frac{1}{n^{c-5/2}/8 - 1}
\end{aligned}$$

and thereby

$$\left\{ \frac{\Pi[\Delta(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y) S(\mathcal{T} \to \mathcal{T}^-)} \times \max\left[1, \frac{\Pi(\mathcal{T} \mid Y) S(\mathcal{T} \to \mathcal{T}^-)}{\Pi(\mathcal{T}^- \mid Y) S(\mathcal{T}^- \to \mathcal{T})}\right] \right\} \leq 2|\mathcal{T}_{int}|(1 + o(1)).$$

S5.3.2. When $\mathcal{T}^* \not\subset \mathcal{T}$ (The Underfitted Case)

If the tree underfits and contains extra nodes, those are deleted first which coincides with the previous case. For those underfitted trees such that $\mathcal{T} \subset \mathcal{T}^*$, the internal nodes $\mathcal{T}_{int}$ do not include at least one pre-terminal node $\mathcal{P}(\mathcal{T}^*)$. According to the Assumption 1, the pre-terminal nodes have large enough signal, where $|\beta^*_{lk}| > A \log n/\sqrt{n}$ for some $A > 0$ for all $(l, k) \in \mathcal{P}(\mathcal{T}^*)$. Denote with $(l, k) \in \mathcal{P}(\mathcal{T}^*) \backslash \mathcal{T}_{int}$ the deepest rightmost signal pre-terminal node missed by $\mathcal{T}$. Let $(l^*, k^*) \in \mathcal{T}_{ext}$ be the external node of $\mathcal{T}$ that is closest to the signal node (l, k). Then $\mathcal{G}(\mathcal{T})$ is formed by growing a twig $[(l^*, k^*) \leftrightarrow (l, k)]$. In other words, $\mathcal{T}^+ = \mathcal{G}(\mathcal{T})$ is the smallest tree that contains nodes $(l, k) \cup \mathcal{T}_{int}$ inside and $\mathcal{T}^+_{int} = \mathcal{T}_{int} \cup [(l^*, k^*) \leftrightarrow (l, k)]$ has k more internal nodes relative to $\mathcal{T}_{int}$. Then we can write

$$\frac{S(\mathcal{T} \to \mathcal{T}^+)}{S(\mathcal{T}^+ \to \mathcal{T})} \leq 2|\mathcal{T}^+_{int}| \leq n.$$

Using the expression for the posterior ratio in (S10) and (S13) we again find that

$$\frac{\Pi(\mathcal{T} \mid Y)}{\Pi(\mathcal{T}^+ \mid Y)} \frac{S(\mathcal{T} \to \mathcal{T}^+)}{S(\mathcal{T}^+ \to \mathcal{T})} \leq n \mathrm{e}^{-(A^2/8)\log^2 n} = o(1).$$

We now proceed similarly as in Section S4.4.2. The precedents $\Delta(\mathcal{T})$ in (S28) of an underfitted model $\mathcal{T}$ are again divided into mutually exclusive categories defined in Section S4.4.2, i.e. $\Delta(\mathcal{T}) = \{\mathcal{T}\} \cup \mathcal{U}_1(\mathcal{T}) \cup \mathcal{U}_2(\mathcal{T})$. We again denote with $\Delta(\mathcal{T}, K) \subset \Delta(\mathcal{T})$ those precedents that are K steps away from $\mathcal{T}$ on a canonical path. Note that trees inside $\mathcal{U}_1(\mathcal{T}) \cap \Delta(\mathcal{T}, K)$ have at least K fewer internal nodes compared to $\mathcal{T}$ and trees inside $\mathcal{U}_2(\mathcal{T}) \cap \Delta(\mathcal{T}, K)$ have at least K extra internal nodes compared to $\mathcal{T}$.

Each tree in $\Delta(\mathcal{T}, K) \cap \mathcal{U}_1(\mathcal{T})$ misses K preterminal nodes in $\mathcal{P}(\mathcal{T}^*)$ (and thereby at least K internal nodes relative to $\mathcal{T}^*$). The cardinality $\mathrm{card}[\Delta(\mathcal{T}, K) \cap \mathcal{U}_1(\mathcal{T})]$ is thereby at most the number of binary trees with $|\mathcal{T}^*_{int}| - K$ internal nodes which equals the Catalan number $\mathbb{C}_{|\mathcal{T}^*_{int}|-K}$. By Lemma 6 in [9], we have $\mathbb{C}_K \asymp 4^K / K^{3/2}$ and using the expressions in (S10) we obtain for large enough $A > 0$ and $|\mathcal{T}^*_{int}| \lesssim \log^2 n$

$$\frac{\Pi[\mathcal{U}_1(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)} \leq \sum_{K=1}^{|\mathcal{P}(\mathcal{T}^*)|} \sum_{\mathcal{T}' \in \Delta(\mathcal{T}, K) \cap \mathcal{U}_1(\mathcal{T})} \frac{\Pi(\mathcal{T}' \mid Y)}{\Pi(\mathcal{T} \mid Y)} \lesssim |\mathcal{T}^*_{int}| \times 4^{|\mathcal{T}^*_{int}|} \times \mathrm{e}^{-(A^2/8)\log^2 n} = o(1). \tag{S36}$$

For the second type of underfitting precedents, we follow the same arguments as in Section S4.4.2 to conclude

$$\frac{\Pi[\mathcal{U}_2(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)} \leq \frac{1}{n^{c-5/2}/8 - 1} \frac{\Pi[\mathcal{U}_1(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)}$$

and thereby

$$\frac{\Pi[\Delta(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y) S(\mathcal{T} \to \mathcal{T}^+)} \leq \frac{2^L (D^L - 1)}{D - 1} (1 + o(1)).$$

The bound for the second underfitting case when $\mathcal{T} \not\subset \mathcal{T}^*$ and, at the same time, $\mathcal{T} \not\supset \mathcal{T}^*$ proceeds analogously, only without the set $\mathcal{U}_1(\mathcal{T})$ that is empty.

Putting it all together, the bound in (S30) yields $\rho(\mathcal{E}) \leq \frac{(D^L-1)}{D-1} 2^{L+1}(1 + o(1))$ for $c > 5/2 + \log D$.

The mixing bound (23) for Twiggy Bayesian CART is then obtained by the sandwich relation (19) with (21) in the same way as in the proof of Theorem 5.2.

S6. Proof of Theorem 5.4 (Mixing Upper Bound for Locally Informed Versions)

The proof of Theorem 5.4 rests on the two drift condition argument developed by [63]. In the next section, we provide details of the two drift functions chosen for our tree regression setting.

S6.1. Two Drift Conditions

Up to now, we have relied on the canonical path argument to upper-bound the mixing rates. In order to show linear mixing, we apply the two-drift condition framework developed by [63]. We say that a drift condition is satisfied on $A \subset \mathbb{T}_L$ when there exists a function $V : \mathbb{T}_L \to [1, \infty)$ and a constant $\lambda \in (0, 1)$ such that

$$(PV)(\mathcal{T}) \leq \lambda V(\mathcal{T}) \quad \text{for all} \quad \mathcal{T} \in A, \quad \text{where} \quad (PV)(\mathcal{T}) = \sum_{\widetilde{\mathcal{T}} \in \mathbb{T}_L} V(\widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}).$$

Similarly to the canonical path construction [60], [63] observe that in Bayesian variable selection, the chain tends to escape underfitted states. If it escapes to an overfitted state rather than the true covariate vector, then again the chain tends to escape to the true covariate vector. This idea was formalized by using two drift functions; drifting first to the non-underfittted states (an overfitted state or the true covariates), and then drifting to the true covariate vector. We also apply the same idea in the settings of Bayesian CART and Twiggy Bayesian CART.

Definition S1. We define two drift functions as

$$V_1(\mathcal{T}) = \exp\left\{ \frac{1}{2^L (C_{f_0} + 2)^2 (n+1)} (Y'(I - P_{\mathcal{T}}/n)Y) \right\}, \tag{S37}$$

$$V_2(\mathcal{T}) = \exp\left\{ \frac{1}{2^L} \left(|\mathcal{T}_{int} \backslash \mathcal{T}^*_{int}| + (|\mathcal{T}^*_{int} \backslash \mathcal{T}_{int}| \wedge 1) \times (2^L - |\mathcal{T}_{int} \cup \mathcal{T}^*_{int}|) \right) \right\}, \tag{S38}$$

where $P_{\mathcal{T}} = \boldsymbol{X}_{\mathcal{T}}\boldsymbol{X}'_{\mathcal{T}}$ so that $P_{\mathcal{T}}/n$ denotes the projection onto the column space spanned by $\boldsymbol{X}_{\mathcal{T}}$ in regular designs. The drift ratios are defined as

$$R_i(\mathcal{T}, \widetilde{\mathcal{T}}) = V_i(\widetilde{\mathcal{T}})/V_i(\mathcal{T}) - 1 \text{ for } i = 1, 2.$$

Remark S1. The second drift function V_2 is designed so that for any overfitting $\mathcal{T}$ we have $V_2(\mathcal{T}) = \exp\{|\mathcal{T}_{int} \backslash \mathcal{T}^*_{int}|/2^L\}$, while V_2 is a constant function on non-overfitting (underfitting) trees as $V_2(\mathcal{T}) = \exp\{1 - |\mathcal{T}^*_{int}|/2^L\}$. Therefore, for any $\mathcal{T}, \widetilde{\mathcal{T}} \in \mathbb{T}_L$ such that $\mathcal{T} \supset \mathcal{T}^*$ and $\widetilde{\mathcal{T}} \not\supset \mathcal{T}^*$, we can guarantee $V_2(\mathcal{T}) \leq V_2(\widetilde{\mathcal{T}})$ since $|\mathcal{T}_{int} \backslash \mathcal{T}^*_{int}| + |\mathcal{T}^*_{int}| = |\mathcal{T}_{int}| \leq 2^L$. The following lemma characterizes the chosen drift functions and the drift ratios for V_1 and V_2.

The first drift condition guarantees that the chain will frequently visit overfitted states, while the second condition guarantees that within the overfitted states, the chain will consistently attempt to hit the true tree $\mathcal{T}^*$. The following proposition is used to obtain the bound in Theorem 5.4.

Proposition S2. *Under the same assumptions of Theorem 5.4, with probability at least* $1 - 4/n - \mathrm{e}^{-n/8}$ *and with* $c > 5/2$ *we have the following properties of the drift functions:*

(i) For any underfitted tree $\mathcal{T} \in \mathbb{T}_L$ *such that* $\mathcal{T} \not\supseteq \mathcal{T}^*$,

$$\frac{(PV_1)(\mathcal{T})}{V_1(\mathcal{T})} \leq 1 - \frac{A^2}{2^{L+5}(C_{f_0} + 2)^2} \frac{\log^2 n}{n} + \frac{\mathrm{e} - 1}{2n^{(A^2/8)\log n - 1}}.$$

(ii) For any overfitted tree $\mathcal{T} \in \mathbb{T}_L$ *such that* $\mathcal{T} \supset \mathcal{T}^*$,

$$\frac{(PV_2)(\mathcal{T})}{V_2(\mathcal{T})} \leq 1 - \frac{1}{2^{L+2}} \frac{1}{(1 + n^{5/2-c})} + \frac{M}{n^{c-3/2}} + n^{1-(A^2 \log n)/8},$$

where $M = 1$ *for the Bayesian CART and* $M = 2L$ *for the Twiggy Bayesian CART.*

Proof. See Section S6.3.

S6.2. Proof of Theorem 5.4

We will use a similar strategy as in [63]. We first state the general two-stage drift condition theorem and its corollary, which are a slight modification from [63].

Theorem S3. *Consider a Markov chain $(X_t)_{t\in\mathbb{N}}$ on a state space $(\mathcal{X},\mathcal{E})$ where the σ-algebra $\mathcal{E}$ is countably generated. Assume a transition kernel P that is reversible with respect to a stationary distribution π and that P has a non-negative eigenspectrum. Suppose that there exist two drift functions $V_1, V_2 : \mathcal{X} \to [1,\infty)$ with constants $\lambda_1, \lambda_2 \in (0,1)$, a set $A \in \mathcal{E}$ and a point $x^* \in A$ such that*

(i) $PV_1 \le \lambda_1 V_1$ *on* A^c,

(ii) $PV_2 \le \lambda_2 V_2$ *on* $A\backslash\{x^*\}$, *and*

Further, suppose that A satisfies the following conditions for some finite constants $K_1 \ge 2$, $K_2 \ge 1$.

(iii) For any $x \in A$, $V_1(x) \le K_1/2$, and if $P(x,A^c) > 0$, $\mathbb{E}_x[V_1(X_1)|X_1 \in A^c] \le K_1/2$.

(iv) For any $x \in A$, $V_2(x) \le K_2$, and if $P(x,A^c) > 0$, $\mathbb{E}_x[V_2(X_1)|X_1 \in A^c] \ge V_2(x)$.

(v) For any $x \in A$, $P(x,A^c) \le q$ for some constant $q < \min\{1-\lambda_1, (1-\lambda_2)/K_2\}$.

Then, for every $x \in \mathcal{X}$ and $t \in \mathbb{N}$, we have

$$\|P^t(x,\cdot) - \pi\|_{TV} \le 4\alpha^{t+1}\left(1 + \frac{V_1(x)}{K_1}\right),$$

where α is a constant that satisfies

$$\alpha = \frac{1+\rho^m}{2} = \frac{1+K_1^m/u}{2},\ \rho = \frac{qK_2}{1-\lambda_2},\ u = \frac{1}{1-q/2},\ m = \frac{\log u}{\log(K_1/\rho)}.$$

Corollary S4. *Recall the definition of the ϵ-mixing time in (18). In the setting of Theorem S3, assume that $\lambda_1, \lambda_2 \to 1$ and $q \le \min\{1-\lambda_1, (1-\lambda_2)/C_2K_2\}$ for some universal constant $C_2 > 1$. With $K_1 = 2\sup_{x\in\mathcal{X}} V_1(x)$, for sufficiently large n, we have*

$$\tau_\epsilon \lesssim \frac{4\log(6/\epsilon)}{\log C_2}\log(C_2K_1)\max\left\{\frac{1}{1-\lambda_1}, \frac{C_2K_2}{1-\lambda_2}\right\}.$$

S6.2.1. Application of general two-stage drift condition

First, we introduce some notation. For $\mathcal{T}, \widetilde{\mathcal{T}} \in \mathbb{T}_L$, denote by $B(\mathcal{T}, \widetilde{\mathcal{T}})$ the posterior ratio $\Pi(\widetilde{\mathcal{T}} \mid Y)/\Pi(\mathcal{T} \mid Y)$. If $\widetilde{\mathcal{T}} \subset \mathcal{T}$ and $|\mathcal{T}\backslash\widetilde{\mathcal{T}}| = K$, we say $\widetilde{\mathcal{T}}$ is a K-node sub tree of $\mathcal{T}$.

In what follows, we show how the general two drift conditions of [63] can be applied in the context of regression trees. First, we consider the case of Bayesian CART. We will check the conditions in Theorem S3. To ensure a non-negative spectrum of the transition matrix, we consider the lazy version P_{lazy}, defined as $(P + I)/2$. We can account for this by scaling the terms added to 1 in Proposition S2 by a factor of $1/2$[4]. Therefore, the bounds in Proposition S2 become in a following manner. For any underfitted tree $\mathcal{T} \in \mathbb{T}_L$,

$$\frac{(P_{lazy}V_1)(\mathcal{T})}{V_1(\mathcal{T})} \leq 1 - \frac{A^2}{2^{L+6}(C_{f_0}+2)^2}\frac{\log^2 n}{n} + \frac{\mathrm{e}-1}{4n^{(A^2/8)\log n - 1}},$$

and for any overfitted tree $\mathcal{T} \in \mathbb{T}_L$ such that $\mathcal{T} \neq \mathcal{T}^*$,

$$\frac{(P_{lazy}V_2)(\mathcal{T})}{V_2(\mathcal{T})} \leq 1 - \frac{1}{2^{L+3}}\frac{1}{(1+n^{5/2-c})} + \frac{M}{2n^{c-3/2}} + \frac{1}{2}n^{1-(A^2\log n)/8}, \tag{S39}$$

where M is set to 1. We assign the values λ_1 and λ_2 to correspond to V_1 and V_2, where

$$\begin{aligned}\lambda_1 &= 1 - \frac{A^2}{72(C_{f_0}+2)^2}\frac{\log^2 n}{2^L n},\\ \lambda_2 &= 1 - \frac{1}{2^{L+7/2}}.\end{aligned} \tag{S40}$$

Let $A \subset \mathbb{T}_L$ be a set of overfitted trees and $\mathcal{T}^* \in A$ is the true tree. Then, conditions (i) and (ii) of Theorem S3 are satisfied with the above λ_1 and λ_2 for a large enough n and $c > 3$. Let $K_1 = 2\mathrm{e}, K_2 = \mathrm{e}$, and then by Lemma S5 (i), conditions (iii) and (iv) are satisfied; The second condition of (iv) is satisfied because for any proposal $\widetilde{\mathcal{T}} \in A^c$ from the state $\mathcal{T} \in A$, we have $V_2(\mathcal{T}) \leq V_2(\widetilde{\mathcal{T}})$ (see, Remark S1). To check condition (v), we set $C_2 = 2\mathrm{e}$ as a universal constant in Corollary S4. We want to see if $P_{lazy}(\mathcal{T}, A^c)$ for $\mathcal{T} \in A$ is smaller than $q = \min(1-\lambda_1, (1-\lambda_2)/C_2K_2)$. Note that the only transition that makes an overfitted tree to underfitted one is the PRUNE movement. Also, by (S42) and by the definition of P_{lazy}, we

[4]When $(PV)(\mathcal{T}) \leq (1-\delta)V(\mathcal{T})$ for some $\delta \in (0,1)$ and a drift function V, we have $(PV)(\mathcal{T})/2 \leq (1/2 - \delta/2)V(\mathcal{T})$. Therefore, $(P_{lazy}V)(\mathcal{T}) = (PV)(\mathcal{T})/2 + V(\mathcal{T})/2 \leq (1-\delta/2)V(\mathcal{T})$.

have $P_{lazy}(\mathcal{T}, A^c) \le B(\mathcal{T}, A^c)/2$. Therefore, by applying (S44), we have

$$\frac{B(\mathcal{T}, A^c)}{2} = \sum_{\widetilde{\mathcal{T}} \in A^c \cap N_p(\mathcal{T})} \frac{B(\mathcal{T}, \widetilde{\mathcal{T}})}{2} \le \sum_{\widetilde{\mathcal{T}} \in A^c \cap N_p(\mathcal{T})} \frac{1}{2n^{(A^2 \log n)/8}}$$
$$\le \frac{1}{2n^{(A^2 \log n)/8 - 1}} \le q = \min\left(\frac{A^2}{72(C_{f_0}+2)^2} \frac{\log^2 n}{2^L n}, \frac{1}{e^2 2^{L+9/2}}\right), \tag{S41}$$

for large enough A and n. Therefore, condition (v) is satisfied. Now, by applying Corollary S4, we have

$$\tau_\epsilon \lesssim \frac{4 \log(6/\epsilon)}{\log C_2} \log(2C_2 e) \max\left\{\frac{72(C_{f_0}+2)^2}{A^2} \frac{2^L n}{\log^2 n}, C_2 2e 2^{L+7/2}\right\}$$
$$\lesssim \log(6/\epsilon) \max\left(\frac{9\,(C_{f_0}+2)^2}{A^2} \frac{2^L n}{\log^2 n}, 2^{L+5}\right).$$

Lastly, when it comes to the Twiggy Bayesian CART, the only difference is that we have $M = 2L$ in (S39) instead of $M = 1$. This change does not affect the above proof because λ_2 in (S40) is still valid. Therefore, we finish our proof.

S6.3. Proof of Proposition S2

The proof is based on the key decomposition characterized in [63] as (for $i = 1, 2$)

$$\frac{(PV_i)(\mathcal{T})}{V_i(\mathcal{T})} = 1 + \sum_{\widetilde{\mathcal{T}} \neq \mathcal{T}} R_i(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}})$$
$$= 1 + \sum_{\star = g, p} \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_\star(\mathcal{T})} R_i(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}),$$

and on a useful bound for the transition probability (for any $\widetilde{\mathcal{T}} \neq \mathcal{T} \in \mathbb{T}_L$)

$$P(\mathcal{T}, \widetilde{\mathcal{T}}) = \min\{S(\mathcal{T} \to \widetilde{\mathcal{T}}), B(\mathcal{T}, \widetilde{\mathcal{T}}) S(\widetilde{\mathcal{T}} \to \mathcal{T})\} \le B(\mathcal{T}, \widetilde{\mathcal{T}}). \tag{S42}$$

Here, we consider both the Bayesian and Twiggy CART together in one place. This is possible by observing the following commonalities. (1) The neighbor sizes for both algorithms can be bounded by $|\mathcal{N}_p(\mathcal{T})| \le 2^L \le n/2$ and $|\mathcal{N}_g(\mathcal{T})| \le 2^L \le n/2$. For the Bayesian CART, $|\mathcal{N}_g(\mathcal{T})| = |\mathcal{T}_{ext}| \le 2^{L-1} \le n/2$. In the case of the Twiggy CART, $|\mathcal{N}_g(\mathcal{T})| =$

$|\mathcal{T}^L_{full,int}\backslash\mathcal{T}_{int}| \le 2^L \le n/2$. (2) The internal tree size difference between the existing tree and the proposed one is $k \ge 1$ for the Twiggy CART, while the Bayesian CART is a special case with $k = 1$. These commonalities allow for a unified framework to prove both algorithms.

The unimodal shape of the posterior is crucial for guaranteeing the linear mixing rate of LIT-MH. Therefore, we first characterize the posterior landscape, which implies the posterior unimodality given (Twiggy) GROW and PRUNE movements. Recall that on the event $\mathcal{A}_n$ defined in (S1), we have two prior ratios. First, similar to (S4) for any overfitted trees $\mathcal{T} \subset \widetilde{\mathcal{T}} \in \mathbb{T}_L$ such that $\mathcal{T} \supseteq \mathcal{T}^*$ and $|\widetilde{\mathcal{T}}\backslash\mathcal{T}| = K$,

$$\frac{\Pi(\widetilde{\mathcal{T}} \mid Y)}{\Pi(\mathcal{T} \mid Y)} \le n^{-K(c-3/2)}. \tag{S43}$$

Second, due to Assumption 1, for any underfitted tree $\mathcal{T} \in \mathbb{T}_L$, there exists a tree $\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})$ containing (at least) one extra signal node, which may not be unique. Such $\widetilde{\mathcal{T}}$ should have one extra node than $\mathcal{T}$ for the Bayesian CART or $k \ge 1$ extra nodes for the Twiggy Bayesian CART. For any such $\widetilde{\mathcal{T}}$, from (S13) we have

$$\frac{\Pi(\widetilde{\mathcal{T}} \mid Y)}{\Pi(\mathcal{T} \mid Y)} \ge n^{(A^2 \log n)/8}. \tag{S44}$$

Now, we characterize the properties of the two drift functions.

Lemma S5. *Under the same assumptions of Theorem 5.4, for any $\mathcal{T}, \widetilde{\mathcal{T}} \in \mathbb{T}_L$, the following statements hold with probability at least $1 - 4/n - \mathrm{e}^{-n/8}$.*

(i) $1 \le V_1(\mathcal{T}) \le \mathrm{e}$ *and* $1 \le V_2(\mathcal{T}) \le \mathrm{e}$.

(ii) When $\widetilde{\mathcal{T}} \supset \mathcal{T}$,

$$R_1(\mathcal{T}, \widetilde{\mathcal{T}}) \le 0, \quad R_1(\widetilde{\mathcal{T}}, \mathcal{T}) \ge 0.$$

(iii) When $\widetilde{\mathcal{T}} \supset \mathcal{T}$ and $|\widetilde{\mathcal{T}}_{int}\backslash\mathcal{T}_{int}| = k$, where $\mathcal{T} \supseteq \mathcal{T}^*$,

$$R_2(\mathcal{T}, \widetilde{\mathcal{T}}) \le \frac{2k}{2^L}, \quad R_2(\widetilde{\mathcal{T}}, \mathcal{T}) \le -\frac{k}{2^{L+1}}.$$

Proof. For part (i), we first show the upper bound of $V_1(\mathcal{T})$. We will work on $\mathcal{A}'_n = \mathcal{A}_n \cap \{\varepsilon : \|\varepsilon\|_2^2 \le 2n\}$, where $\mathcal{A}_n$ is defined in (S1). Since $\|\varepsilon\|_2^2 \sim \chi^2(n)$, by applying the tail bound in

[25] (Theorem 1), we have $\mathbb{P}(\|\varepsilon\|_2^2 > 2n) \le \mathrm{e}^{-n/8}$. Therefore, $\mathbb{P}(\mathcal{A}_n') \ge 1 - 4/n - \mathrm{e}^{-n/8}$. As $\boldsymbol{\nu} = \varepsilon$, we obtain the bound by observing that on the event $\mathcal{A}_n'$ with $p := 2^L$, the following holds.

$$\begin{aligned} Y'Y &= \|\boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + \|\boldsymbol{\nu}\|_2^2 + 2\boldsymbol{\nu}'\boldsymbol{X}\boldsymbol{\beta}^* \\ &\le n\|\boldsymbol{\beta}^*\|_2^2 + 2n + 2|\boldsymbol{\nu}'\boldsymbol{X}\boldsymbol{\beta}^*| \\ &\le n\,p\,C_{f_0}^2 + 2n + 4\|\boldsymbol{\beta}^*\|_2\sqrt{n^2\log p} \\ &\le n\,p\left(C_{f_0}^2 + 2/p + 4C_{f_0}\sqrt{\frac{\log p}{p}}\right) \le n\,2^L\,(C_{f_0}+2)^2, \end{aligned}$$

where we use the assumption that $|\beta_{lk}^*| \le C_{f_0}$. The other upper bound in part (i) is trivial since for any tree $\mathcal{T} \in \mathbb{T}_L$ we have $\mathcal{T}_{int} \le 2^L$. For part (ii), we observe that the column space spanned by $\boldsymbol{X}_{\mathcal{T}}$ is a subspace of the column space spanned by $\boldsymbol{X}_{\widetilde{\mathcal{T}}}$. Therefore,

$$V_1(\widetilde{\mathcal{T}})/V_1(\mathcal{T}) = \exp\left\{\frac{1}{2^L\,(C_{f_0}+2)^2(n+1)}\left(Y'(P_{\mathcal{T}}/n - P_{\widetilde{\mathcal{T}}}/n)Y\right)\right\} \le 1.$$

For part (iii), we have $|\widetilde{\mathcal{T}}\backslash\mathcal{T}^*| - |\mathcal{T}\backslash\mathcal{T}^*| = k \le 2^L$, and $V_2(\widetilde{\mathcal{T}})/V_2(\mathcal{T}) = e^{k/2^L}$. The result follows by using the two inequalities as in [63]

$$\mathrm{e}^x \le 1 + 2x, \;\; \mathrm{e}^{-x} \le 1 - \frac{x}{2}, \;\; \forall x \in [0,1]. \tag{S45}$$

□

S6.3.1. Drift condition for overfitted models (R_2)

Lemma S6. *Recall the definition of w_p and w_g in (16). Under the same assumptions of Theorem 5.4, for any overfitted tree $\mathcal{T} \in \mathbb{T}_L$,*

(i) $Z_g(\mathcal{T}) \le \frac{n^{-(c-5/2)}}{2}$.

(ii) For any subtree $\widetilde{\mathcal{T}} \subset \mathcal{T}$, $w_p(\widetilde{\mathcal{T}}|\mathcal{T}) = n^{c-3/2}$ if $\widetilde{\mathcal{T}}$ contains all the signal nodes, i.e., $\widetilde{\mathcal{T}} \supset \mathcal{T}^*$, and otherwise, $w_p(\widetilde{\mathcal{T}}|\mathcal{T}) = 1$.

(iii) $Z_p(\mathcal{T}) \le |a_{\mathcal{T}}| + |b_{\mathcal{T}}|\,n^{c-3/2}$, *where $a_{\mathcal{T}}$ and $b_{\mathcal{T}}$ in the decomposition $N_p(\mathcal{T}) = a_{\mathcal{T}} \cup b_{\mathcal{T}}$ are defined as follows.*

(a) *(Classical)* $a_{\mathcal{T}} = \mathcal{P}(\mathcal{T}) \cap \mathcal{T}^*$ *and* $b_{\mathcal{T}} = \mathcal{P}(\mathcal{T}) \backslash \mathcal{T}^*$.

(b) *(Twiggy) Denote by* $W(\mathcal{T})$ *all the twigs existing on* $\mathcal{T}$ *that end at a pre-terminal node (the Twiggy prune candidates).*

$a_{\mathcal{T}} = \{W \in W(\mathcal{T}) | W \cap \mathcal{T}^* \neq \emptyset\}$ *and* $b_{\mathcal{T}} = \{W \in W(\mathcal{T}) | W \cap \mathcal{T}^* = \emptyset\}$.

Proof. (i) By (S43),

$$Z_g(\mathcal{T}) = \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} \left(B(\mathcal{T}, \widetilde{\mathcal{T}}) \wedge n^{(A^2 \log n)/2} \right) \leq |\mathcal{N}_g(\mathcal{T})| n^{-(c-3/2)} \leq \frac{n^{-(c-5/2)}}{2}.$$

(ii) When $\widetilde{\mathcal{T}} \supset \mathcal{T}^*$, by applying (S43), we get $B(\mathcal{T}, \widetilde{\mathcal{T}}) \geq n^{c-3/2}$, and thus $w_p(\widetilde{\mathcal{T}}|\mathcal{T}) = n^{c-3/2}$ by definition in (16). Likewise, when $\widetilde{\mathcal{T}}$ loses a signal node compared with $\mathcal{T}$, by (S44), we have $B(\mathcal{T}, \widetilde{\mathcal{T}}) \leq n^{-(A^2 \log n)/2} \leq 1$. Therefore, it follows from definition (16) that $w_p(\widetilde{\mathcal{T}}|\mathcal{T}) = 1$. (iii) (a) is apparent by definition (16) and that the prune candidates are in $\mathcal{P}(\mathcal{T})$; For $\widetilde{\mathcal{T}} = \mathcal{T} \backslash \{(l,k)\}$, $B(\mathcal{T}, \widetilde{\mathcal{T}}) = 1$ if $(l,k) \in \mathcal{P}(\mathcal{T}) \cap \mathcal{T}^*$ and $B(\mathcal{T}, \widetilde{\mathcal{T}}) = n^{c-3/2}$ if $(l,k) \in \mathcal{P}(\mathcal{T}) \backslash \mathcal{T}^*$. Similarly, for (b), we apply the same reasoning to the twiggy candidate pool $W(\mathcal{T})$. □

Lemma S7. *Under the same assumptions of Theorem 5.4, for any overfitted tree* $\mathcal{T} \in \mathbb{T}_L$ *such that* $\mathcal{T} \neq \mathcal{T}^*$,

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_p(\mathcal{T})} R_2(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq -\frac{1}{2^{L+2}} \frac{1}{(1 + n^{5/2-c})} + n^{1-(A^2 \log n)/8},$$

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} R_2(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq \frac{M}{n^{c-3/2}},$$

where M *is defined as 1 for the Bayesian CART and* $2L$ *for the Twiggy CART.*

Proof. **The PRUNE movement.** Recall the definitions of $\alpha_{\mathcal{T}}$ and $b_{\mathcal{T}}$ in Lemma S6 (iii). First, consider $\widetilde{\mathcal{T}} \in b_{\mathcal{T}}$. We know that $b_{\mathcal{T}}$ is non-empty because $\mathcal{T} \neq \mathcal{T}^*$, which means there exists in $\mathcal{N}_p(\mathcal{T})$ a 1-node (k-node for Twiggy) subtree $\widetilde{\mathcal{T}} \subset \mathcal{T}$ such that $\widetilde{\mathcal{T}}_{int} \supseteq \mathcal{T}^*_{int}$. By (S43), we have $B(\widetilde{\mathcal{T}}, \mathcal{T}) \leq n^{-(c-3/2)} \leq n^{(A^2 \log n)/8}$. Therefore, for such $\widetilde{\mathcal{T}}$, $w_g(\mathcal{T}|\widetilde{\mathcal{T}}) = B(\widetilde{\mathcal{T}}, \mathcal{T})$, and

thus by applying Lemma S6 (i), we have

$$B(\mathcal{T},\widetilde{\mathcal{T}})S(\widetilde{\mathcal{T}} \to \mathcal{T}) \geq B(\mathcal{T},\widetilde{\mathcal{T}})\frac{w_g(\mathcal{T}|\widetilde{\mathcal{T}})}{2Z_g(\widetilde{\mathcal{T}})} = \frac{1}{2Z_g(\widetilde{\mathcal{T}})} \geq n^{c-5/2} \geq 1.$$

Therefore, by (S42), we have $P(\mathcal{T},\widetilde{\mathcal{T}}) = S(\mathcal{T} \to \widetilde{\mathcal{T}})$. Since the true signals contained in $\mathcal{T}$ and $\widetilde{\mathcal{T}}$ are the same, by definition (15) and (16), we have $S(\mathcal{T} \to \widetilde{\mathcal{T}}) \geq S_{PRUNE}(\mathcal{T} \to \widetilde{\mathcal{T}})/2 = n^{c-3/2}/2Z_p(\mathcal{T})$. Then, applying Lemma S5 (iii), and then Lemma S6 (iii), we find that

$$-R_2(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}}) \geq \frac{(|\mathcal{T}| - |\widetilde{\mathcal{T}}|)}{2^{L+1}}S(\mathcal{T} \to \widetilde{\mathcal{T}}) \geq \frac{n^{c-3/2}}{2^{L+2}(|a_{\mathcal{T}}| + |b_{\mathcal{T}}|\, n^{c-3/2})}.$$

Since $|b_{\mathcal{T}}| \geq 1$, we have for $c > 5/2$,

$$-\sum_{\widetilde{\mathcal{T}} \in b_{\mathcal{T}}} R_2(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}}) \geq \frac{|b_{\mathcal{T}}|n^{c-3/2}}{2^{L+2}(n + |b_{\mathcal{T}}|\, n^{c-3/2})} \geq \frac{1}{2^{L+2}}\frac{1}{1 + n^{5/2-c}}. \tag{S46}$$

Note that from (S42) and (S44), we have

$$\sum_{\widetilde{\mathcal{T}} \in a_{\mathcal{T}}} R_2(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}}) \leq |a_{\mathcal{T}}|\,(\mathrm{e}-1)\, n^{-(A^2 \log n)/8} \leq n^{1-(A^2 \log n)/8},$$

where we used $|a_{\mathcal{T}}| \leq 2^L \leq n/2$. Since $\mathcal{N}_p(\mathcal{T}) = a_{\mathcal{T}} \cup b_{\mathcal{T}}$, we have the result of the lemma.

The GROW movement. There is no additional signal node that can be added by GROW when the current state $\mathcal{T} \in \mathbb{T}_L$ is overfitted. Therefore, for any $\widetilde{\mathcal{T}} \supset \mathcal{T}$, from (S42) and (S43),

$$P(\mathcal{T},\widetilde{\mathcal{T}}) \leq B(\mathcal{T},\widetilde{\mathcal{T}}) \leq \frac{1}{n^{c-3/2}}. \tag{S47}$$

With Lemma S5 (iii) with $k = |\widetilde{\mathcal{T}}_{int}| - |\mathcal{T}_{int}|$, we obtain that

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} R_2(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}}) \leq \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} \frac{2k}{2^L}\frac{1}{n^{c-3/2}}. \tag{S48}$$

Since $|\mathcal{N}_g(\mathcal{T})| \leq 2^{L-1}$, and $k = 1$ in the Bayesian CART ($M = 1$), and $|\mathcal{N}_g(\mathcal{T})| \leq 2^L$ and $k \leq L$ in the Twiggy CART ($M = 2L$), we get the results. □

S6.3.2. Drift condition for underfitted models (R_1)

This section shares the same proof process for both the Twiggy CART and Bayesian CART algorithms, based on the observation at the beginning of Section S6.3.

Lemma S8. *Under the same assumptions of Theorem 5.4, for any underfitted tree $\mathcal{T} \in \mathbb{T}_L$ i.e., for any $\mathcal{T} \not\supset \mathcal{T}^*$,*

(i) $Z_g(\mathcal{T}) \geq n^{(A^2 \log n)/8}$.

(ii) $Z_p(\mathcal{T}) \leq n^{c-1/2}$.

Proof. (i) By (S44), we can always find a proposal $\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})$ such that $B(\mathcal{T}, \widetilde{\mathcal{T}}) \geq n^{(A^2 \log n)/8}$. (ii) is apparent by definition (16) and that $|\mathcal{N}_p(\mathcal{T})| \leq n$ for both Bayesian and Twiggy CART. □

Lemma S9. *Suppose $B(\mathcal{T}, \widetilde{\mathcal{T}}) \geq n^a$ for some $a \in \mathbb{R}$ and define*

$$b = \frac{1}{2^{L-1}(C_{f_0}+2)^2 n}\left(a \log n - \log\left(\frac{\Pi(\widetilde{\mathcal{T}})}{\Pi(\mathcal{T})}\right) - \frac{|\mathcal{T}_{ext}| - |\widetilde{\mathcal{T}}_{ext}|}{2}\log(1+n)\right). \tag{S49}$$

If $b \in [0,1]$, then $-R_1(\mathcal{T}, \widetilde{\mathcal{T}}) \geq b/2$.

Proof. From the posterior in (8), we relate B_1 to R_1 by

$$B(\mathcal{T}, \widetilde{\mathcal{T}}) = \frac{\Pi(\widetilde{\mathcal{T}})}{\Pi(\mathcal{T})}(1+n)^{\frac{|\mathcal{T}_{ext}|-|\widetilde{\mathcal{T}}_{ext}|}{2}}\left(\frac{V_1(\widetilde{\mathcal{T}})}{V_1(\mathcal{T})}\right)^{-n\,2^{L-1}(C_{f_0}+2)^2}.$$

Therefore, it follows by the assumption $\log B(\mathcal{T}, \widetilde{\mathcal{T}}) \geq a \log n$ that

$$a \log n \leq \log\left(\frac{\Pi(\widetilde{\mathcal{T}})}{\Pi(\mathcal{T})}\right) + \frac{|\mathcal{T}_{ext}| - |\widetilde{\mathcal{T}}_{ext}|}{2}\log(1+n) - n\,2^{L-1}(C_{f_0}+2)^2 \log(1+R_1(\mathcal{T}, \widetilde{\mathcal{T}})).$$

Therefore, $\log(1+R_1(\mathcal{T}, \widetilde{\mathcal{T}})) \leq -b$, which means $-R_1(\mathcal{T}, \widetilde{\mathcal{T}}) \geq 1 - e^{-b}$. If $b \in [0,1]$, we apply the second inequality in (S45) to get $-R_1(\mathcal{T}, \widetilde{\mathcal{T}}) \geq b/2$. □

Lemma S10. *Under the same assumptions of Theorem 5.4, for any underfitted tree $\mathcal{T} \in \mathbb{T}_L$*

i.e., for any $\mathcal{T} \not\supset \mathcal{T}^*$,

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \le -\frac{(A^2/8)\log^2 n}{2^{L+2} n (C_{f_0}+2)^2},$$

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_p(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \le \frac{\mathrm{e}-1}{2n^{(A^2/8)\log n - 1}}.$$

The bounds are for both the Bayesian and Twiggy CART.

Proof. **The GROW movement.** By (S44), there exists some tree $\mathcal{G}(\mathcal{T}) = \widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})$ containing at least one extra signal node, such that $B(\mathcal{T}, \mathcal{G}(\mathcal{T})) \ge n^{(A^2 \log n)/8}$. By Lemma S9 with $a = (A^2 \log n)/8$ and with large enough n so that b in (S49) is less than 1[5], we find that

$$\begin{aligned} -R_1(\mathcal{T}, \mathcal{G}(\mathcal{T})) &\ge \frac{1}{2n\, 2^{L-1}(C_{f_0}+2)^2} \left(a \log n - \log\left(\frac{\Pi(\mathcal{G}(\mathcal{T}))}{\Pi(\mathcal{T})}\right) - \frac{|\mathcal{T}_{ext}| - |\mathcal{G}(\mathcal{T})_{ext}|}{2} \log(1+n) \right) \\ &\ge \frac{1}{2^L n (C_{f_0}+2)^2} \left(a \log n - k \log\left(n^{-c}(1-n^{-c})\right) + \frac{k}{2}\log(1+n) \right) \\ &\ge \frac{1}{2^L n (C_{f_0}+2)^2}\, a \log n, \end{aligned} \tag{S50}$$

where $k = |\mathcal{G}(\mathcal{T})_{ext}| - |\mathcal{T}_{ext}| \ge 1$. Now, for some $V \ge 1$, consider a set of good GROW moves as

$$\mathcal{D} = \mathcal{D}(\mathcal{T}) = \{\widetilde{\mathcal{T}} \supset \mathcal{T} : B(\mathcal{T}, \widetilde{\mathcal{T}}) \ge n^{(A^2 \log n)/2 - V}\}.$$

Again using Lemma S9, we have for all $\widetilde{\mathcal{T}} \in \mathcal{D}(\mathcal{T})$,

$$\begin{aligned} -R_1(\mathcal{T}, \widetilde{\mathcal{T}}) &\ge \frac{1}{2n\, 2^{L-1}(C_{f_0}+2)^2} \left((a-V)\log n - \log\left(\frac{\Pi(\widetilde{\mathcal{T}})}{\Pi(\mathcal{T})}\right) - \frac{k}{2}\log(1+n) \right) \\ &= \frac{1}{2^L n (C_{f_0}+2)^2} \left((a-V)\log n - k \log\left(n^{-c}(1-n^{-c})\right) + \frac{k}{2}\log(1+n) \right) \\ &\ge \frac{1}{2^L n (C_{f_0}+2)^2} (a-V)\log n, \end{aligned} \tag{S51}$$

where $k = |\widetilde{\mathcal{T}}_{ext}| - |\mathcal{T}_{ext}| \ge 1$. Now we bound $P(\mathcal{T}, \widetilde{\mathcal{T}})$ for $\widetilde{\mathcal{T}} \in \mathcal{D}$. By the definition of $w_p(\mathcal{T}|\widetilde{\mathcal{T}})$, for any $\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})$,

$$S(\widetilde{\mathcal{T}} \to \mathcal{T}) \ge S_{PRUNE}(\widetilde{\mathcal{T}} \to \mathcal{T})/2 = \frac{w_p(\mathcal{T}|\widetilde{\mathcal{T}})}{2Z_p(\widetilde{\mathcal{T}})} \ge \frac{1}{2Z_p(\widetilde{\mathcal{T}})}. \tag{S52}$$

[5] When $\widetilde{\mathcal{T}} \supset \mathcal{T}$, it is apparent $b \ge 0$ because $k = |\widetilde{\mathcal{T}}_{ext}| - |\mathcal{T}_{ext}| \ge 0$ and $\log\left(\frac{\Pi(\widetilde{\mathcal{T}})}{\Pi(\mathcal{T})}\right) = k \log n^{-c}(1-n^{-c}) \le 0$.

This lower bound of $S(\widetilde{\mathcal{T}} \to \mathcal{T})$ in (S52) is why the two sided threshold in (16) is crucial in showing the mixing rate. Due to the two sided threshold, $S(\widetilde{\mathcal{T}} \to \mathcal{T})$ is not too small so that the transition kernel $P(\mathcal{T}, \widetilde{\mathcal{T}})$ is also not too small as the following: For any $\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})$,

$$\begin{aligned} P(\mathcal{T}, \widetilde{\mathcal{T}}) &= \min\{S(\mathcal{T} \to \widetilde{\mathcal{T}}), B(\mathcal{T}, \widetilde{\mathcal{T}}) S(\widetilde{\mathcal{T}} \to \mathcal{T})\} \\ &\geq \min\{\frac{w_g(\widetilde{\mathcal{T}}|\mathcal{T})}{2Z_g(\mathcal{T})}, \frac{B(\mathcal{T}, \widetilde{\mathcal{T}})}{2Z_p(\widetilde{\mathcal{T}})}\} \\ &\geq w_g(\widetilde{\mathcal{T}}|\mathcal{T}) \min\{\frac{1}{2Z_g(\mathcal{T})}, \frac{1}{2Z_p(\widetilde{\mathcal{T}})}\} \\ &\geq \frac{w_g(\widetilde{\mathcal{T}}|\mathcal{T})}{2Z_g(\mathcal{T})}. \end{aligned} \tag{S53}$$

In the last inequality, we used Lemma S8 (i) and (ii), for a large enough A,

$$Z_g(\mathcal{T}) \geq n^{(A^2 \log n)/8} \geq n^{c-1/2} \geq Z_p(\widetilde{\mathcal{T}}).$$

Define $\mathcal{D}' = \mathcal{D} \backslash \{\mathcal{G}(\mathcal{T})\}$, which may be empty. Let $W = \sum_{\widetilde{\mathcal{T}} \in \mathcal{D}'} w_g(\widetilde{\mathcal{T}}|\mathcal{T})$. Then, since $|\mathcal{N}_g(\mathcal{T}) \backslash \mathcal{D}| \leq n$, and for $\widetilde{\mathcal{T}} \notin \mathcal{D}$, $B(\mathcal{T}, \widetilde{\mathcal{T}}) \geq n^{(A^2 \log n)/8 - V}$, we have

$$\begin{aligned} Z_g(\mathcal{T}) &= \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} w_g(\widetilde{\mathcal{T}}|\mathcal{T}) \\ &= w_g(\mathcal{G}(\mathcal{T})|\mathcal{T}) + \sum_{\widetilde{\mathcal{T}} \in \mathcal{D}'} w_g(\widetilde{\mathcal{T}}|\mathcal{T}) + \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T}) \backslash \mathcal{D}} w_g(\widetilde{\mathcal{T}}|\mathcal{T}) \\ &= n^{(A^2 \log n)/8} + W + n^{(A^2 \log n)/8 - V + 1} \\ &\leq W + 2n^{(A^2 \log n)/8}. \end{aligned} \tag{S54}$$

Now, putting all things together using Lemma S5 (ii), (S50), (S51), (S53), and (S54), and recalling $a = (A^2 \log n)/8$, we get

$$\begin{aligned} &- \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \geq - \sum_{\widetilde{\mathcal{T}} \in \mathcal{D}(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \\ &\geq \frac{\log n}{2^L n (C_{f_0} + 2)^2} \left(a \frac{n^{(A^2 \log n)/8}}{2Z_g(\mathcal{T})} + (a - V) \sum_{\widetilde{\mathcal{T}} \in \mathcal{D}'(\mathcal{T})} \frac{w_g(\widetilde{\mathcal{T}}|\mathcal{T})}{2Z_g(\mathcal{T})} \right) \\ &\geq \frac{a \log n}{2^{L+2} n (C_{f_0} + 2)^2} \frac{n^{(A^2 \log n)/8} + (1 - V/a) W}{n^{(A^2 \log n)/8} + W/2}. \end{aligned} \tag{S55}$$

Therefore, as long as $a \geq 2V$, we have

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq -\frac{a \log n}{2^{L+2} n (C_{f_0} + 2)^2} = -\frac{(A^2/8) \log^2 n}{2^{L+2} n (C_{f_0} + 2)^2}.$$

The PRUNE movement. By applying Lemma S8 (i), we have for any 1-node (k-node for Twiggy) subtree $\widetilde{\mathcal{T}} \subset \mathcal{T}$

$$\begin{aligned} B(\mathcal{T}, \widetilde{\mathcal{T}}) S(\widetilde{\mathcal{T}} \to \mathcal{T}) &\leq B(\mathcal{T}, \widetilde{\mathcal{T}}) \frac{w_g(\mathcal{T} | \widetilde{\mathcal{T}})}{Z_g(\widetilde{\mathcal{T}})} \\ &\leq \frac{1}{Z_g(\widetilde{\mathcal{T}})} \leq \frac{1}{n^{(A^2 \log n)/8}}. \end{aligned}$$

Therefore, by (S42), we have

$$R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq R_1(\mathcal{T}, \widetilde{\mathcal{T}}) B(\mathcal{T}, \widetilde{\mathcal{T}}) S(\widetilde{\mathcal{T}} \to \mathcal{T}) \leq \frac{R_1(\mathcal{T}, \widetilde{\mathcal{T}})}{n^{(A^2 \log n)/8}}.$$

By Lemma S5 (i), $R_1(\mathcal{T}, \widetilde{\mathcal{T}}) \leq \mathrm{e} - 1$, and the pool size is $|\mathcal{N}_p(\mathcal{T})| \leq n/2$. Therefore,

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_p(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq \frac{\mathrm{e} - 1}{2 n^{(A^2/8) \log n - 1}}.$$

□

S6.4. Proof of Remark 10

Here, we present the non-informed counterpart of Proposition S2. To achieve this, we modify V_1 as

$$V_1(\mathcal{T}) = \exp\left\{ \frac{1}{2^L C_{f_0}^2 n} ((\boldsymbol{X}\boldsymbol{\beta}^*)'(I - P_{\mathcal{T}}/n)\boldsymbol{X}\boldsymbol{\beta}^*) \right\}, \tag{S56}$$

which is designed to ignore the error terms. This is to guarantee $R_1(\mathcal{T}, \widetilde{\mathcal{T}}) = 0$ for $\widetilde{\mathcal{T}}$ obtained by pruning non-signals from $\mathcal{T} \in \mathbb{T}_L$. All the properties of Lemma S5 can be shown to apply to the new V_1 on the event $\mathcal{A}_n$ (with probability at least $1 - 4/n$). For example, for Lemma S5 (i), we obtain the bound by

$$\|\boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 = n\|\boldsymbol{\beta}^*\|_2^2 \leq n\, 2^L C_{f_0}^2. \tag{S57}$$

Proposition S11. *Under the same assumptions of Theorem 5.4, for the Bayesian CART and Twiggy Bayesian CART algorithms described in Section 2.1.2 and Section 3.1, with probability at least* $1 - 4/n$ *we have the following properties of the drift functions.*

(i) For any underfitted tree $\mathcal{T} \in \mathbb{T}_L$,

$$\frac{(PV_1)(\mathcal{T})}{V_1(\mathcal{T})} \leq 1 - \frac{\delta_1 A^2 \log^2 n}{2^{2L+2} C_{f_0}^2 n} + \frac{\mathrm{e} - 1}{2n^{(A^2/8)\log n - 1}}.$$

(ii) For any overfitted tree $\mathcal{T} \in \mathbb{T}_L$ *such that* $\mathcal{T} \neq \mathcal{T}^*$, *for* $c > 3/2$,

$$\frac{(PV_2)(\mathcal{T})}{V_2(\mathcal{T})} \leq 1 - \frac{1}{2^{2L+2}} + \frac{M}{n^{c-3/2}} + n^{1-(A^2 \log n)/8}, \tag{S58}$$

where $M = \delta_1 = 1$ *for the Bayesian CART and* $M = 2L$, $\delta_1 = \frac{2(D-1)}{D^L - 1}$ *for the Twiggy Bayesian CART.*

To ensure that the upper bound in (S58) is less than 1, we impose a stronger condition on c, requiring $c \geq 4$. This is because if $c = 7/2$, we may have $1/2^{2L+1} \asymp \frac{n^{c-3/2}}{2}$, for example when $L = L_{max}$. Now, with $\lambda_1 = 1 - \frac{\delta_1 A^2 \log^2 n}{2^{2L+4} C_{f_0}^2 n}$ and $\lambda_2 = 1 - \frac{1}{2^{2L+4}}$, it is straightforward to extend Section S6.2.1 (the application of the two-drift condition) to this case, obtaining the bound in Remark 10. Proposition S11 is derived from the non-informed counterpart of Lemma S7 and Lemma S10 presented below.

Lemma S12. *Under the same assumptions of Theorem 5.4, for the Bayesian CART and Twiggy Bayesian CART algorithms described in Section 2.1.2 and Section 3.1, for any overfitted tree* $\mathcal{T} \in \mathbb{T}_L$, *such that* $\mathcal{T} \neq \mathcal{T}^*$,

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_p(\mathcal{T})} R_2(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq -\frac{1}{2^{2L+2}} + n^{1-(A^2 \log n)/8}, \tag{S59}$$

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} R_2(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq \frac{M}{n^{c-3/2}}, \tag{S60}$$

where $M = 1$ *for the Bayesian CART, and* $M = 2L$ *for the Twiggy CART.*

Proof. **The PRUNE movement.** First, we consider the case of the Bayesian CART. The proof is the same as in Lemma S7, except for the bound on $-\sum_{\widetilde{\mathcal{T}} \in b_{\mathcal{T}}} R_2(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}})$. Since $\mathcal{T}$ is

an overfitted tree and $\mathcal{T} \neq \mathcal{T}^*$, we have $|b_{\mathcal{T}}| \geq 1$. We take any $\widetilde{\mathcal{T}} \in b_{\mathcal{T}}$. By (S43), we have $B(\mathcal{T}, \widetilde{\mathcal{T}}) \geq n^{(c-3/2)}$ and $n^{(c-3/2)} S(\widetilde{\mathcal{T}} \to \mathcal{T})/S(\mathcal{T} \to \widetilde{\mathcal{T}}) \geq 1$. This results in the acceptance rate of 1, implying that for such $\widetilde{\mathcal{T}}$, $P(\mathcal{T}, \widetilde{\mathcal{T}}) = S(\mathcal{T} \to \widetilde{\mathcal{T}})$, and thus by applying Lemma S5 (iii), and then Lemma S6 (iii), we find that

$$-R_2(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \geq \frac{(|\mathcal{T}| - |\widetilde{\mathcal{T}}|)}{2^{L+1}} S(\mathcal{T} \to \widetilde{\mathcal{T}}) \geq \frac{(|\mathcal{T}| - |\widetilde{\mathcal{T}}|)}{2^{L+1} \times 2^{L+1}} \geq \frac{1}{2^{2L+2}}. \tag{S61}$$

The other parts of the proof in Lemma S7 do not depend on the choice of the proposal probability $S(\cdot \to \cdot)$. Therefore, we have the result.

Now, for the Twiggy Bayesian CART, the only difference is in the lower bound of $P(\mathcal{T}, \widetilde{\mathcal{T}})$. By (13), (14), (S42), and $D \leq \mathrm{e}$,

$$\begin{aligned} P(\mathcal{T}, \widetilde{\mathcal{T}}) &= \min\{S(\mathcal{T} \to \widetilde{\mathcal{T}}), B(\mathcal{T}, \widetilde{\mathcal{T}}) S(\widetilde{\mathcal{T}} \to \mathcal{T})\} \\ &\geq \min\left\{\frac{1}{2^{L+1}}, \frac{D-1}{2^L (D^L - 1)} n^{c-3/2}\right\} = \frac{1}{2^{L+1}}. \end{aligned}$$

Therefore, by proceeding as above, we obtain the bound.

The GROW movement. The bound in (S60) is the same as the informed case in Lemma S7. The proof of Lemma S7 does not depend on a specific choice of $S(\cdot \to \cdot)$ but uses only (S47), which is obtained by (S42). Since the non-informed version shares all the same movement neighbor and posterior ratios, the proof also results in (S60) in the current lemma.

Lemma S13. *Under the same assumptions of Theorem 5.4, for the Bayesian CART and Twiggy Bayesian CART algorithms described in Section 2.1.2 and Section 3.1, for any underfitted tree $\mathcal{T} \in \mathbb{T}_L$ i.e., for any $\mathcal{T} \not\supset \mathcal{T}^*$,*

$$\begin{aligned} \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) &\leq -\frac{\delta_1 A^2 \log^2 n}{2^{2L+2} C_{f_0}^2 n}, \\ \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_p(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) &\leq \frac{\mathrm{e} - 1}{2n^{(A^2/8)\log n - 1}}, \end{aligned}$$

where $\delta_1 = 1$ for Bayesian CART and $\delta_1 = \frac{2(D-1)}{D^L - 1}$ for Twiggy Bayesian CART.

Proof. **The GROW movement.** Consider first the case of the Bayesian CART. As in Lemma S10, by (S44) there exists a tree $\mathcal{G}(\mathcal{T}) \supset \mathcal{T}$ containing at least one extra signal node, such that $B(\mathcal{T}, \mathcal{G}(\mathcal{T})) \geq n^{(A^2 \log n)/2}$. This large posterior rate implies $P(\mathcal{T}, \mathcal{G}(\mathcal{T})) = S_{GROW}(\mathcal{T} \to \mathcal{G}(\mathcal{T}))/2 \geq 1/2^{L+1}$. By inequality (S45), and with a decomposition $P_{\mathcal{G}(\mathcal{T})} = P_{\mathcal{T}} + P_{\mathcal{G}(\mathcal{T}) \setminus \mathcal{T}}$,

$$\begin{aligned} -R_1(\mathcal{T}, \mathcal{G}(\mathcal{T})) &\geq \frac{1}{2} \frac{1}{C_{f_0}^2 n\, 2^L} \left((\boldsymbol{X}\boldsymbol{\beta}^*)' (P_{\mathcal{G}(\mathcal{T})}/n - P_{\mathcal{T}}/n) \boldsymbol{X}\boldsymbol{\beta}^* \right) \\ &\geq \frac{1}{2^{L+1}} \frac{1}{C_{f_0}^2 n} \left((\boldsymbol{X}\boldsymbol{\beta}^*)' (P_{\mathcal{G}(\mathcal{T}) \setminus \mathcal{T}}/n) \boldsymbol{X}\boldsymbol{\beta}^* \right) \\ &\geq \frac{1}{2^{L+1}\, n\, C_{f_0}^2} n \frac{A^2 \log^2 n}{n} = \frac{A^2 \log^2 n}{2^{L+1}\, n\, C_{f_0}^2}. \end{aligned}$$

Besides $R_1(\mathcal{T}, \widetilde{\mathcal{T}}) \leq 0$ for all $\widetilde{\mathcal{T}} \in N_G(\mathcal{T})$ by Lemma S5 (ii). Therefore, by considering this movement to $\mathcal{G}(\mathcal{T})$, we obtain

$$\begin{aligned} &- \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \geq -R_1(\mathcal{T}, \mathcal{G}(\mathcal{T})) P(\mathcal{T}, \mathcal{G}(\mathcal{T})) \\ &\geq \frac{A^2 \log^2 n}{2^{L+1}\, n\, C_{f_0}^2} P(\mathcal{T}, \mathcal{G}(\mathcal{T})) \geq \frac{A^2 \log^2 n}{2^{2L+2}\, n\, C_{f_0}^2}. \end{aligned} \tag{S62}$$

Now, we consider the case of the Twiggy Bayesian CART. The only change in the above calculation is the lower bound for $P(\mathcal{T}, \mathcal{G}(\mathcal{T}))$. By (13), (14), and (S42),

$$\begin{aligned} P(\mathcal{T}, \mathcal{G}(\mathcal{T})) &= \min\{S(\mathcal{T} \to \mathcal{G}(\mathcal{T})), B(\mathcal{T}, \mathcal{G}(\mathcal{T})) S(\mathcal{G}(\mathcal{T}) \to \mathcal{T})\} \\ &\geq \min\left\{ \frac{D-1}{2^L (D^L - 1)}, \frac{n^{(A^2 \log n)/2)}}{2^{L+1}} \right\} = \frac{D-1}{2^L (D^L - 1)} = \frac{\delta_1}{2^{L+1}}. \end{aligned}$$

Therefore, by proceeding as above, we obtain the bound.

The PRUNE movement. Consider first the case of the Bayesian CART. There are two cases of a 1-node subtree $\widetilde{\mathcal{T}} \subset \mathcal{T}$. First, when $\widetilde{\mathcal{T}}$ is made by pruning a *non-signal* from $\mathcal{T}$: Due to the modification of the new V_1 in (S56), we have $R_1(\mathcal{T}, \widetilde{\mathcal{T}}) = 0$. Second, when $\widetilde{\mathcal{T}}$ is made by pruning a *signal* from $\mathcal{T}$: We have from (S44), $B(\mathcal{T}, \widetilde{\mathcal{T}}) \leq n^{-(A^2 \log n)/8}$, and by Lemma S5 (i), $R_1(\mathcal{T}, \widetilde{\mathcal{T}}) \leq \mathrm{e} - 1$. From (S42), we have $P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq B(\mathcal{T}, \widetilde{\mathcal{T}})$. Therefore, considering the maximum possible value of $R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}})$ and since the pool size is $|\mathcal{N}_p(\mathcal{T})| \leq 2^L \leq n/2$,

we have

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_p(\mathcal{T})} R_1(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq \frac{\mathrm{e} - 1}{2n^{(A^2/8)\log n - 1}}.$$

Now, when it comes to the case of the Twiggy Bayesian CART, there are two cases of a k-node subtree $\widetilde{\mathcal{T}} \subset \mathcal{T}$. First, when all nodes of $\mathcal{T} \backslash \widetilde{\mathcal{T}}$ are non-signals, and second when $\mathcal{T} \backslash \widetilde{\mathcal{T}}$ contains at least one signal. In these cases, the above reasoning applies in the same way.

Remark S2. The only algorithmic difference between the informed versions and their non-informed counterparts is the proposal distribution $S(\cdot \to \cdot)$, whether proposing uniformly or informatively. This difference brings two major benefits compared to the original non-informed algorithms. First, the proposal probability of $\mathcal{G}(\mathcal{T})$ from $\mathcal{T}$ in the canonical path, or namely, the best movement, is significantly improved. Note that for any MH-algorithm, $P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq S(\mathcal{T} \to \widetilde{\mathcal{T}})$. Therefore, no matter how much posterior increase can be brought by the best movement, its transition probability is still can be as small as $1/2^L$ in the non-informed algorithms. This contrast is highlighted by comparing the large proposal probability bound in (S46) with the small lower bound (of a uniform proposal) in (S61). Second, the change to the informed proposal increases the transition probability of a set of movements that reduce the drift function values, or namely, good movements. This plays an important role especially when handling underfitted tree cases (GROW) as in (S55) and the following display, which exploit that the transition probability of good movements is more than $1/4$. Although there is no guarantee that there will be multiple good movements other than the best movement, even when there is only a single best movement, (S55) implies then its transition probability is greater than $1/4$. In the proving technique of two-drift conditions, movements that have a small drift ratio (R_1 and R_2) are good movements. Here, such many good movements collectively reduce the expectation of the ratio in the next MCMC step. On the contrary, in the above proof of Remark 10, we considered only a single best movement when handling underfitted tree cases (GROW) as in (S62). Note that this consideration was unavoidable. In the non-informed setting, like in the informed setting, guaranteeing multi-

ple good movements (here, in the sense of posterior increase) is difficult other than a single best movement. However, unlike the case of the informed setting, the uniform proposal only guarantees that the transition probability of the best movement (signal obtaining) can be as small as $1/2^{L+1}$. Therefore, the upper bound in Remark 10 slower than that of the informed algorithms is not because only a single movement was considered in (S62). Rather, this is due to the difference in the proposal distributions.

S7. Comparison to [60]

[60] showed rapid mixing of MH algorithm of a Bayesian variable selection problem for a standard linear model

$$Y = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\omega},$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the unknown regression vector and $\boldsymbol{\omega} \sim \mathcal{N}(0, \sigma_0 I_n)$. Here p is the number of covariates and n is the sample size. Denote by $\boldsymbol{\gamma} \in \{0,1\}^p$ the vector of indicators for influential regression weights in $\boldsymbol{\beta}^*$. A coefficient $\beta_j^* \in \boldsymbol{\beta}^*$ is considered influential if $|\beta_j^*| \geq C_\beta$ for a constant $C_\beta > 0$ that depends on (σ_0, n, p). The MH algorithm for Bayesian variable selection generates its proposal by randomly swapping two indicators or adding/removing one indicator in $\boldsymbol{\gamma}$. The Bayesian variable selection problem is highly connected to our tree sampling. However, there is no requirement on the selected variables in $\boldsymbol{\gamma}$ to maintain a systematic structure. This is an important contrast from our setting, where the selected nodes should compose a valid tree shape. Therefore, it is an interesting question whether this imposed tree structure would encourage even more rapid mixing in comparison with the standard Bayesian variable selection problem.

In answering this question, we introduce some notations in [60] and compare them with our settings. Denote the design matrix of the selected columns by $\boldsymbol{X}_{\boldsymbol{\gamma}} \in \mathbb{R}^{n\times|\boldsymbol{\gamma}|}$. The Bayesian

hierarchical model considered in [60] is

$$\begin{aligned} \boldsymbol{\omega} &\sim \mathcal{N}(0, \phi^{-1} I_n) \\ \pi(\phi) &\propto \frac{1}{\phi}, \\ \boldsymbol{\beta}|\boldsymbol{\gamma} &\sim \mathcal{N}(0, g\phi^{-1}(\boldsymbol{X}'_{\boldsymbol{\gamma}}\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}), \\ \Pi(\boldsymbol{\gamma}) &\propto \left(\frac{1}{p}\right)^{\kappa|\boldsymbol{\gamma}|} \mathbb{I}[|\boldsymbol{\gamma}| \leq s_0], \end{aligned} \tag{S63}$$

where s_0 is the upper bound on the maximum number of important covariates, $g > 0$ is the degree of dispersion in the regression prior, and κ is the model size penalty. A hyperparameter $\alpha \geq 1/2$ is used to constraint the relationship between g and p by $g \asymp p^{2\alpha}$. Our setting corresponds to when $p = n/2$, $C_\beta = A \log n/\sqrt{n}$, $g = n$, $s_0 = 2^L$, $\sigma_0^2 = 1$, $\kappa = c$, and $\alpha = 1/2$. The consistency condition in [60] ((9a), High SNR condition) is satisfied if $A^2 \geq 30(4.5 + \kappa)$ given $L_{max} \geq 5$. Due to the orthogonality of our design matrix $\boldsymbol{X}'\boldsymbol{X} = nI_p$ and Assumption 1, these hyperparameter settings meet their regularity conditions (Assumption A to D in [60]) by additionally assuming $C_{f_0}^2 2^L \leq \log(n/2)$, $c \geq 17 + 1/2$ and $L \leq L_{max} - \log_2 L_{max} - 4$ as follows.

Assumption A) The condition (7a) is written as $C_{f_0}^2 2^L \leq \log(n/2)$, which leads to satisfy $\|\frac{1}{\sqrt{n}}\boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 \leq \log n$. The other condition (7b) is trivial since non-influential nodes are regarded to have zero coefficients, with $\tilde{L} = 0$.

Assumption B) The lower restricted eigenvalue condition is met for any $\nu \in (0, 1]$ since $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = I_p$. Due to the orthogonality of $\boldsymbol{X}$ as discussed in [60], the sparse projection condition is always satisfied when $L = 4\nu^{-1}$. We set $\nu = 1$ and $L = 4$ since smaller L is a less restrictive condition.

Assumption C) It is trivial with the hyperparameter settings of $g = n, p = n/2, \alpha = 1/2$ and $c = \kappa \geq 17 + 1/2$.

Assumption D) Version $D(s_0)$ should be met for the Theorem 2 in [60], which is

$$\max\{1, (2\nu^{-2}\omega(X) + 1)s^*\} \leq s_0 \leq \frac{1}{32}\left\{\frac{n}{\log p} - 8\tilde{L}\right\},$$

where $\omega := \max_{\mathcal{T}} \|(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}})^{-1}\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}^*\setminus\mathcal{T}}\|_{\text{op}}^2$. In our translation, $s_0 = 2^L$ and $s^* = |\mathcal{T}^*_{int}|$. The lower bound condition on s_0 is trivial because X is orthogonal, i.e., $w(X) = 0$. Since $\tilde{L} = 0$, the right upper bound is satisfied when $L \le L_{max} - \log_2 L_{max} - 4$.

Therefore, the rapid mixing guarantee (Theorem 2 in [60]) is translated as follows.

Theorem S1 ([60] Theorem 2)**.** *Assume the model* (3) *with Assumption 1 and the spike-and-slab prior in* (S63) *with* $\kappa = c \ge 17 + 1/2$. *Consider the Spike-and-Slab MH algorithm in [60] without a tree structure restriction (*$\boldsymbol{\gamma}$ *is the vectorized* $\mathcal{T} \in \mathbb{T}_L$*). Assume* $C_{f_0}^2 2^L \le \log(n/2)$ *and* $1 \le L \le L_{max} - \log_2 L_{max} - 4$. *With a large enough constant* $A > 0$*, with probability at least* $1 - c_3 p^{-c_4}$,

$$\tau_\epsilon \le 3 \times 2^{2L} n \left[n \log(n/2) + (1 + 4c) 2^L \log(n/2)) + \log(2/\epsilon) \right], \tag{S64}$$

for some c_3, *and* c_4.

Now, for the comparison purpose, we match the settings by applying the sparsity prior in (S63) to our result instead of the classical Bayesian CART prior in 2.1.1. That is, for the comparison, we use the prior $\Pi(\mathcal{T}) \propto \widetilde{p}_{lk}^{(-|\mathcal{T}_{int}|)} \mathbb{I}[|\mathcal{T}_{int}| \le 2^L]$, where $\widetilde{p}_{lk} = (n/2)^{-c}$. Because $\widetilde{p}_{lk} = 2^c p_{lk}$, it is easy to verify the consistency and the mixing rate results in Theorem 2.2 and Theorem 5.2 for $c > 7/2$ as long as $\widetilde{p}_{lk} < 1/2$. Therefore, we can compare our upper bound in (22) against the bound in (S64).

S8. Additional Visualizations

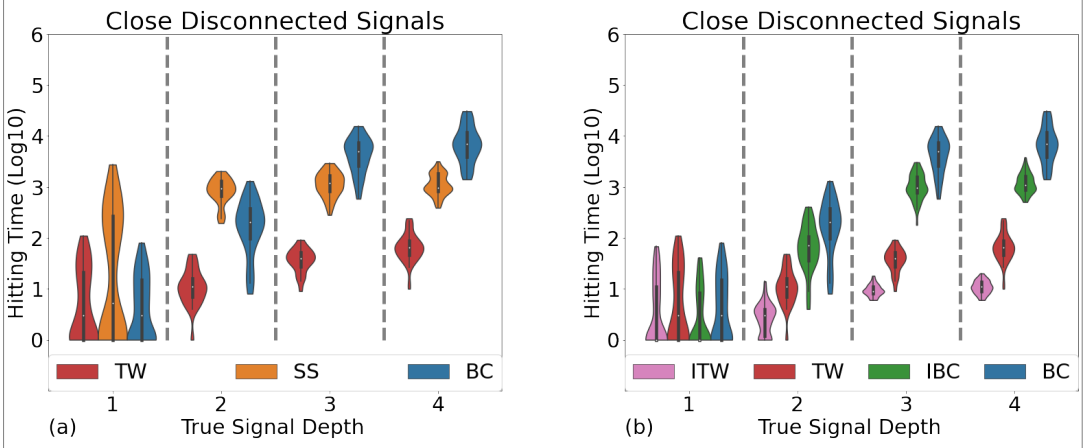

Fig S1. *Hitting time* $\tau = \min_{t\geq 0}\{\mathcal{B} \subset \mathcal{T}_{int}^t\}$ *when true tree gets deeper of Case (3) (by gradually making the deeper part of the tree). (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: original and informed Twiggy Bayesian CART, ss: Spike-and-Slab with prior* p_1^{ss}*. (a) Twiggy Bayesian CART < Bayesian CART. Spike-and-Slab performance is consistent across the true tree depth. (b) Informed (Twiggy) Bayesian CART hits the true signals faster than (Twiggy) Bayesian CART. However, informed Bayesian CART does not hit faster than Twiggy Bayesian CART.*

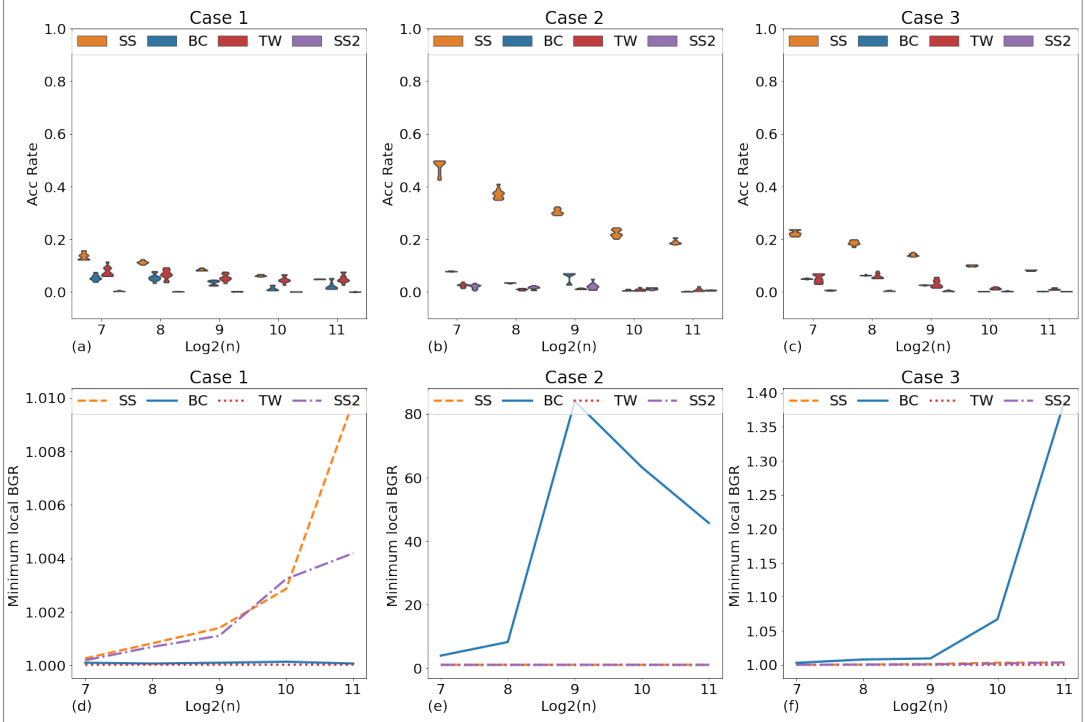

Fig S2. *The acceptance rates and minimum local BGRs. (Legend) BC: Bayesian CART, TW: Twiggy Bayesian CART, SS and SS2: Spike-and-Slab with prior* p_1^{ss} *and* p_2^{ss} *respectively.*

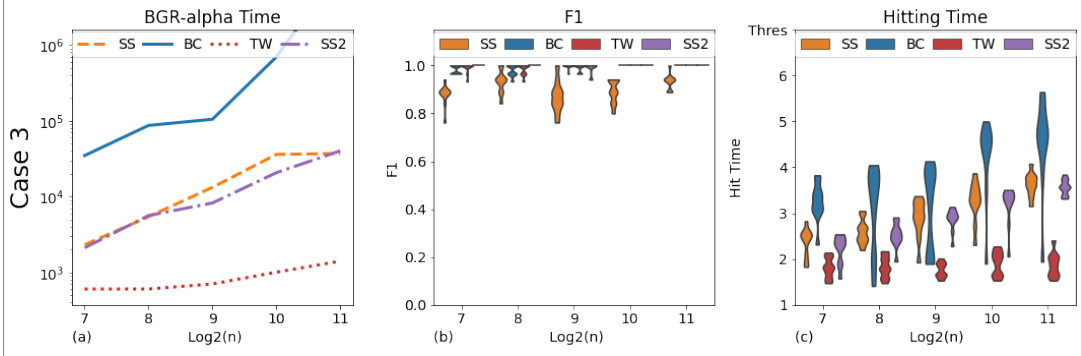

Fig S3. *The MCMC performance measures for Case (3). (Legend) BC: Bayesian CART, TW: Twiggy Bayesian CART, SS and SS2: Spike-and-Slab with prior* p_1^{ss} *and* p_2^{ss} *respectively.*

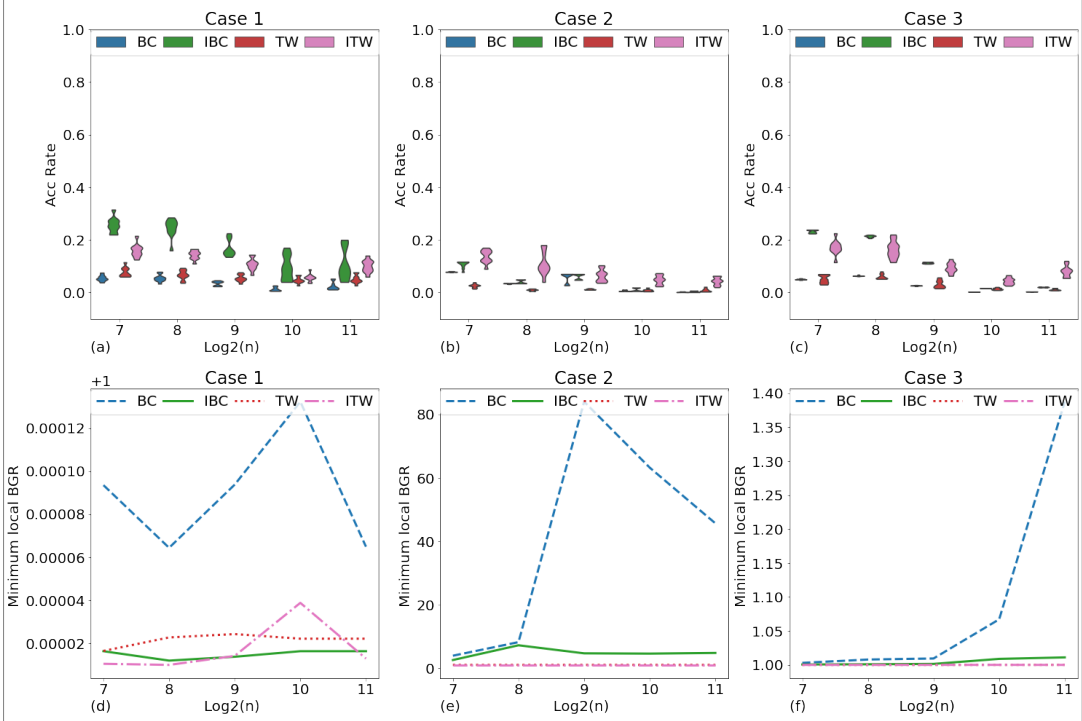

Fig S4. *The acceptance rates, hit time and minimum local BGRs for Case (1) and (2). (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: original and informed Twiggy Bayesian CART.*

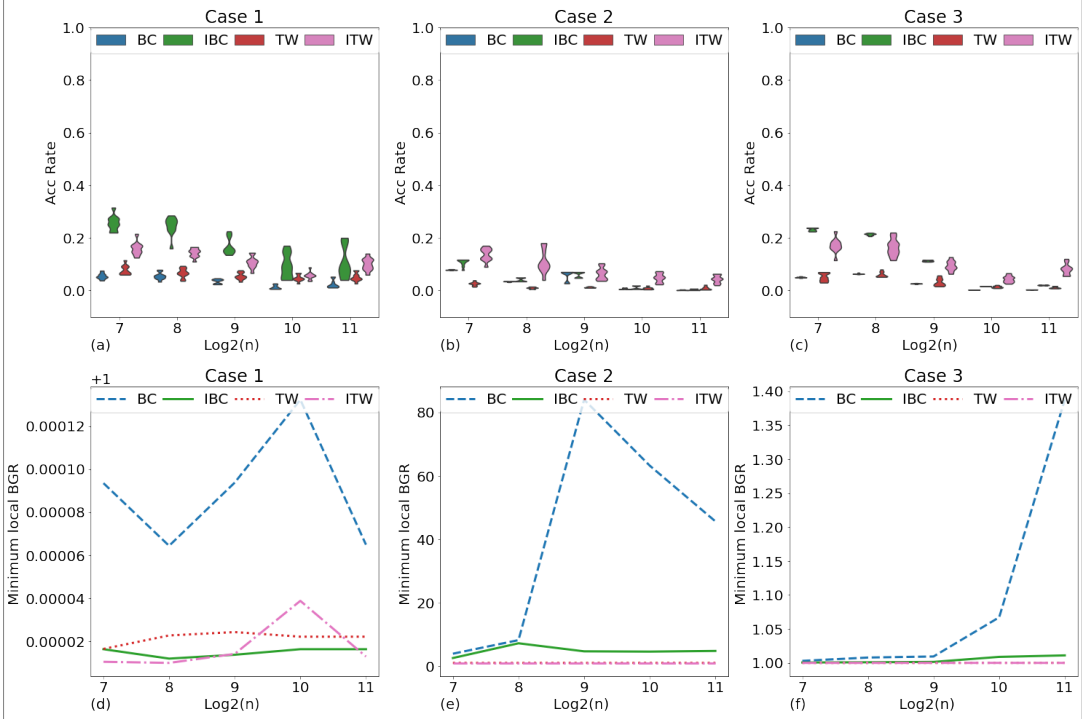

Fig S5. *The MCMC performance measures for Case (3). (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: original and informed Twiggy Bayesian CART.*

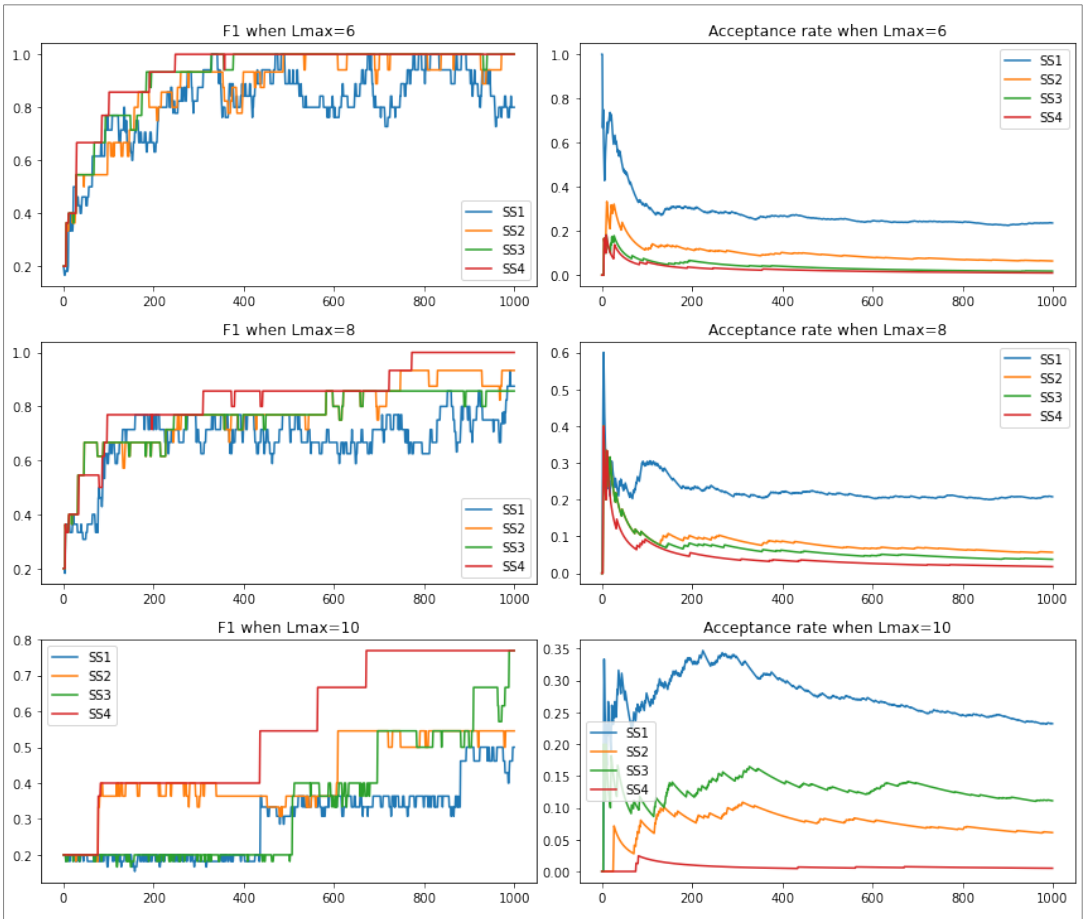

Fig S6. *The behavior of Spike-and-Slab for Case (3) for different node inclusion priors for increasing data size (L_{max}). The x-axis in all plots are the number of iterations. SS1: $p_{lk} = 0.25/2^{L_{max}-6}$. SS2: $p_{lk} = 0.05/2^{L_{max}-6}$. SS3: $0.01\, n^{1/4} 2^{-l/2}$. SS4: $0.01\, n^{1/4} 6^{-l/2}$. SS4 has the smallest node inclusion prior, and so the smallest acceptance rate. However, in terms of grabbing the true signals without overfitting, SS4 shows the best performance.*

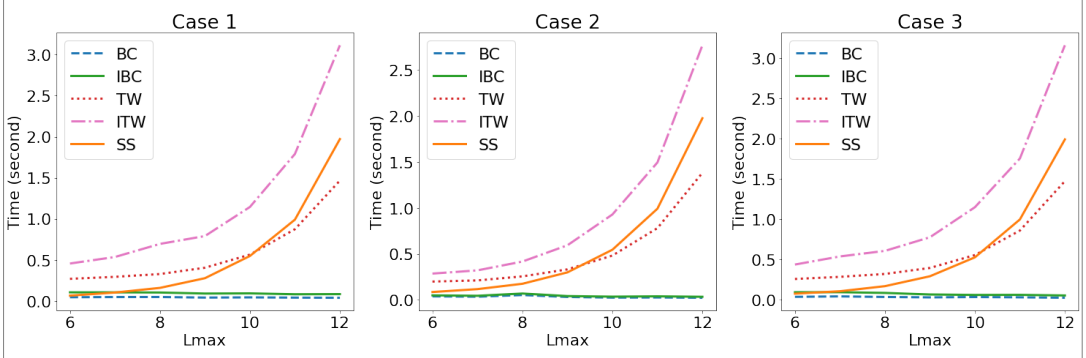

Fig S7. *The computational times. Informed Bayesian CART is generally slower than Bayesian CART due to the time calculating the proposal probabilities (e.g., (16)).*

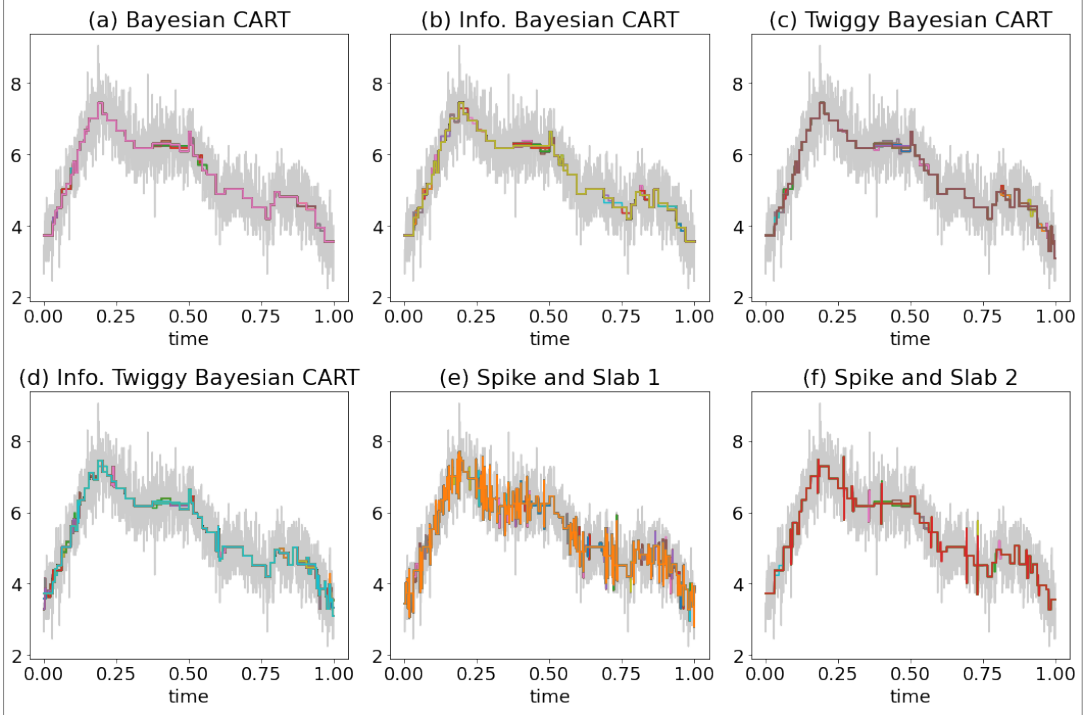

Fig S8. *The visualization of 1000 samples after 10,000 burn-in of the MCMC chains on Call Center Data. The gray lines are the data. (a) Bayesian CART (b) informed Bayesian CART (c) Twiggy Bayesian CART (d) informed Twiggy Bayesian CART (e) Spike-and-Slab (prior:* $p_{lk}^{ss,1} = 0.01$*) (f) Spike-and-Slab (prior:* $p_{lk}^{ss,2} = 0.01 \times 6^{-l/2}$*)*