On Mixing Rates for Bayesian CART

Jungeum Kim[*] and Veronika Ročková[†]

Booth School of Business, University of Chicago

Abstract

The success of Bayesian inference with MCMC depends critically on Markov chains rapidly reaching the posterior distribution. Despite the plentitude of inferential theory for posteriors in Bayesian non-parametrics, convergence properties of MCMC algorithms that simulate from such ideal inferential targets are not thoroughly understood. This work focuses on the Bayesian CART algorithm which forms a building block of Bayesian Additive Regression Trees (BART). We derive upper bounds on mixing times for typical posteriors under various proposal distributions. Exploiting the wavelet representation of trees, we provide sufficient conditions for Bayesian CART to mix well (polynomially) under certain hierarchical connectivity restrictions on the signal. We also derive a negative result showing that Bayesian CART (based on simple grow and prune steps) cannot reach deep isolated signals in faster than exponential mixing time. To remediate myopic tree exploration, we propose Twiggy Bayesian CART which attaches/detaches entire twigs (not just single nodes) in the proposal distribution. We show polynomial mixing of Twiggy Bayesian CART without assuming that the signal is connected on a tree. Going further, we show that informed variants achieve even faster mixing. A thorough simulation study highlights discrepancies between spike-and-slab priors and Bayesian CART under a variety of proposals.

1 Introduction

The advent of Markov Chain Monte Carlo (MCMC) has accelerated the widespread adoption of Bayesian methods in practice. Bayesian inference via MCMC simulation, however, depends critically on Markov chains reaching their stationary distribution reasonably fast. The folk wisdom is that MCMC is far slower than optimization and is only warranted when uncertainty quantification is desperately needed [38]. Positive findings have nevertheless been reported where rapid (polynomial) MCMC mixing times are, in fact, attainable in complex combinatorial problems (such as Bayesian variable selection [53]). This paper aims to create similar reasons for optimism (as well as caution) in the context for Bayesian tree-based regression.

[*] Jungeum Kim is a postdoctoral researcher at the Booth School of Business of the University of Chicago (jungeum.kim@chicagobooth.edu)

[†] Veronika Ročková is Professor in Econometrics and Statistics and James S. Kemper Faculty Scholar at the Booth School of Business of the University of Chicago (veronika.rockova@chicagobooth.edu). The author gratefully acknowledges support from the James S. Kemper Foundation Faculty Research Fund at the Booth School of Business as well as the National Science Foundation (DMS:1944740).

Bayesian tree-based regression (Bayesian CART of [13, 10] and BART of [11]) is one of the most popular machine learning tools in practice today. A host of frequentist theory now exists to certify their inferential validity [8, 45, 26, 34]. While estimation and inferential theory already exists, properties of MCMC approximations to these ideal inferential targets are conspicuously missing. This paper addresses computational properties of the Bayesian CART algorithm [13, 10] as opposed to statistical properties of the Bayesian CART posterior. We attempt to quantify (with lower and upper bounds) the speed at which practically used MCMC algorithms converge to the ideal inferential targets. Characterizations of MCMC mixing times for Bayesian CART (besides a lower bound in a recent independent paper [47]) have been unavailable.

There is an apparent disconnect between theory for optimization and sampling [38] and between theory for posterior distributions and their MCMC approximations. Bayesian CART implementations [13, 10] are instantiations of the Metropolis-Hastings (MH) algorithm [40] with local grow and prune proposal steps for addition or deletion of a node. The BART algorithm is essentially a Bayesian back-fitting extension where Bayesian CART is applied to the residuals for each individual tree. In spite of widespread popularity, difficulties in mixing have been reported [10, 52, 43, 33, 25]. Several enhancements have been proposed such as modifications of the proposal [52, 43], "warm start" initializations [24], or running multiple chains [7, 17]. This work attempts to characterize the computational bottlenecks of Bayesian CART and performs a comparative study of various proposal distributions in terms of mixing times.

Our computational complexity analysis builds on several fundamental papers studying Metropolis procedures [36, 18, 39]. Notably, [39] derive necessary and sufficient conditions for MH algorithms (with independent or symmetric proposals) to converge at a geometric rate to a prescribed continuous distribution. [3] study computational complexity of MCMC based on Metropolis random walks as both the sample and parameter dimensions grow to infinity for non-concave and possibly non-smooth likelihoods. We focus on a spectral bound approach suitable for Markov chains whose states are combinatorial structures. For finite-state Markov chains, the spectral gap can be bounded in terms of quantities associated with its graph [28, 14, 19]. Perhaps the first systematic approach to handling spectral bounds was developed in [31] using the conductance concept due to [9]. Conductance is a measure of edge expansion of the Markov chain, see [35] who proved the connection between conductance and convergence for the continuous state space. Lower bounds on the conductance, which give upper mixing bounds, are typically obtained by a technique of canonical paths where the idea is to find a set of paths such that no edge is very heavily congested. By using the canonical path argument, [53] show the rapid mixing of Bayesian variable selection. This bound is improved in [55] in the context of informed MCMC (that uses posterior information in the proposal) using the drift condition of [27] rooted in the coupling inequality [42, 32]. We consider locally informed proposals as well and, using similar drift conditions, we conclude linear mixing in n. Our work draws parallels between tree-based regression and structured wavelet shrinkage [8]. The wavelet representation of dyadic trees turns the tree selection problem into a variable selection problem with hierarchical constraints. The constraint creates certain reachability barriers and requires more sophisticated movements across the state space and a more careful design of canonical paths. In this work, we navigate the complex relationship between the MH proposal distribution and the mixing rate. While [53] used the deterministic stepwise selection algorithm as an inspiration to construct canonical paths, we use the CART algorithm [4] as an inspiration.

We primarily focus on a one-dimensional setting with dyadic splits (noting that non-dyadic CART can be analyzed in a similar manner as in [8]) where the MH proposal distribution consists of a simple attachment of a terminal node (GROW) or a detachment of two sister bottom nodes (PRUNE) [13, 10]. This algorithm is used in practice, for example in the context of uncertainty quantification for non-parametric regression with spatially varying smoothness [45]. Rapid mixing rate bounds in [53] and [55] critically rely on an asymptotic unimodality of the posterior distribution which can be translated in our context as model selection consistency. We first characterize sufficient conditions for tree selection consistency. Second, we show a negative result (an exponential mixing lower bound) where Bayesian CART fails at reaching deep isolated signals obscured by layers of noise. This motivates our proposal of Twiggy Bayesian CART, a new MH proposal distribution which attaches and deletes twigs (as opposed to individual nodes) to extend reachability. We show that Twiggy Bayesian CART attains polynomial mixing in non-parametric regression when the truth is a step function. To dilute the negative message about Bayesian CART, we show that it, in fact, achieves rapid mixing when the truth consists of wavelet signals that are connected along a tree. This is expected since myopic additions and deletions can reach deep signal through intermediate steps. It is interesting that the upper bound for Bayesian CART is then faster by a factor of n relative to spike-and-slab priors [53]. This may indicate smoothing benefits of tree-shaped regularization that avoids the addition of spurious high-resolution signals. Finally, using the two-drift condition argument [55], we show linear mixing of Markov chains under locally informed proposals.

Recently, independently from our work, [47] studied Bayesian CART with PRUNE and GROW movements in a multi-dimensional setting, where a lower bound that scales exponentially with n is shown exploiting the bottleneck that happens when one splits on a wrong variable early in the tree. Our work differs in several aspects: we exploit the wavelet formulation of trees to show consistency and upper bounds on mixing. The paper [47] only discusses a lower bound. Our lower bound is for the univariate case and focuses on the bottleneck that happens when deep signal is surrounded by noise.

The paper is structured as follows. In Section 2, we provide a brief review of the Bayesian CART and establish its tree selection consistency. The Twiggy Bayesian CART and the informed variations are introduced in Section 3. The theoretical framework and analysis of the mixing rates are presented in Section 4 and Section 5. The numerical study in Section 6 reinforces our theoretical findings on both simulation and real datasets. The paper concludes with Section 7.

2 Bayesian CART

Regression trees perform structured wavelet shrinkage [16, 8], where the underlying tree provides a skeleton for signal coefficients. This regression re-interpretation of Bayesian CART allows for straightforward implementations of the Bayesian CART algorithm through closed-form tree posterior probabilities.

2.1 Trees as Wavelets

We assume that observed continuous outcomes $Y = (Y_1, \dots, Y_n)'$ arise from

$$Y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i = 1, \dots, n = 2^{L_{max}+1} \tag{1}$$

where $\mathcal{X} = \{x_i = i/n : 1 \leq i \leq n\}$ are fixed observations on a regular[1] grid. We focus on wavelet reconstructions of f_0 using the standard Haar wavelet basis $\psi_{-10}(x) = I_{[0,1]}(x)$ and $\psi_{lk}(x) = 2^{l/2}\psi(2^l x - k)$ obtained with orthonormal dilation-translations of $\psi = I_{(0,1/2]} - I_{(1/2,1]}$. Denote with $\boldsymbol{X} = (x_{ij})$ the $(n \times p)$ regression matrix of $p = 2^{L_{max}} = n/2$ regressors constructed from Haar wavelets ψ_{lk} up to the maximal resolution L_{max}, i.e.

$$x_{ij} = \begin{cases} \psi_{-10}(x_i) = 1 & \text{for} \quad j = 1 \\ \psi_{lk}(x_i) & \text{for} \quad j = 2^l + k + 1. \end{cases} \tag{2}$$

We assume that the columns of $\boldsymbol{X}$ have been ordered according to the index $2^l + k$ (increasing ordering). We denote with $F_0 = (f_0(x_1), \dots, f_0(x_n))'$ the vector of realized values of the true regression function at design points. The non-parametric regression model (1) can be written in a matrix form

$$Y = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\nu}, \quad \text{where} \quad \boldsymbol{\nu} = F_0 - \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \text{ with } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n), \tag{3}$$

where $\boldsymbol{\beta}^*$ is an ordered vector of wavelet coefficients $\beta^*_{lk} = \langle \psi_{lk}, f_0 \rangle$. Bayesian dyadic CART (with splits at dyadic rationals) corresponds to tree-shaped wavelet reconstructions [8], as we re-iterate in Section 2.1.1 below.

Definition 1. (Tree) By a tree $\mathcal{T}$, we understand a collection of hierarchically organized nodes (l,k) such that $(l,k) \in \mathcal{T} \Rightarrow (j, \lfloor k/2^{l-j} \rfloor) \in \mathcal{T}$ for $j = 0, \dots, l-1$. We distinguish between two types of nodes: <u>internal</u> ones $\mathcal{T}_{int} = \{(l,k) \in \mathcal{T} : \{(l+1, 2k), (l+1, 2k+1)\} \in$

[1] The fixed grid assumption $x_i = i/n$ could be avoided using either unbalanced Haar wavelets [20] or regularity relaxations [45].

$\mathcal{T}\} \cup (-1, 0)$ and external ones $\mathcal{T}_{ext} = \mathcal{T} \backslash \mathcal{T}_{int}$ which are at the bottom of the tree. We define a set of pre-terminal nodes $\mathcal{P}(\mathcal{T}) = \{(l, k) \in \mathcal{T}_{int} : \{(l+1, 2k), (l+1, 2k+1)\} \in \mathcal{T}_{ext}\}$ as those internal nodes whose children are external. The null tree is defined as $\mathcal{T}_{null} = \{(-1, 0)\}$ and the full tree at the level L is defined as $\mathcal{T}^L_{full} = \{(l, k) : l < L\}$.

We will often denote with $\boldsymbol{\beta}_{\mathcal{T}} = (\beta_{lk} : (l, k) \in \mathcal{T}_{int})'$ the vector of ordered coefficients inside the tree[2] and with $\boldsymbol{\beta}_{\backslash\mathcal{T}}$ the complement. Similarly, for a given tree structure $\mathcal{T}$ we often split the design matrix $\boldsymbol{X}$ into active covariates $\boldsymbol{X}_{\mathcal{T}}$ (that correspond to $(l, k) \in \mathcal{T}_{int}$) and the complementary inactive ones $\boldsymbol{X}_{\backslash\mathcal{T}}$.

2.1.1 The Bayesian CART Posterior

The distinguishing feature of Bayesian CART, compared to selective wavelet reconstructions such as RiskShrink of [15], is that the pattern of sparsity has a tree structure. Namely, for a chosen maximal tree depth $L \leq L_{max}$, we assume the tree-shaped wavelet shrinkage prior [8]

$$\mathcal{T} \quad \sim \quad \Pi(\mathcal{T}) \tag{4}$$

$$\{\beta_{lk}\}_{l<L,k} \,|\, \mathcal{T} \;\sim\; \Pi(\boldsymbol{\beta}_{\mathcal{T}}) \otimes \bigotimes_{(l,k)\notin\mathcal{T}_{int}} \delta_0(\beta_{lk}). \tag{5}$$

Similarly as in [45], we consider the unit information g-prior $\boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}(0, g_n(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}})^{-1})$ with $g_n = n$ which coincides with the standard Gaussian prior $\Pi(\boldsymbol{\beta}_{\mathcal{T}}) = \prod_{(l,k)\in\mathcal{T}_{int}} \phi(\beta_{lk}; 0, 1)$ in regular designs.

The integral component of Bayesian CART is the tree prior $\Pi(\mathcal{T})$ over a set $\mathbb{T}_L$ of all trees up to the maximal chosen depth $L \leq L_{max}$. The Bayesian CART prior in [10] uses the heterogeneous Galton-Watson (GW) process (see Section 2.1 in [8] and [46]) with node split probabilities

$$p_{lk} = \mathbb{P}[(l, k) \in \mathcal{T}_{int}] \tag{6}$$

which need to be small in order to prevent the trees from growing indefinitely. While [10] suggest $p_{lk} = \alpha/(1+l)^\gamma$ for some $\alpha \in (0, 1)$ and $\gamma > 0$, we will assume that p_{lk} decays faster, potentially depending on n. Given p_{lk}, the tree prior probability for $\mathcal{T} \in \mathbb{T}_L$ satisfies

$$\Pi(\mathcal{T}) \propto \prod_{(l,k)\in\mathcal{T}_{int}} p_{lk} \prod_{(l,k)\in\mathcal{T}_{ext}} (1 - p_{lk}). \tag{7}$$

The conditional conjugacy of the Gaussian prior yields tractable posterior (up to multiplication) which is useful for Metropolis-Hastings implementations. In particular, for $\Sigma_{\mathcal{T}} = c_n(\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{X}_{\mathcal{T}})^{-1}$ with $c_n = n/(n+1)$ we have $\Pi(\mathcal{T} \,|\, Y) \propto \Pi(\mathcal{T}) \times N_Y(\mathcal{T})$, where

$$N_Y(\mathcal{T}) = \frac{\exp\left\{-\frac{1}{2}Y'[I - \boldsymbol{X}_{\mathcal{T}}\Sigma_{\mathcal{T}}\boldsymbol{X}'_{\mathcal{T}}]Y\right\}}{(2\pi)^{n/2}(1+n)^{|\mathcal{T}_{ext}|/2}}. \tag{8}$$

[2] there are $|\mathcal{T}_{ext}|$ of those

The Bayesian CART posterior has many favorable properties, such as near-minimax rate adaptation under the supremum loss [8] for α-Hölderian functions with $\alpha \leq 1$. This work focuses on computational (not statistical) properties of Bayesian CART. The mixing rate of the Bayesian CART MCMC algorithm [13, 10], however, ultimately depends on the structure of the underlying truth f_0. For clearer exposition of our findings, we focus on the following two assumptions on f_0, which are consonant with the tree (step function) model.

Assumption 1. *Assume that f_0 in (1) satisfies $f_0(x) = \sum_{(l,k)\in\mathcal{B}} \psi_{lk}(x)\beta^*_{lk}$, for some subset $\mathcal{B} \subseteq \{(l,k) : l < L\}$ such that $A \log n/\sqrt{n} < |\beta^*_{lk}| < C_{f_0}$ for all $(l,k) \in \mathcal{B}$ for some $A > 0$ and $C_{f_0} > 0$. Define $\mathcal{T}^* \in \mathbb{T}_L$ as the smallest tree that contains all signal nodes in $\mathcal{B}$ as internal nodes.*

(a) Assume that $\mathcal{B} \in \mathbb{T}_L$, i.e. $\mathcal{T}^*_{int} = \mathcal{B}$.

(b) Assume that $\mathcal{B} \notin \mathbb{T}_L$.

Remark 1. A class of tree-sparse functions compatible with Assumption 1 (a) is discussed in [2]. For example, signal discontinuity gives rise to a chain of large wavelet coefficients connected in the wavelet tree from the root to a leaf ([2], Figure 2). The connected signal property has been leveraged in a myriad of wavelet-based processing and compression algorithms [48, 12]. Assumption 1 (a) is intentionally optimistic in the sense that Bayesian CART is expected do well on a tree-shaped truth compared to, for example, spike-and-slab priors that do not have structured regularization. We will see this superiority in both our numerical as well as theoretical study.

Remark 2. Unlike previous investigations of Bayesian CART [45, 8], we do not assume Hölderian f_0 which alone does not guarantee tree selection consistency. Our results can be however replicated for structured Hölderian signals under suitable signal gap assumption for coefficients inside and outside $\mathcal{T}^*$.

An essential first step towards obtaining upper bounds on Markov chain mixing times is tree selection consistency. The following Theorem shows that under Assumption 1 the posterior concentrates on $\mathcal{T}^*$, the minimal tree spanning over signal. Similar consistency requirements (or log-concavity and asymptotic normality assumptions) have been required to obtain rapid convergence rate statements for Markov chains [1, 37, 53, 55]. While our theory has been derived for the regular fixed design, similar theoretical conclusions can be obtained also for fixed irregular design as in [45] using the unit information g-prior.

Theorem 1. *(Tree Selection Consistency) Assume the model (1), the Bayesian CART prior from Section 2.1.1 with $p_{lk} = n^{-c}$ for $c > 5/2$. Under Assumption 1 for large enough $A > 0$ we have with probability at least $1 - 4/n$*

$$\Pi(\mathcal{T}^* \mid Y) \geq 1 - \frac{1}{n^{c-5/2} - 1} - \frac{1}{n^{A^2/8 \log n}}.$$

Proof. Section 9.

Remark 3. In the context of Bayesian inference with phylogenetic trees, [41] show that when the data are generated by a mixture of two trees, many of the popular Markov chain take exponentially long to reach stationarity. Lemma 1 focuses on the less adverse situations when a single generative model is present that can be identified by the posterior.

Remark 4. The consistency result in Theorem 1 is different from posterior concentration rate results in [8] and [45] for Hölderian functions f_0 under the supremum loss. Due to the step function Assumption 1, we require a more aggressive split probability $p_{lk} = n^{-c}$ in Lemma 1 because we cannot leverage the decaying property of wavelet coefficients.

Much of the value of the optimality properties of the Bayesian CART posterior (e.g. adaptation to local smoothness [45] and frequentist validity of inference about certain f_0 [8]) hinges on the ability to approximate this posterior well.

2.1.2 The Bayesian CART Algorithm

The Bayesian CART algorithm is devised to explore the space of regression tree topologies by sequential sampling from the tree posterior distribution determined by (8). The two original algorithms [13, 10] are based on Metropolis-Hastings ideas with an accept-reject proposal mechanism consisting of four basic proposal moves (add a node, delete a node, change a variable and change a split-point). Many variations were later proposed with more intricate moves, such as tree rotations [23, 43], to better explore the tree space.

The Bayesian CART algorithm generates a chain of trees $\mathcal{T}^0, \mathcal{T}^1, \dots$ which will gravitate toward regions charged with posterior probability. Starting with an initial tree $\mathcal{T}^0$, transitions from $\mathcal{T}^i$ to $\mathcal{T}^{i+1}$ proceed in two steps: (1) generate a candidate value $\widetilde{\mathcal{T}}$ from a proposal distribution $S(\mathcal{T}^i \to \widetilde{\mathcal{T}})$ and (2) accept the proposal (i.e. $\mathcal{T}^{i+1} = \widetilde{\mathcal{T}}$) with a probability

$$\alpha(\mathcal{T}^i, \widetilde{\mathcal{T}}) = \min\left\{1, \frac{\Pi(\widetilde{\mathcal{T}} \,|\, Y) S(\widetilde{\mathcal{T}} \to \mathcal{T}^i)}{\Pi(\mathcal{T}^i \,|\, Y) S(\mathcal{T}^i \to \widetilde{\mathcal{T}})}\right\} \tag{9}$$

and set $\mathcal{T}^{i+1} = \mathcal{T}^i$ otherwise.

Under weak conditions (Section 7.4 of [44]), the sequence obtained by this algorithm will be an irreducible and aperiodic Markov chain with a limiting distribution $\Pi(\mathcal{T} \,|Y)$. Below, we will describe a dyadic one-dimensional version of Bayesian CART [10] which deploys a kernel $S(\mathcal{T}^i \to \widetilde{\mathcal{T}})$ that generates $\widetilde{\mathcal{T}}$ from $\mathcal{T}^i$ by randomly choosing among two steps (GROW and PRUNE). The algorithmic description of dyadic Bayesian CART we study is in Algorithm 1. We describe the algorithm using our wavelet tree representation.

The GROW movement chooses (uniformly at random) one terminal node, say $(\widetilde{l}, \widetilde{k})$, and splits it. In particular, we have $\widetilde{\mathcal{T}}_{int} = \mathcal{T}^i_{int} \cup \{(\widetilde{l}, \widetilde{k})\}$ and $\widetilde{\mathcal{T}}_{ext} = \mathcal{T}^i_{ext} \cup \{(\widetilde{l}+1, 2\widetilde{k}), (\widetilde{l}+1, 2\widetilde{k}+$

$1)\}\backslash\{(\widetilde{l},\widetilde{k})\}$ and

$$S_{GROW}(\mathcal{T}^i \to \widetilde{\mathcal{T}}) = \frac{1}{|\mathcal{T}^i_{ext}|}. \tag{10}$$

The PRUNE movement reverses the GROW move by choosing (uniformly at random) one pre-terminal node, $(\widetilde{l},\widetilde{k}) \in \mathcal{P}(\mathcal{T}^i)$, and by turning it into a terminal node. In particular, we have $\widetilde{\mathcal{T}}_{int} = \mathcal{T}^i_{int}\backslash\{(\widetilde{l},\widetilde{k})\}$ and $\widetilde{\mathcal{T}}_{ext} = \mathcal{T}^i_{ext}\cup\{(\widetilde{l},\widetilde{k})\}\backslash\{(\widetilde{l}+1,2\widetilde{k}),(\widetilde{l}+1,2\widetilde{k}+1)\}$ and

$$S_{PRUNE}(\mathcal{T}^i \to \widetilde{\mathcal{T}}) = \frac{1}{|\mathcal{P}(\mathcal{T}^i)|}. \tag{11}$$

Combining the two moves, dyadic Bayesian CART has the following proposal distribution

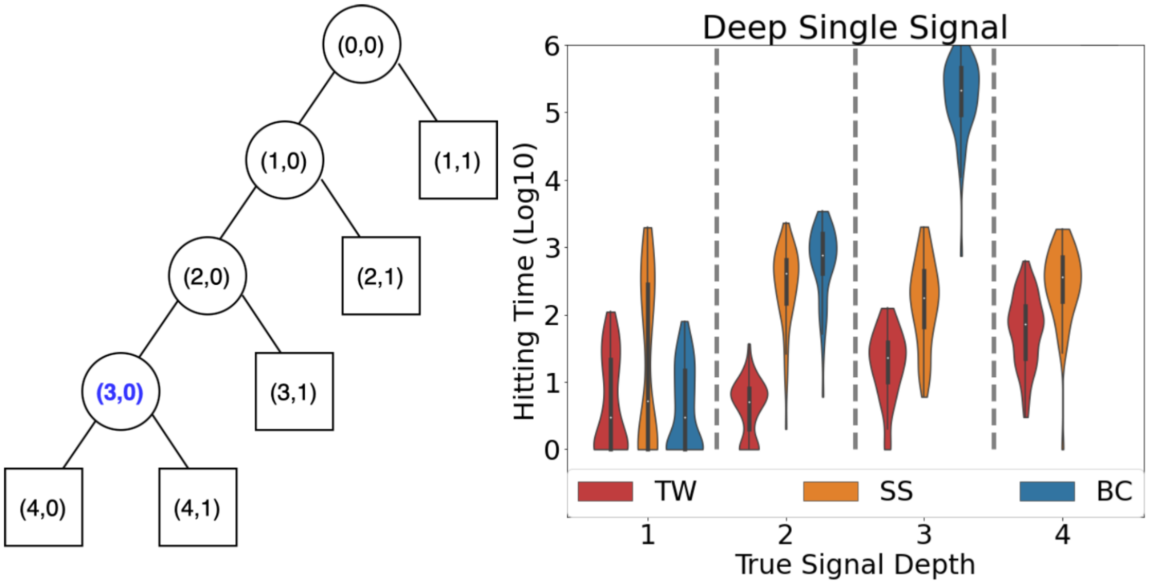
$$S(\mathcal{T} \to \widetilde{\mathcal{T}}) = \Gamma(\mathcal{T}) \times S_{GROW}(\mathcal{T} \to \widetilde{\mathcal{T}}) + [1-\Gamma(\mathcal{T})] \times S_{PRUNE}(\mathcal{T} \to \widetilde{\mathcal{T}}), \tag{12}$$

where $\Gamma(\mathcal{T})$ is the grow binary indicator with $P[\Gamma(\mathcal{T})=1] = 1/2$ for $\mathcal{T} \notin \{\mathcal{T}_{null}, \mathcal{T}^L_{full}\}$ and $P[\Gamma(\mathcal{T}_{null})] = 1 - P[\Gamma(\mathcal{T}^L_{full})] = 1$. The dyadic Bayesian CART algorithm was successfully deployed for estimating functions with spatially varying smoothness and for the construction of valid confidence sets [45]. Despite a simplified version of the full-blown Bayesian CART, this toy algorithm will give us many useful insights about computational bottlenecks. The GROW/PRUNE transition kernel performs only very local moves, not allowing bushy trees to be substantially restructured. This property makes this generic sampler susceptible to myopic encasement if initialized far away from high-posterior regions. While its poor mixing has been widely recognized in empirical studies [10, 52, 43, 33, 25], limited theoretical studies of the mixing times have been available [47].

3 Bayesian CART with a Twist

Local Metropolis-Hastings proposals are known to induce poor mixing [52, 43] which may result in misleading under-representations of uncertainty. In the context of trees, [23] remediate this issue by applying a rotation algorithm [50] while [43] proposes various elaborate moves for radical restructuring (see also [52]). Another way to prevent single trees from getting stuck is by adding them up and by performing Bayesian back-fitting (see the BART method of [11]). Alternatives to MH samplers have also been recently explored, see [29] for Sequential Monte Carlo approach and [30] for a particle Gibbs algorithm. We focus on the original Bayesian CART (dyadic version). One source of mixing issues for Bayesian CART is illustrated in a cautionary tale example below.

Example 1. (The Pitfalls of Bayesian CART) Consider $f_0 : [0,1] \to \mathbb{R}$ which satisfies Assumption 1 (b) where $\mathcal{B} = \{(j,0)\}$ and $\beta^*_{j,0} = 2$. We also assume $n = 2^{L_{max}+1}$ with $L_{max} = 8$. We consider the cases where the true signal depth grows, i.e. $j \in \{1,2,3,4\}$. We found that once the chain hits the signal node, it tends to stay around the minimal spanning tree $\mathcal{T}^*$ (plotted in Figure 1(a) for $j = 3$). Therefore, as a proxy to mixing time,



(a)

(b)

Figure 1: (a) An example of the minimal spanning tree $\mathcal{T}^*$ from Example 1 with oval internal and square external nodes. The signal node is marked in blue. (b) The hitting time of 50 chains initialized at $(0,0)$. TW: Twiggy Bayesian CART, SS: Spike-and-Slab, BC: Bayesian CART. The deeper the signal is, the slower (exponentially) the hitting time of Bayesian CART. We investigate this phenomenon theoretically in Section 5.1.

we measure the hitting time defined as $\tau = \min_{t\geq 0}\{\mathcal{B} \subset \mathcal{T}_{int}^{t}\}$. We run 50 chains for three algorithms: Bayesian CART, Twiggy Bayesian CART (to be introduced later), and Spike-and-Slab (one-site Metropolis-Hastings), where for all methods we use $p_{lk} = 0.1$. All chains are initialized at the root node $(0, 0)$. The violin plots of the hitting times are in Figure 1 (b). We see how the hitting time of Bayesian CART slows down exponentially as the depth of the signal increases. When the signal depth is 4, none of the 50 Bayesian CART chains hit within 1,000,000 iterations. In conclusion, Bayesian CART may not be able to capture signal if there are layers of noise separating the initialization and the signal. We will prove this theoretically later in Section 5.1. On the other hand, as Spike-and-Slab does not have a tree structure, its performance is consistent across different signal depth levels.

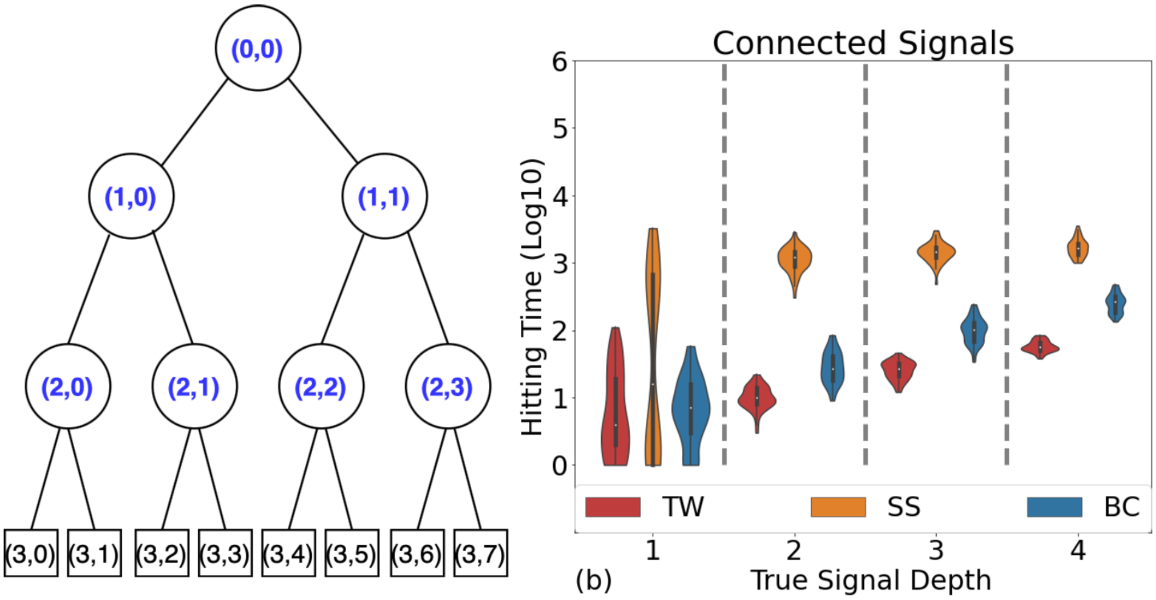
Example 1 may be unnecessarily pessimistic for Bayesian CART. The following example demonstrates that Bayesian CART actually mixes well when the signal is connected on a tree.

Example 2. (The Benefits of Bayesian CART) In contrast with Example 1, we now consider Assumption 1 (a) where $\mathcal{B} = \mathcal{T}_{full}^{j}$ for $j \in \{1, 2, 3, 4\}$ (plotted in Figure 2(a) for $j = 3$). We consider the same simulation settings as in Example 1. The violin plots of hitting times (for the entire set $\mathcal{B}$) in Figure 2 (b) show superiority of tree-shaped regularization where spike-and-slab takes longer to hit the entire group of connected signals. The stable increase of the hitting time of Bayesian CART is in sheer contrast with the exponential slowdown in Figure 1 (b). We investigate mixing of Bayesian CART theoretically for situations like this one in Section 5.2.

This work is not necessarily aimed at establishing the new methodological gold standard for MH tree proposal distributions. Instead, it is aimed at formalizing computational bottlenecks of Bayesian CART by performing a theoretical study of the default approach used in practice. During our theoretical investigation, however, one natural modification of the classical Bayesian CART resurfaced. In Figure 1(b), we showed a variant of Bayesian CART (called Twiggy Bayesian CART) which had more favorable hitting times. We now describe this new twist on an old classic. Later in Section 3.2 we describe another enhancement using locally informed proposals.

3.1 Twiggy Bayesian CART

To extend the reachability of trees in situations such as Example 1, we modify the GROW and PRUNE proposals. The GROW proposal attaches a twig to a chosen terminal node (rather than just splitting it into two nodes). The reverse move is then removing an entire branch (twig) in a tree rather than just collapsing two sibling bottom nodes. We call this variant Twiggy Bayesian CART. A twig is a portion of a tree that has at most one internal node for each level (as formalized below).



(a) (b)

Figure 2: (a) An example of the minimal spanning tree $\mathcal{T}^*$ from Example 2 with oval internal and square external nodes. (b) The hitting time of 50 chains initialized at $(0,0)$. TW: Twiggy Bayesian CART, SS: Spike-and-Slab, BC: Bayesian CART.

Definition 2. (Twig) For two distinct nodes (l,k) and (l',k') where (l,k) is an ancestor of (l',k') and $l \leq l' < L$, we define a twig $[(l,k) \leftrightarrow (l',k')] = \{(l,k), ...(l'-1, \lfloor k'/2 \rfloor), (l',k')\}$ as the collection of nodes on the unique shortest path connecting (l,k) and (l',k') in a full tree $\mathcal{T}^L_{full}$. A twig of length one is simply $[(l,k) \leftrightarrow (l,k)] = \{(l,k)\}$.

Definition 3. (Ancestors and Descendants) Given $\mathcal{T} \in \mathbb{T}_L$ and an internal node $(l,k) \in \mathcal{T}_{int}$, we define ancestors of (l,k) inside $\mathcal{T}$ as $A_{lk}(\mathcal{T}) = \{(l',k') \in \mathcal{T}_{int} : \exists j \in \{0,1,\dots,L-1\}\ s.t.\ (l',k') = (l-j, \lfloor k/2^j \rfloor)\}$. Descendants of (l,k) inside $\mathcal{T}$ are defined as $D_{lk}(\mathcal{T}) = \{(l',k') \in \mathcal{T}_{int} : (l,k) \in A_{l'k'}(\mathcal{T})\}$.

Given the current state $\mathcal{T}^i$, the GROW proposal of Twiggy Bayesian CART picks a node (l^*,k^*) in $\mathcal{T}^{L\,full}_{int} \backslash \mathcal{T}^i_{int}$ and grows a twig from an external node $(\widetilde{l},\widetilde{k}) \in \mathcal{T}^i_{ext}$ that is closest to (l^*,k^*). In particular, we have

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

If we considered a uniform proposal for (l^*,k^*), we would often pick a node from the deepest allowed layer $L-1$ (because there are 2^{L-1} of nodes). Instead, we penalize the inclusion of deep nodes by first picking a layer l^* from eligible layers $E^i = \{l < L : \exists (l,k) \notin \mathcal{T}^i_{int}\}$ with probabilities $d_{l^*} = D^{-l^*} / \sum_{l \in E_l} D^{-l}$ for $D > 1$ and by considering a uniform

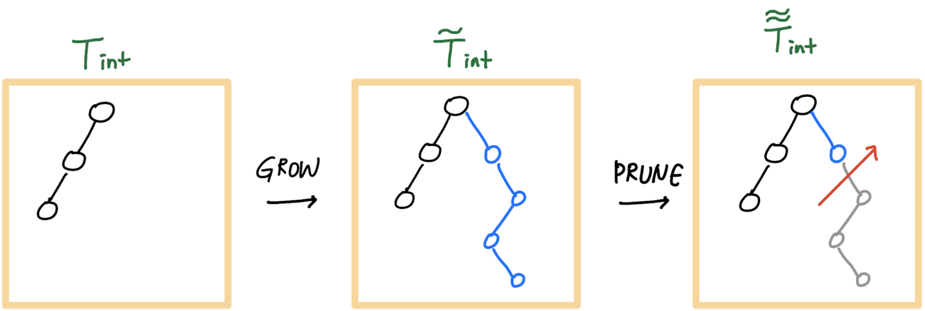


Figure 3: The Twiggy GROW and PRUNE.

proposal within the chosen layer, i.e.

$$\left(\frac{D-1}{D^L-1}\right)\frac{1}{2^{L-1}} \leq S_{GROW}(\mathcal{T}^i \to \widetilde{\mathcal{T}}) = \frac{d_{l^*}}{|\mathcal{K}_{l^*}|}, \tag{13}$$

where $\mathcal{K}_{l^*} = \{k : (l^*, k) \notin \mathcal{T}^i_{int}\}$. Note that larger D provides a stronger penalty that prevents the proposal from stretching towards nodes that are too deep.

The PRUNE step uniformly picks an internal node $(\widetilde{l}, \widetilde{k}) \in \mathcal{T}^i_{int}$ such that its entire branch below is a twig, i.e. $D_{lk} = [(\widetilde{l}, \widetilde{k}) \leftrightarrow (l^*, k^*)]$ for some $(l^*, k^*) \in \mathcal{T}^i_{int}$. The entire branch below the node is then removed to obtain $\widetilde{\mathcal{T}}$. In particular, we have

$$\widetilde{\mathcal{T}}_{int} = \mathcal{T}^i_{int} \backslash [(\widetilde{l}, \widetilde{k}) \leftrightarrow (l^*, k^*)]$$

Since the proposal candidates are all the nodes that have a twig below (including all per-terminal nodes $\mathcal{P}(\mathcal{T}^i)$), the proposal probability is bounded by

$$\frac{1}{|\mathcal{T}^i_{int}|} \leq S_{PRUNE}(\mathcal{T}^i \to \widetilde{\mathcal{T}}) \leq \frac{1}{|\mathcal{P}(\mathcal{T}^i)|}. \tag{14}$$

A cartoon of the twig proposals is depicted in Figure 3. It can be easily verified that the Twiggy Bayesian CART also yields an irreducible Markov Chain (i.e. all states communicate (see Section 6.3.1 in [44]). However, due to denser connectivity among trees we expect fewer bottlenecks.

3.2 Locally Informed Bayesian CART

The proposal distribution $S(\cdot \to \cdot)$ for Bayesian CART and Twiggy Bayesian CART ignores posterior information which might be useful in guiding the chain towards high-posterior zones. To accelerate MCMC over general discrete state spaces, [54] proposed locally informed proposal schemes that leverage posterior information in the vicinity of the current state $\mathcal{T}^i$ to propose the next state $\widetilde{\mathcal{T}}$. In particular, the proposal assigns a weight to each neighboring state $\mathcal{T}$ that depends on the posterior ratio $\Pi(\mathcal{T} \mid Y)/\Pi(\mathcal{T}^i \mid Y)$. Intuitively, we may expect that a large-posterior candidate is more likely to be accepted. Interestingly, [55] point out that this expectation is not always met and, as a remedy, threshold the posterior ratio in the proposal probability calculation. This approach is called LIT-MH (Metropolis-Hastings with Locally Informed and Thresholded proposal distributions). In the context of Bayesian variable selection, [55] show that LIT-MH significantly improves the mixing rate. Inspired by this finding, we also consider LIT-MH variants for Bayesian CART and Twiggy Bayesian CART and, later in Section 5.3, show that their mixing rate is linear in problem parameters.

Denote by $\mathcal{N}_g(\mathcal{T}^i) = \{\mathcal{T}' \supset \mathcal{T}^i : |\mathcal{T}'_{int} \backslash \mathcal{T}^i_{int}| = 1\}$ and $\mathcal{N}_p(\mathcal{T}^i) = \{\mathcal{T}' \subset \mathcal{T}^i : |\mathcal{T}^i_{int} \backslash \mathcal{T}'_{int}| = 1\}$ the GROW and PRUNE candidates from the current state $\mathcal{T}^i$ of the Bayesian CART algorithm. For these neighbor candidate trees, we define an intelligent movement rule instead of just a random walk. The proposal distribution for the LIT-MH proposal for Bayesian CART consists of

$$\begin{aligned} S_{GROW}(\mathcal{T}^i \to \widetilde{\mathcal{T}}) &= \frac{w_g(\widetilde{\mathcal{T}} \mid \mathcal{T}^i)}{Z_g(\mathcal{T}^i)} \mathbb{I}_{\mathcal{N}_g(\mathcal{T}^i)}(\widetilde{\mathcal{T}}), \\ S_{PRUNE}(\mathcal{T}^i \to \widetilde{\mathcal{T}}) &= \frac{w_p(\widetilde{\mathcal{T}} \mid \mathcal{T}^i)}{Z_p(\mathcal{T}^i)} \mathbb{I}_{\mathcal{N}_p(\mathcal{T}^i)}(\widetilde{\mathcal{T}}), \end{aligned} \tag{15}$$

where the weighting functions are defined for suitable $A > 0$ and $c > 3/2$ as

$$w_g(\widetilde{\mathcal{T}} \mid \mathcal{T}) = \frac{\Pi(\widetilde{\mathcal{T}} \mid Y)}{\Pi(\mathcal{T} \mid Y)} \wedge n^{(A^2 \log n)/8} \quad \text{and} \quad w_p(\widetilde{\mathcal{T}} \mid \mathcal{T}) \quad = 1 \vee \frac{\Pi(\widetilde{\mathcal{T}} \mid Y)}{\Pi(\mathcal{T} \mid Y)} \wedge n^{c-3/2}, \tag{16}$$

and the corresponding normalizing constants are

$$Z_g(\mathcal{T}) = \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} w_g(\widetilde{\mathcal{T}} \mid \mathcal{T}) \quad \text{and} \quad Z_p(\mathcal{T}) = \sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_p(\mathcal{T})} w_p(\widetilde{\mathcal{T}} \mid \mathcal{T}).$$

We call by Informed (Twiggy) Bayesian CART the variant with proposal probabilities in (15). The Informed Twiggy Bayesian CART has more neighbors $\mathcal{N}_g(\mathcal{T})$ and $\mathcal{N}_p(\mathcal{T})$ compared to the Informed Bayesian CART.

4 On Mixing Rates of Markov Chains

This section revisits several known facts about Markov chains whose states are combinatorial structures. In our work, bounds on mixing rates will be obtained by inspecting the eigenspectrum of the transition matrix. We denote with P the transition matrix on the state space $\mathbb{T}_L$ whose entries $P(\mathcal{T}_i, \mathcal{T}_j)$ quantify the probability of the move $\mathcal{T}_i \to \mathcal{T}_j$. For a given $\mathcal{T} \in \mathbb{T}_L$, we denote with $\mathcal{N}(\mathcal{T}) = \{\mathcal{T}' : S(\mathcal{T} \to \mathcal{T}') \neq 0\}$ the neighborhood of $\mathcal{T}$ consisting of all trees $\mathcal{T}'$ which can reach $\mathcal{T}$ in one step. Under the MH algorithm, we can write

$$P(\mathcal{T}, \mathcal{T}') = \begin{cases} S(\mathcal{T} \to \mathcal{T}')\alpha(\mathcal{T}, \mathcal{T}') & \text{if} \quad \mathcal{T}' \in \mathcal{N}(\mathcal{T}) \\ 0 & \text{if} \quad \mathcal{T}' \notin \mathcal{N}(\mathcal{T}) \cup \{\mathcal{T}\}, \\ 1 - \sum_{\widetilde{\mathcal{T}} \neq \mathcal{T}} P(\mathcal{T}, \widetilde{\mathcal{T}}) & \text{if} \quad \mathcal{T}' = \mathcal{T}. \end{cases}$$

where $\alpha(\cdot,\cdot)$ is the MH acceptance probability and $S(\cdot \to \cdot)$ is the proposal probability in (12). Moreover, the chain is reversible with respect to the probability distribution $\Pi(\mathcal{T} \mid Y)$ as it satisfies the detailed balance condition

$$Q(\mathcal{T},\mathcal{T}') \equiv \Pi(\mathcal{T} \mid Y)P(\mathcal{T},\mathcal{T}') = \Pi(\mathcal{T}' \mid Y)P(\mathcal{T}',\mathcal{T}) \quad \text{for all} \quad \mathcal{T},\mathcal{T}' \in \mathbb{T}_L.$$

This condition ensures that $\Pi(\cdot \mid Y)$ is the stationary distribution for P. It will be useful to associate the Markov chain with a weighted undirected graph on the vertex set $\mathbb{T}_L$ where the weight between two connecting (neighboring) vertices $\mathcal{T}$ and $\mathcal{T}'$ equals $Q(\mathcal{T},\mathcal{T}')$. We denote such a weighted undirected graph by G. Recall that two vertices $\mathcal{T}$ and $\mathcal{T}'$ are connected if and only if $Q(\mathcal{T},\mathcal{T}') > 0$. For an initial state $\mathcal{T}$ of the Markov chain at time $t = 0$, the total variation distance to the stationary distribution after t iterations satisfies

$$\Delta_{\mathcal{T}}(t) = \|P^t(\mathcal{T},\cdot) - \Pi[\cdot \mid Y]\|_{TV} \equiv \max_{S \subset \mathbb{T}_L} |P^t(\mathcal{T},S) - \Pi[S \mid Y]|, \tag{17}$$

where $P^t(\mathcal{T},S) \equiv \sum_{\mathcal{T}' \in S} P^t(\mathcal{T},\mathcal{T}')$ and where $P^t(\mathcal{T},\cdot)$ denotes the distribution of the state at time t with an initial condition $\mathcal{T}$. We now recall the formal definition of a mixing time.

Definition 4. The ϵ-mixing time of the Markov chain is defined as

$$\tau_\epsilon \equiv \max_{\mathcal{T} \in \mathbb{T}_L} \min\{t \in \mathbb{N} : \Delta_{\mathcal{T}}(t') \leq \epsilon \quad \text{for all} \quad t' \geq t\}, \tag{18}$$

where $\Delta_{\mathcal{T}}(t)$ is as in (17).

For an ergodic chain (whose states are aperiodic and positively recurrent), the rate of convergence to $\Pi(\cdot \mid Y)$ is governed by the spectral gap of P. Defining $\lambda_{max} = \max\{\lambda_1, |\lambda_{|\mathbb{T}_L|-1}|\}$, the spectral gap is defined as $Gap(P) = 1 - \lambda_{max}$. The following sandwich relation shows that the mixing time τ_ϵ and the spectral gap are related ([49], equation 2.9 in [51])

$$\frac{1 - Gap(P)}{2 \times Gap(P)} \log\left[\frac{1}{2\epsilon}\right] \leq \tau_\epsilon \leq \frac{\log[1/\min_{\mathcal{T} \in \mathbb{T}_L} \Pi(\mathcal{T} \mid Y)] + \log 1/\epsilon}{Gap(P)}. \tag{19}$$

For our theoretical study, we will work with a modified transition matrix (as suggested in [49]) which adds self-loops of weight 1/2 to each state. This so called "lazy" Markov chain does not significantly affect the mixing times. We denote with $\widetilde{P}$ the transition matrix of the original sampler and with $P \equiv \widetilde{P}/2 + I/2$ the modified matrix. This modification ensures that all eigenvalues are non-negative where the spectral gap satisfies $Gap(P) = 1 - \lambda_1$. Beyond the connection in (19), the second eigenvalue λ_1 (or the spectral gap) controls the information flow through the graph or, in other words, the conductance of the Markov chain.

4.1 Canonical Paths and Conductance

Some of the earliest spectral gap lower bounds were based on the concept of conductance [28]. In particular, Theorem 2 in [49] shows that in a reversible Markov chain

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

is the conductance which measures the ability of the chain to escape from any small region of the state space (and make rapid progress to equilibrium). The idea behind conductance is that chains with fewer bottlenecks will mix faster. While conductance can sometimes be estimated directly, in many applications the better approach to upper-bound the spectral gap is with edge overload on canonical paths [49].

Definition 5. (Canonical Path Ensemble) For any distinct pair of trees $\mathcal{T}, \mathcal{T}' \in \mathbb{T}_L$ we denote with $T_{\mathcal{T},\mathcal{T}'}$ a simple path running from $\mathcal{T}$ to $\mathcal{T}'$ through adjacent states in the state space graph G. A canonical path ensemble $\mathcal{E} = \{T_{\mathcal{T},\mathcal{T}'} : (\mathcal{T}, \mathcal{T}') \in \mathbb{T}_L \times \mathbb{T}_L\}$ is then a collection of such simple paths, one for each (ordered) pair of distinct vertices in G.

For any reversible Markov chain and any choice of a canonical path ensemble $\mathcal{E}$, the spectral gap of P can be lower-bounded with (Corollary 6 of [49])

$$Gap(P) \geq \frac{1}{l(\mathcal{E})\rho(\mathcal{E})}, \tag{20}$$

where $l(\mathcal{E})$ is the length of the longest path in $\mathcal{E}$ and

$$\rho(\mathcal{E}) = \max_{e \in \mathcal{E}} \frac{1}{Q(e)} \sum_{(\mathcal{T},\mathcal{T}'): e \in T_{\mathcal{T},\mathcal{T}'}} \Pi(\mathcal{T} \mid Y)\Pi(\mathcal{T}' \mid Y) \tag{21}$$

is the path congestion parameter. For the edge e in between two adjacent states $\mathcal{T}$ and $\mathcal{T}'$, the quantity $Q(e) \equiv Q(\mathcal{T}, \mathcal{T}') = \Pi(\mathcal{T} \mid Y)P(\mathcal{T}, \mathcal{T}')$ measures the natural capacity of the edge e or, in other words, how much traffic it would normally experience in the stationary state. The sum in (21) then counts the flow of the edge in the given family of canonical paths. The congestion is the maximum load of any edge of the state space graph as a fraction of its capacity. In order to find an upper bound on the mixing time using (19), in Section 5.2.1 we construct a canonical path ensemble and find a lower bound on the conductance (21).

5 Mixing Rates for Bayesian CART

This section presents some positive as well as negative findings for Bayesian CART in the context of Assumption 1. The signal assumptions (a) and (b) are qualitatively rather different and we will be able to appreciate the importance of less myopic proposals in the structure-less signal (b). Without the tree skeleton, local moves of Bayesian CART may not be able to reach all signals.

5.1 Bayesian CART Can Mix Poorly

We continue our cautionary tale from Example 1 showing that isolated signals are out of reach for initializations which need grow through noise to catch them. We now characterize the inability of the Markov chain to reach the posterior distribution reasonably (polynomially) fast. By finding an upper bound for the spectral gap in a counterexample f_0 constructed according to the Example 1, we show that the mixing lower bound increases exponentially in n.

Theorem 2. *Assume the model (1) with the Bayesian CART prior from Section 2.1.1 with $L \geq 2$ and $p_{lk} = n^{-c}$ with $c > 5/2$. There exists f_0 that satisfies Assumption 1 (b) such that, with probability at least $1 - 4/n$, the Bayesian CART mixing time satisfies for some $C > 1$*

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

This bound is exponential in n when $L = L_{max} \sim \log(n/2)$.

Proof. Section 10.

Remark 5. In work independent from ours, [47] provided a lower bound result showing that a simplified version of BART [11] mixes poorly (at least exponentially in n). In particular, [47] considered a single tree with prune and grow movements in a multi-dimensional setting. The key idea behind the lower bound is that the first split direction causes a serious bottleneck. To move between two trees that differ in their first split direction, one must prune all the way up to the root tree to replace the first split. We consider a perhaps more simplified scenario by exploiting wavelet representations and we show slow mixing even in a one-dimensional setting.

5.2 Bayesian CART Can Mix Well

We now establish sufficient conditions for classical Bayesian CART to mix "well", i.e the number of iterations required to converge to an ϵ-ball of the stationary distribution grows only polynomially in the problem parameters. We will inspect various components of the sandwich relation presented earlier in (19).

The following theorem provides a polynomial upper bound for the speed of MCMC convergence of classical Bayesian CART for connected signals (Assumption 1 (a)).

Theorem 3. *Assume the model (1) with the Bayesian CART prior with $p_{lk} = n^{-c}$ with $c > 5/2$. Under Assumption 1 (a) with a large enough constant $A > 0$, with probability at least $1 - 4/n$ the Bayesian CART algorithm from Section 2.1.2 satisfies*

$$\tau_\epsilon \leq 2^{2L+3}\left\{n\left[\left(c+\frac{1}{2}\right)\log(1+n) + |\mathcal{T}^*_{int}|C^2_{f_0} + 1\right] + 4\,|\mathcal{T}^*_{int}|\log n + \log\left(\frac{2}{\epsilon}\right)\right\} \tag{22}$$

where $C_{f_0} > 0$ is the constant from Assumption 1.

Proof. See Section 12.

Remark 6. In the bound (22), we intentionally separated the influence of model complexity (captured by the maximal allowed depth L) and the sample size n. In practice, the most reasonable choice for L is the maximal allowed resolution $L = L_{max}$ which will give us cubic mixing in n. If we were confident that the posterior over trees deeper than L goes to zero, we can always devise a Markov chain with a smaller state space (trees up to level L). For example, for α-Hölderian function choosing $L \propto (n/\log n)^{1/(2\alpha+1)}$ yields $\mathbb{P}_{f_0}(\mathbb{T}_L^c) = o(1)$.

Remark 7. (Comparison with Spike-and-Slab) The tree-structured signal in Assumption 1 (a) is particularly flattering for Bayesian CART. It is interesting to compare this approach with a spike-and-slab prior and a one-site Metropolis-Hastings proposal. [53] showed rapid mixing of the MH algorithm in a high-dimensional linear model (i.e. (3) with p covariates and $\boldsymbol{\nu} = \boldsymbol{\varepsilon}$) and a g-prior (which coincides with our prior in orthogonal designs). We attempt to rephrase their result in the context of our wavelet regression matrix where $p = n/2$. The point-mass spike-and-slab prior in [53] assumes $\Pi(\boldsymbol{\gamma}) \propto \left(\frac{1}{p}\right)^{\kappa|\boldsymbol{\gamma}|} \mathbb{I}[|\boldsymbol{\gamma}| \leq s_0]$, where s_0 is a chosen upper bound on the model complexity and κ is a model-size penalty. We show (Theorem 7 in the Supplement Section 15) that our upper bound on Bayesian CART mixing time is tighter by n than the upper bound for Spike-and-Slab MH due to a tighter bound on the congestion parameter.

The proof of Theorem 3 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 show Lemma 1 and Lemma 2 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 Lemma 1. In the next section, we provide details of the canonical path construction and describe basic properties of our canonical path ensemble. Similarly as in [53], whose canonical path architecture was inspired by stepwise variable selection, our construction was inspired by the CART algorithm [4].

5.2.1 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 $\mathcal{T}^*$ $\underline{\text{along edges}}$ in the graph with a transition matrix P. To this end, we introduce the $\underline{\text{transition function}}$ $\mathcal{G} : \mathbb{T}_L \backslash \mathcal{T}^* \to \mathbb{T}_L$ 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 $\underline{\text{“closer”}}$ to $\mathcal{T}^*$, where closeness is determined by the Hamming distance $h(\mathcal{T}, \mathcal{T}^*)$ between binary tree encodings[3]. 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

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

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

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.

(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

$$\mathcal{T}^-_{int}=\mathcal{T}_{int}\backslash\{(l,k)\}\quad\text{and}\quad\mathcal{T}^-_{ext}=\mathcal{T}_{ext}\backslash\{(l+1,2k),(l+1,2k+1)\}\cup\{(l,k)\}\tag{23}$$

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

(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

$$\mathcal{T}^+_{int}=\mathcal{T}_{int}\cup\{(l,k)\}\quad\text{and}\quad\mathcal{T}^+_{ext}=\mathcal{T}_{ext}\cup\{(l+1,2k),(l+1,2k+1)\}\backslash\{(l,k)\}\tag{24}$$

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

(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 (23).

Definition 6. 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 $T_{\mathcal{T}\backslash\mathcal{T}'}:=T_{\mathcal{T},\mathcal{T}^*}\backslash(T_{\mathcal{T},\mathcal{T}^*}\cap T_{\mathcal{T}',\mathcal{T}^*})$[4].

Below, we characterize important properties of $\mathcal{E}$ which are instrumental in the sandwich relation (19) and in the proof of Theorem 3.

Lemma 1. *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 $\ell(\mathcal{E})\equiv\max\limits_{\mathcal{T},\mathcal{T}'\in\mathbb{T}_L}|T_{\mathcal{T},\mathcal{T}'}|\leq 2^{L+1}$.*

Proof. See Section 11.1.

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

[4] By construction each step in $T_{\mathcal{T},\mathcal{T}^*}$ reduces the Hamming distance, and thus we can show similarly to [53] that $\mathcal{E}$ is an ensemble of <u>simple</u> paths.

Lemma 2. *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 11.2.

5.3 Twiggy Bayesian CART Mixes Well

In Section 5.2 we established encouraging results for Bayesian CART with PRUNE and GROW steps under the connected signal Assumption 1 (a). We have also seen that under Assumption 1 (b), where signals are not connected, Bayesian CART can mix poorly (Theorem 2). We now investigate mixing of Twiggy Bayesian CART in the context of the unstructured signal in Assumption 1 (b). Moving from Bayesian CART to Twiggy Bayesian CART extends signal reachability where trees can become more competitive with the Spike-and-Slab approach [53] when signal is obscured by layers of noise.

Theorem 4. *Assume the model* (1) *with the Bayesian CART prior from with* $p_{lk} = n^{-c}$ *and* $c > 5/2 + \log D$*. Under Assumption 1 (b) with* $|\mathcal{T}^*_{int}| \lesssim \log^2 n$ *and large enough* $A > 0$*, the Twiggy Bayesian CART algorithm in Section 3.1 (with* $D > 1$*) satisfies with probability at least* $1 - 4/n$

$$\tau_\epsilon \leq \frac{(D^L - 1)}{D - 1} \times 2^{2L+3} \left\{ n \left[\left(c + \frac{1}{2} \right) \log(1+n) + |\mathcal{T}^*_{int}| C^2_{f_0} + 1 \right] + 4\, |\mathcal{T}^*_{int}| \log n + \log\left(\frac{2}{\epsilon} \right) \right\} \tag{25}$$

Proof. See Section 13.

5.4 Locally Informed Versions Mix Even Better

For the informed versions of Bayesian CART and Twiggy Bayesian CART described in Section 3.2, we obtain the following upper bound on the mixing time that is only linear in 2^L, where $L \leq L_{max}$ is required to go to infinity as $n \to \infty$. This speedup is likely a consequence of the posterior-informed proposal defined in (15).

Theorem 5. *Assume the model* (1) *and the Bayesian CART prior with* $p_{lk} = n^{-c}$ *and* $c > 3$*. Consider the Twiggy Bayesian CART with an informed proposal in* (15)*. Under Assumption 1 (a) or (b), for a large enough constant* $A > 0$ *and* $L \leq L_{max}$ *such that* $L \to \infty$ *as* $n \to \infty$*, we have with probability at least* $1 - 4/n - \mathrm{e}^{-n/8}$*,*

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

For the informed Bayesian CART, the same bound holds but only under Assumption 1 (a).

Proof. See Section 14.

Remark 8. Theorem 5 provides at most linear in n mixing for Bayesian CART only under Assumption 1 (a). The exponential mixing rate lower bound in Theorem 2 still applies to the informed Bayesian CART[5]. Therefore, the proposal informativeness alone does not solve the myopic problem of Bayesian CART.

Remark 9. It is worthwhile to point out that the linear mixing in Theorem 5 is truly a consequence of the informed proposal as opposed to the proving technique (two-drift condition as opposed to canonical path argument). By using the two-drift condition proving technique (for $c \geq 4$ and $D \leq \mathrm{e}$), as opposed to the canonical path argument, we can slightly improve the mixing rate upper bound in (22) for Bayesian CART and for Twiggy Bayesian CART (the original non-informed versions) to $\tau_\epsilon \lesssim \log(6/\epsilon) \times \max\left(\frac{C_{f_0}^2}{\delta_1 A^2} \frac{2^{2L} n}{\log^2 n}, 2^{2L+1}\right)$, where $\delta_1 = 1$ for the Bayesian CART, and $\delta_1 = \frac{2(D-1)}{D^L - 1}$ for the Twiggy Bayesian CART. Compared with the bound (22) obtained by the canonical path argument, this bound has a slight improvement by a logarithmic factor when $|\mathcal{T}_{int}^*|$ is fixed. For more explanation, see the discussion in Remark 11. The proof is in Section 14.3.

The proof of Theorem 5 rests on the two drift condition argument developed by [55]. In the next section, we provide details of the two drift functions chosen for our tree regression setting.

5.4.1 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 [55]. 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 [53], [55] 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.

[5] See that (45) in the proof is valid as long as the proposal neighbor is the same as the Bayesian CART.

Definition 7. We define two drift functions as

$$V_1(\mathcal{T}) = \exp\left\{\frac{1}{2^L\,(C_{f_0}+2)^2\,(n+1)}\left(Y'(I - P_{\mathcal{T}}/n)Y\right)\right\}, \tag{26}$$

$$V_2(\mathcal{T}) = \exp\left\{\frac{1}{2^L}\left(|\mathcal{T}_{int}\backslash\mathcal{T}^*_{int}| + (|\mathcal{T}^*_{int}\backslash\mathcal{T}_{int}| \wedge 1) \times (2^L - |\mathcal{T}_{int} \cup \mathcal{T}^*_{int}|)\right)\right\}, \tag{27}$$

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 10. 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.

Proposition 1. *Under the same assumptions of Theorem 5, 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 14.2.

6 Performance Evaluation

We compare Bayesian CART, Twiggy Bayesian CART (with $D = 2$) and their informed versions[6] on simulated data as well as a real dataset. To appreciate the effect of tree-shaped regularization, we also compare Bayesian CART with the Metropolis-Hastings one-site sampler for Spike-and-Slab priors.

[6] The upper thresholds of the informed algorithms in (16), we use e^{10} without tuning.

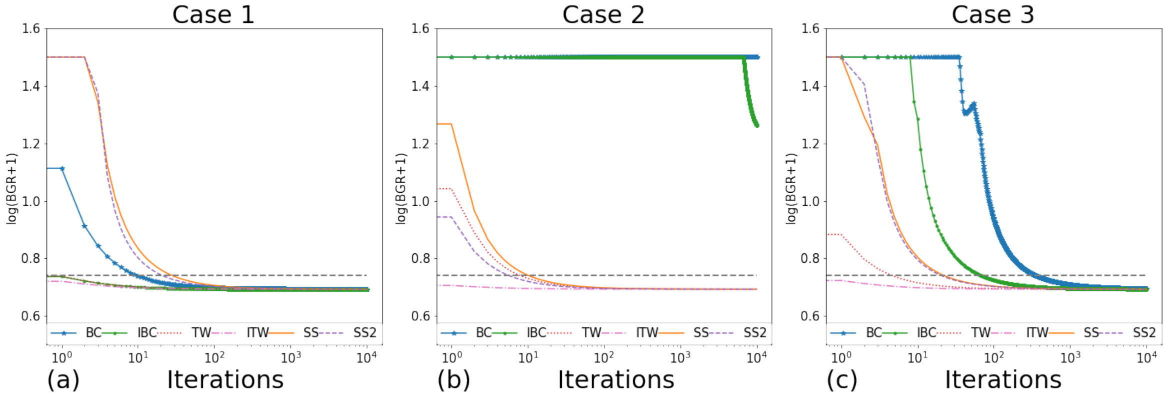


Figure 4: The local BGRs (log-transformed) when $n = 2^7$. The local BGR values tend to decrease during the course of MCMC sampling. We have capped the values at a threshold $log(Y + 1) = 1.5$ for clearer visualization. The horizontal grey dotted line corresponds to the local BGR value 1.1. (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: original and informed Twiggy Bayesian CART, SS and SS2: Spike-and-Slab with prior p_1^{ss} and p_2^{ss} respectively.

6.1 Simulation Study

Data We generate simulated data from the model (1) with three true signal skeletons. Given the skeleton, all true coefficients are set equal to 2. (1) Case 1 (fully connected signal) is a full tree of internal depth 3, i.e. $\mathcal{T}_{int}^* = \mathcal{B} = \{(-1, 0)\} \cup_{l=0}^{3} \cup_{k=0}^{2^l-1} \{(l, k)\}$. (2) Case 2 is a single, disconnected and isolated deep signal at $\mathcal{B} = \{(4, 0)\}$. (3) Case 3 is a mixed signal consisting of several isolated nodes with $\mathcal{B} = \{(2, 0), (2, 3), (3, 2), (3, 3), (3, 4), (3, 5), (4, 15)\}$. Recall that $\mathcal{T}^*$ is the smallest tree that includes $\mathcal{B}$ as its internal nodes.

Prior As the split probability in (7) for Bayesian CART (using $L = L_{max}$), we use $p_{lk} = \alpha n^{-c}$, which was used in our theoretical studies up to a constant factor. We choose α so that $p_{lk} = 0.25/2^{L_{max}-6}$ for $L_{max} \in [6, ..., 11]$. For the Spike-and-Slab prior, we consider $\frac{\Pi(\mathcal{T} \cup (l,k))}{\Pi(\mathcal{T})} = p_l^{ss}$, where we consider two cases $p_1^{ss} = \alpha n^{-c} = 0.25/2^{L_{max}-6}$ and $p_2^{ss} = 0.01\, n^{1/4} 6^{-l}$. Note that p_2^{ss} penalizes deep node inclusion more strongly than p_1^{ss}. As shown Figure 15 in Section 16, several MCMC runs on Case (3) reveal that p_1^{ss} has a higher acceptance rate but p_2^{ss} converges faster to the true signals. The two chosen split probabilities satisfy the sufficient conditions studied in [45] with which a Spike-and-Slab algorithm can achieve locally adaptive minimax rate.

Performance measure We first define a proxy for the mixing time defined in (18) using BGR (Gelman-Rubin diagnostic) [21]. BGR measures the difference between in-chain and across-chain variability when considering multiple initializations. Formally, consider a collection of K chains $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_K\}$. Denoting the L tree samples from each $\mathcal{C}_k$ by $\{\mathcal{T}_1^{\mathcal{C}_k}, ..., \mathcal{T}_L^{\mathcal{C}_k}\}$

and the j-th coefficient sample from tree $\mathcal{T}_i^{\mathcal{C}_k}$ by $\beta_j(\mathcal{T}_i^{\mathcal{C}_k})$, the BGR for β_j is

$$\mathrm{BGR}(\beta_j | \{\mathcal{T}_1^{\mathcal{C}_k}, ..., \mathcal{T}_L^{\mathcal{C}_k}\}_{k=1}^K) = \frac{\frac{L-1}{L} W_j + \frac{1}{L} B_j}{W_j},$$

where $W_j = \frac{1}{K}\sum_{k=1}^K \frac{1}{L-1}\sum_{i=1}^L (\beta_j(\mathcal{T}_i^{\mathcal{C}_k}) - \bar{\beta}_{jk})^2$ and $B_j = \frac{L}{K-1}\sum_{k=1}^K (\bar{\beta}_{jk} - \bar{\beta}_j)^2$, given the in-chain and between-chain coefficient means $\bar{\beta}_{jk} = \frac{1}{L}\sum_{i=1}^L \beta_j(\mathcal{T}_i^{\mathcal{C}_k})$ and $\bar{\beta}_j = \frac{1}{K}\sum_{k=1}^K \bar{\beta}_{jk}$. To reduce the computational cost of monitoring the BGRs for a large L, we modify the BGR as follows. First, we measure the BGR of the estimated coefficient $\hat{\beta}_j(\mathcal{T}_i^{\mathcal{C}_k})$ instead of the sampled $\beta_j(\mathcal{T}_i^{\mathcal{C}_k})$. Given a tree $\mathcal{T}$, the estimated coefficient of the j^{th} node $\hat{\beta}_j(\mathcal{T}) = (X_j' X_j)^{-1} X_j' Y \times \mathbb{I}_{(l_j, k_j) \in \mathcal{T}_{int}}$ only depends on whether the j^{th} node is included in the tree. Therefore, it can be seen that the BGR of $\hat{\beta}_j$ is equal to the BGR of an inclusion indicator, denoted by $I_j \in \{0, 1\}$. Second, because BGR's tend to be smaller for noise coefficients, we monitor the BGR of signal nodes which have larger BGR values in general. Last, we compute BGRs locally; At time t, we consider the most recent 100 samples to calculate BGR. Denote a t^{th} sample of chain $\mathcal{C}_k$ by $\mathcal{T}_t^{\mathcal{C}_k}$. Given $K = 10$ chains, a local BGR of the j^{th} indicator I_j at time t is defined as

$$\mathrm{BGR}(j, \mathcal{C}|t) = \mathrm{BGR}(I_j | \{\mathcal{T}_t^{\mathcal{C}_k}, \mathcal{T}_{t-1}^{\mathcal{C}_k}, ..., \mathcal{T}_{t-99}^{\mathcal{C}_k}\}_{k=1}^{10}). \tag{28}$$

As shown in Figure 4, as the iteration proceeds, the $\mathrm{BGR}(j, \mathcal{C}|t)$ values tend to decrease during the course of MCMC sampling. With this empirical observation, we define a proxy of the mixing time called BGR α-time by

$$\tau_\alpha^{\mathrm{BGR}} = 10^6 \wedge \arg\min_{t \geq 0} \{\max_{j \in S} \mathrm{BGR}(j, \mathcal{C}|t) \leq \alpha\}, \tag{29}$$

where S is the index set of true signals in the data generating mechanism. The inner maximum $\max_{j \in S} \mathrm{BGR}(j, \mathcal{C}|t)$ quantifies the mixing quality of the chains at time t. In this paper, we consider $\alpha = 1.1$, as this number used widely as a criterion of mixing [21].

We threshold the mixing time at 10^6 because beyond this number, it is hard to see that a chain mixes in a reasonable time. In short, we measure the BGR in (29) for every 100 out of $1,000,000$ iterations. Note that the chain may meander towards a poor local neighborgood, far from $\mathcal{T}^*$. Therefore, to quantify the quality of each tree $\mathcal{T}$, we can think of $(l, k) \in \mathcal{T}_{int} \backslash \mathcal{T}_{int}^*$ as a <u>false positive</u> and $(l, k) \in \mathcal{T}_{int}^* \backslash \mathcal{T}_{int}$ as a <u>false negative</u>. A natural quality measure for trees is the F1 score, the harmonic mean of precision and recall (a low precision indicates that the model is overfitting, while a low recall indicates that the model is underfitting). If F1 equals 1 (i.e. both precision and recall are 1) then the tree equals $\mathcal{T}^*$. The F1 values are obtained from the last 100 iterations of each chain. Note that <u>not</u> all nodes in $\mathcal{T}^*$ are necessarily signals. Therefore, when we calculate F1 for Spike-and-Slab, we consider only the true signals $\mathcal{B}$ as the true model. In a similar spirit, we also measure hitting time defined by $\tau_{hit} = \inf_{t \geq 0}\{\mathcal{T}_t = \mathcal{B}\}$. Lastly, we also present the acceptance rates, which may help understand the stickiness of Markov chains.

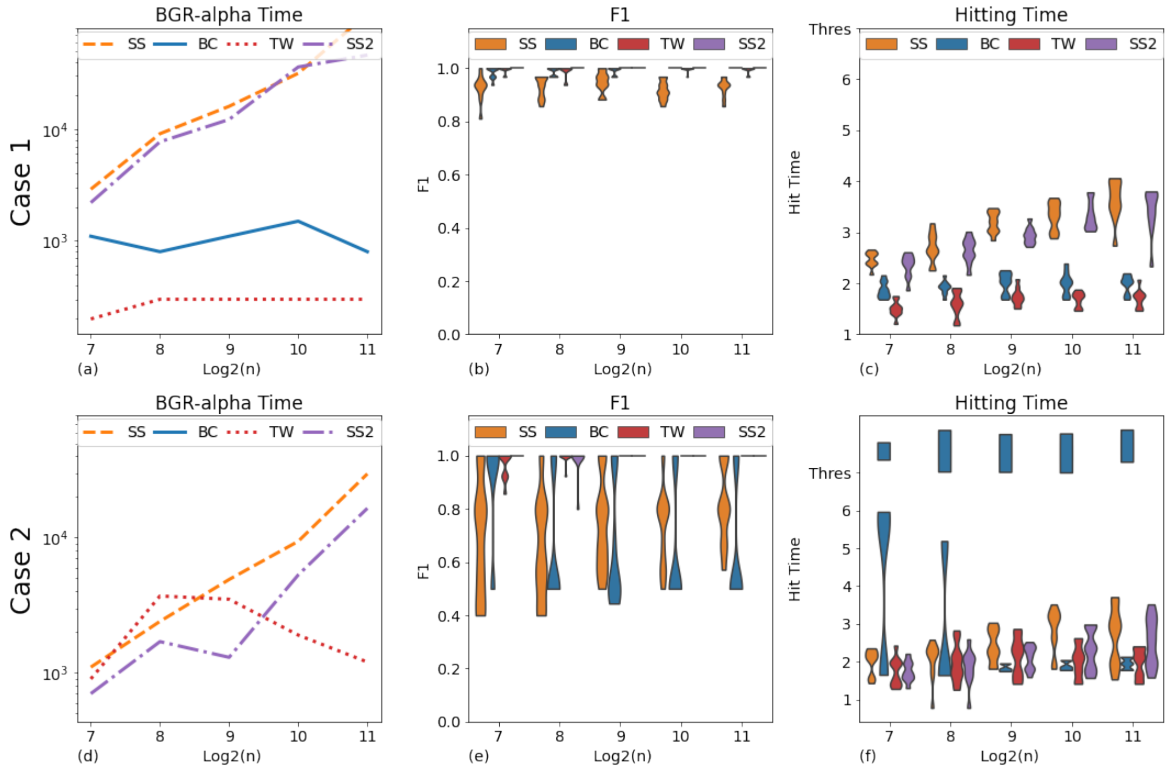


Figure 5: Plots (a) and (d) depict time for achieving the worst case local BGR below 1.1. Plots (b) and (e) show F1 scores. The small F1 value of Spike-and-Slab with p_1^{ss} is due to low precision (overfit). Plots (c) and (f) show hitting times. (Legend) BC: Bayesian CART, TW: Twiggy Bayesian CART, SS and SS2: Spike-and-Slab with prior p_1^{ss} and p_2^{ss} respectively.

6.1.1 The Effect of Signal Structure

Connected signals. The results on Case (1) numerically affirm the sufficient condition for rapid mixing of Bayesian CART in Theorem 3, which says the Bayesian CART mixes rapidly when all signals are connected. By increasing the size of data from $n = 2^7$ to $n = 2^{11}$ (L_{max} from 6 to 10), we measure the BGR-α time $\tau_\alpha^{\mathrm{BGR}}$ for $\alpha = 1.1$ in (29) and F1 as well as τ_{hit}. We run 10 chains for Bayesian CART, Twiggy Bayesian CART and Spike-and-Slab (both p_1^{ss} and p_2^{ss}). For each method, 10 chains are initialized with randomly generated trees. The result is in Figure 5 (a), (b), and (c). We see that Bayesian CART hits the true tree faster than Spike-and-Slab. In addition, we see that Bayesian CART achieves a good BGR-α time as the sample size increases. We observe that Bayesian CART is enjoying this favorable property over the Spike-and-Slab in two ways. First, we see that $\tau_\alpha^{\mathrm{BGR}}$ of Bayesian CART increases more slowly than Spike-and-Slab, and second, the hitting time in Figure 5 (c) is superior over that of Spike-and-Slab. Further investigation revealed that the source of the low F1 values of Spike-and-Slab with p_1^{ss} was low precision i.e., it often overfitted. We notice that the deeper penalty through p_2^{ss} (as opposed to) p_1^{ss} improves Spike-and-Slab in all the performance measures (F1 and BGR-α time, and the hitting time). Lastly, we observe that Twiggy Bayesian CART does similarly well as Bayesian CART. Note that in Figure 4, when $n = 2^7$, the local BGRs of the Twiggy Bayesian CART decrease faster compared to Bayesian CART.

Disconnected signals. Now we numerically affirm the exponential lower bound of Bayesian CART in Theorem 2 in the context of deep isolate signals as in Example 1. On the other hand, Theorem 4 says that Twiggy Bayesian CART still mixes rapidly. These theoretical results are affirmed by the results on Case (2) in Figure 5 (d), (e), and (f). In terms of $\tau_\alpha^{\mathrm{BGR}}$, we see that Twiggy competes with Spike-and-Slab and then performs better when the sample size becomes larger. Bayesian CART did not achieve local BGR under 1.1 in (29) in a given time range (1,000,000 iterations). This is why there is no line for Bayesian CART in Figure 5 (d). Similarly, the hitting time would have to exceed the maximum number of $1,000,000$ iterations. This is marked by the histogram bars above the maximum allowed number of iterations. Besides the BGR-α time displayed in Figure 5 (d), in Section 16 we additionally display the smallest local BGRs in Figure 11 (e). This value is obtained by running all the chains for 10^6 iterations and by taking the minimum over local BGRs defined in (28) for each chain. We observe that the minimum local BGR of Bayesian CART is exceedingly large. This is related to the small F1 of Bayesian CART; further investigation revealed that low Recall values made F1 small, indicating underfit of Bayesian CART. On the other hand, Spike-and-Slab achieves local BGR smaller than 1.1, and using p_2^{ss} resolves the overfitting problem of p_1^{ss} as in Figure 5 (e). However, when n increases, even using p_2^{ss} does not bring up Spike-and-Slab to the speed of Twiggy Bayesian CART in terms of BGR-α time and hitting time (Figure 5 (d) and (f)).

6.1.2 The Effect of Informed MCMC

We repeat the analyses to investigate the improvement brought by the posterior-informed proposals. From Figure 6 and Figure 14 in Section 16, we see that informed versions overall improve on their non-informed counterparts. The BGR-α (Figure 6 (a) and (d), and Figure 14 (a)) and hitting times (Figure 6 (c) and (f), and Figure 14 (c)) become faster. However, in terms of hitting time, the informed Twiggy Bayesian CART does not outperform its non-informed version. This is related to that the informed Twiggy Bayesian CART tends to overfit, thereby having smaller F1 values (Figure 6 (b) and (e), and Figure 14 (b)). It might be understood as a trade-off of having a higher acceptance rate than Twiggy Bayesian CART (Figure 13). However, the informed Bayesian CART also has an increased acceptance rate, but it does not result in overfitting. We think that the decreased precision of informed Twiggy Bayesian CART is due to the combination of more flexible movement and the larger acceptance rate (e.g., the lower bound in (16)).

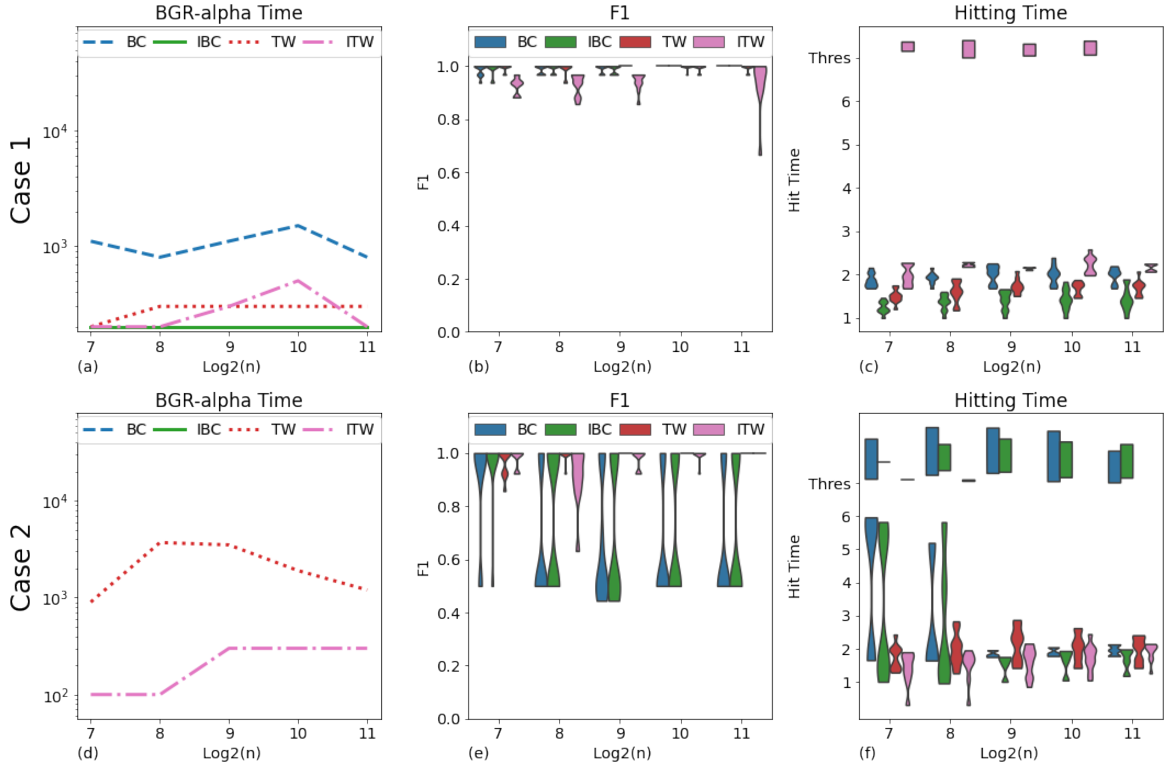


Figure 6: The improvement of the informed variants. (a) and (d) Time for achieving the worst case local BGR below 1.1. (b) and (e) Informed Twiggy Bayesian CART tends to overfit than its non-informed version. (c) and (f) Hitting times. (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: original and informed Twiggy Bayesian CART.

When the signal depth deepens, in the same settings of Example 1 and Example 2, the hitting time results are in Figure 7. We can see that informed versions hit generally faster than their non-informed versions. However, as in Figure 7 (b), the informed Bayesian CART falls short of resolving the hitting-time slow down, which is exponential in the signal depth. This result is consistent with Remark 8, implying that the problem of the myopic movement cannot be overcome even when using informed proposals. For an example where signals are disconnected but not very far from each other, see Figure 10 (b). On the other hand, the informed Twiggy Bayesian CART includes the signal in a single step.

6.2 Call Center Data

The data set, collected by a call center of an Israeli bank, contains arrival times, waiting times and service times. We focus on the arrival times, which can be seen as an inhomogeneous Poisson process with a mean function $\mu(t)$ [5]. This dataset was also studied in [6, 45] in the context of constructing adaptive confidence bands in non-parametric regression. In our analysis, we want to compare the mixing performance of each method studied in our simulations. We preprocess the data following [45] where the response is $Y_i = \sqrt{N_i + 1/4}$ where N_i

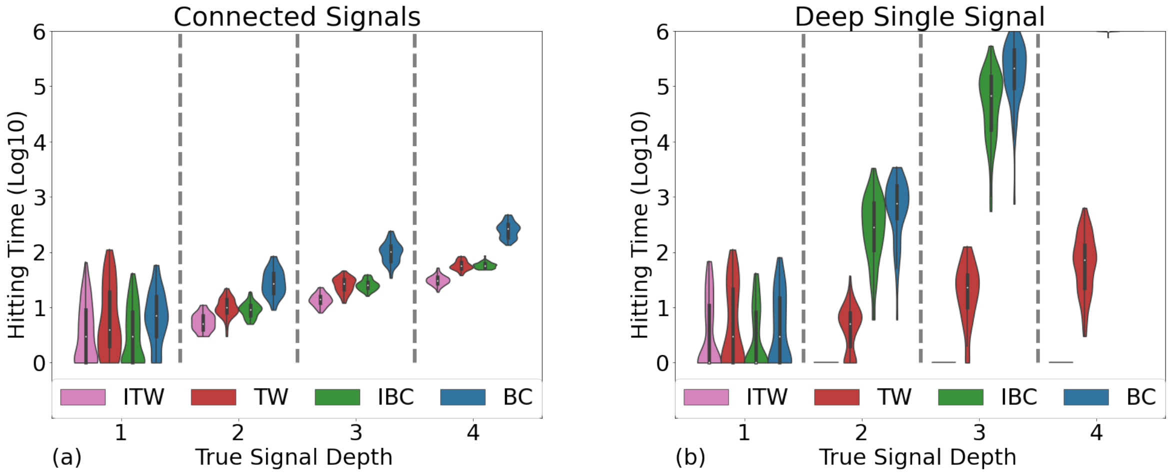


Figure 7: Hitting time when the true tree gets deeper. Informed Bayesian CART hits the true signals faster than Bayesian CART. Likewise, the informed Twiggy Bayesian CART is faster than Twiggy Bayesian CART. (b) However, for an isolated deep signal, informed Bayesian CART does not hit faster than Twiggy Bayesian CART. (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: original and informed Twiggy Bayesian CART.

is the number of calls arriving in the i-th time interval. We have $n = 2048$ equispaced time intervals. As a proxy of mixing, we measure MSE (distance from data to the posterior function draw) and the maximum local BGRs $\max_j \mathrm{BGR}(\beta_j | \{\mathcal{T}_t^{\mathcal{C}_k}, \mathcal{T}_{t-1}^{\mathcal{C}_k}, \dots, \mathcal{T}_{t-99}^{\mathcal{C}_k}\}_{k=1}^{10})$ at time t. As the split probability of the tree based models, we use $p_{lk} = 0.01$. For Spike-and-Slab, we again use two types of priors $p_{lk}^{ss,1} = 0.01$ and $p_{lk}^{ss,2} = 0.01 \times 6^{-l/2}$. As discussed in [6, 45], the data approximately follows the model in (1) with the variance $\sigma^2 = 1/4$. Therefore, we fix the variance at 0.25 in all methods.

In Figure 8, Spike-and-Slab shows a trade-off between overfitting and mixing. When the split probability is low ($p_{lk}^{ss,2}$), the chain may avoid overfitting (Figure 8 (f)) compared with a high split probability $p_{lk}^{ss,1}$ (Figure 8 (e)). However, Figure 9 (a) shows that the speed of the chain's ability to explain the data is much slower. The informed versions of the tree models catch more detailed signals than their non-informed counterparts. The speed of decreasing MSE is informed Twiggy Bayesian CART < Twiggy Bayesian CART < informed Bayesian CART < Bayesian CART < Spike-and-Slab (Figure 9).

7 Concluding Remarks

This work is the first to have described upper bounds on mixing times for Bayesian CART, a simplified version of BART. We focused on one-dimensional setting and various proposal schemes, including our new Twiggy Bayesian CART proposal. We showed rapid mixing of Bayesian CART when the signal is connected on a tree. We also obtained rapid mixing for

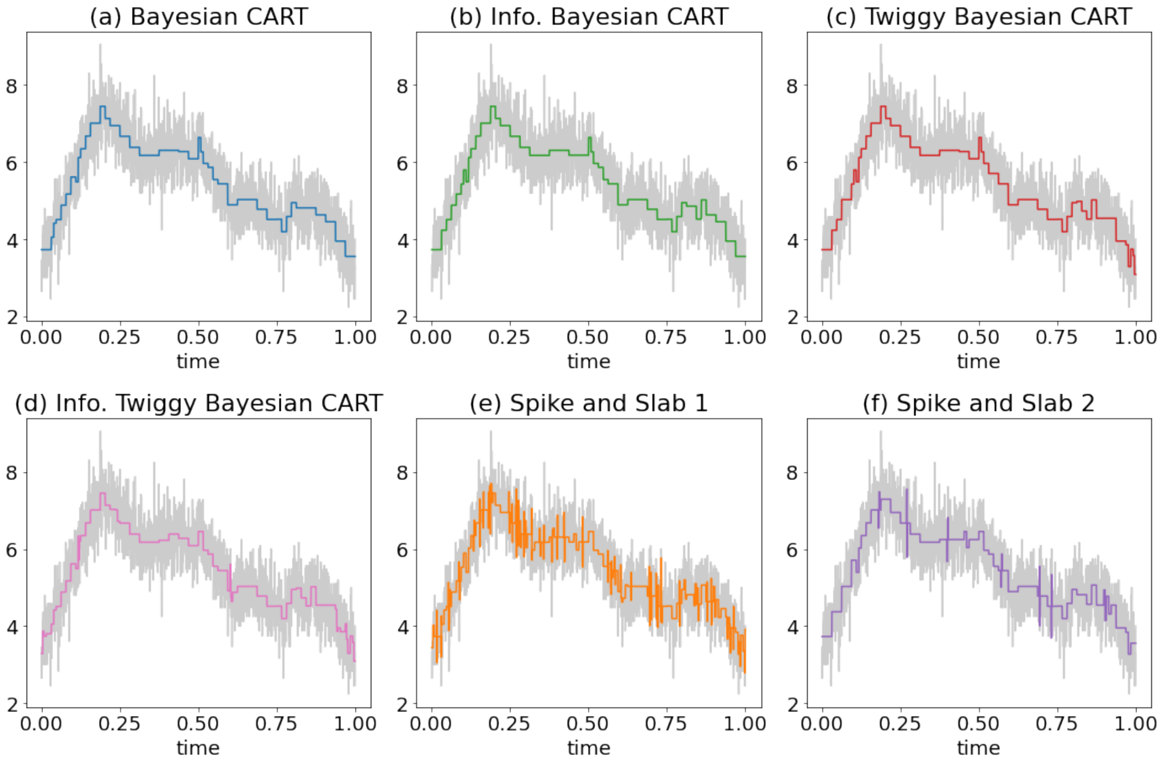


Figure 8: The visualization of MCMC chains on Call Center Data. The colored lines are the median tree fit obtained from 1000 samples after 10,000 burn-in and the gray lines are the data. (a) Bayesian CART (b) informed Bayesian CART (c) Twiggy Bayesian CART (d) informed Twiggy Bayesian CART (e) Spike-and-Slab (prior: $p_{lk}^{ss,1} = 0.01$) (f) Spike-and-Slab (prior: $p_{lk}^{ss,2} = 0.01 \times 6^{-l/2}$)

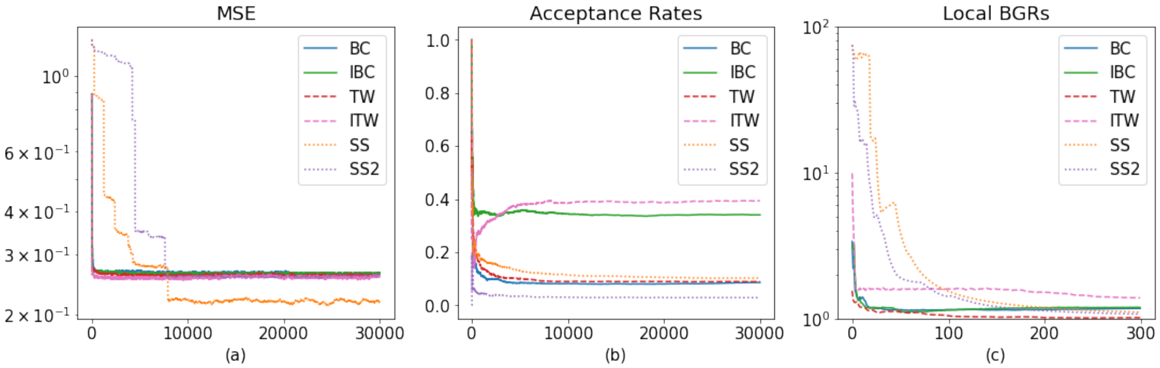


Figure 9: The performance measures on the Call Center data. (a) The MSE (log transformed) over the MCMC iterations (b) Acceptance rates. (c) The local BGRs for every 100 iterations on the Call center data. The minimum local BGRs achieved were Bayesian CART: 3.58, informed Bayesian CART: 1.21, Twiggy Bayesian CART: 1.01, informed Twiggy Bayesian CART: 9.35, Spike-and-Slab ($p_{lk}^{ss,1}$): 1.16, Spike-and-Slab ($p_{lk}^{ss,2}$): 1.17.

Twiggy Bayesian CART which does not require this assumption. We showed that without signal connectivity, Bayesian CART mixes poorly. Extending our conclusions to more dimensions is an interesting problem. The first challenge is the absence of identifiability of the multi-dimensional trees. The non-identifiabiltiy prevents from guaranteeing posterior consistency (e.g., Lemma 1), which is crucial in the canonical path argument. We believe that our results nevertheless serve as a valuable first step towards characterizing mixing of Bayesian Additive Regression Trees which have proven so useful in practice.

Contents

8 Bayesian CART Algorithm

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

9 Proof of Theorem 1 (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

$$\mathcal{A}_n \equiv \{\boldsymbol{\varepsilon} : \|\boldsymbol{X}'\boldsymbol{\varepsilon}\|_\infty \leq 2\|\boldsymbol{X}\|\sqrt{\log p}\}, \tag{30}$$

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.

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

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 $c > 5/2$.

9.0.1 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

$$\frac{\Pi[\Lambda(\mathcal{T}^*, K) \mid Y]}{\Pi(\mathcal{T}^* \mid Y)} = \sum_{\mathcal{T} \in \Lambda(\mathcal{T}^*, K)} \frac{\Pi(\mathcal{T}) N_{\mathcal{T}}(Y)}{\Pi(\mathcal{T}^*) N_{\mathcal{T}^*}(Y)} \tag{31}$$

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

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

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

$$\begin{aligned}
\frac{\Pi(\mathcal{T})N_Y(\mathcal{T})}{\Pi(\mathcal{T}^*)N_Y(\mathcal{T}^*)} &= (1+n)^{-K/2}\prod_{j=1}^K \frac{p_{l_j}}{1-p_{l_j}} \times (1-p_{l_j+1})^2 \times \exp\left\{-\frac{Y'(P_{j-1}-P_j)Y}{2(n+1)}\right\} \\
&= (1+n)^{-K/2}\prod_{j=1}^K \frac{p_{l_j}}{1-p_{l_j}} \times (1-p_{l_j+1})^2 \times \exp\left\{\frac{|X_{[j]}'Y|^2}{2(n+1)}\right\}, \qquad (32)
\end{aligned}$$

where

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

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

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

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

$$\frac{\Pi(\mathcal{T})N_Y(\mathcal{T})}{\Pi(\mathcal{T}^*)N_Y(\mathcal{T}^*)} \leq \exp\left(-\frac{K}{2}\log(1+1/n) - K\,(c-3/2)\log n\right) \leq \mathrm{e}^{-K(c-3/2)\log n}. \qquad (33)$$

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

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

we find an upper bound for (31)

$$\frac{\Pi[\Lambda(\mathcal{T}^*, K)\,|\,Y]}{\Pi(\mathcal{T}^*\,|\,Y)} \leq (|\mathcal{T}_{ext}^*| + K - 1)^K \mathrm{e}^{-K(c-3/2)\log n}. \qquad (34)$$

Since

$$\begin{aligned}
\frac{\Pi(\mathbb{T}_O\,|\,Y)}{\Pi(\mathcal{T}^*\,|\,Y)} &\leq \sum_{K=1}^{2^L} \frac{\Pi(\Lambda(\mathcal{T}^*, K)\,|\,Y)}{\Pi(\mathcal{T}^*\,|\,Y)} < \sum_{K=1}^{2^L} \mathrm{e}^{K\log[(|\mathcal{T}_{ext}^*|+K-1)]}\mathrm{e}^{-K\,(c-3/2)\,\log n} \\
&< \sum_{k=1}^{2^L} \mathrm{e}^{-K(c-5/2)\log n} \leq n^{5/2-c}\frac{1-[n^{(5/2-c)}]^{n/2}}{1-n^{5/2-c}} < \frac{1}{n^{c-5/2}-1} \qquad (35)
\end{aligned}$$

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

9.0.2 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

$$\Pi\left[\mathcal{T} \in \mathbb{T} : \mathcal{B}(A) \not\subseteq \mathcal{T}_{int} \,|\, Y\right] \to 0 \quad \text{as } n \to \infty \tag{36}$$

for

$$\mathcal{B}(A) \equiv \{(l,k) : C_{f_0} > |\beta^*_{lk}| > A \log n/\sqrt{n}\}. \tag{37}$$

The proof of (36) follows the route of Lemma 3 in [8] and Section 9.0.2 in [45]. For simplicity, we have focused in this work on the regular design case where the regression matrix is orthogonal and thereby $\Sigma_{\mathcal{T}} = c_n(X'_{\mathcal{T}}X_{\mathcal{T}})^{-1} = \frac{1}{n+1}I_{|\mathcal{T}_{ext}|}$. 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 [8]. We define $K = |\mathcal{T}^+_{int} \backslash \mathcal{T}_{int}|$ and write (using the expression in (8))

$$\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)} = (1+n)^{K/2} \exp\left\{\frac{1}{2(n+1)} Y'[X_{\mathcal{T}}X'_{\mathcal{T}} - X_{\mathcal{T}^+}X'_{\mathcal{T}^+}]Y\right\}. \tag{38}$$

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

$$\begin{aligned}\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)} &= (1+n)^{K/2} \prod_{j=1}^{K} \exp\left\{\frac{Y'(P_{j-1} - P_j)Y}{2(n+1)}\right\} \\ &= (1+n)^{K/2} \prod_{j=1}^{K} \exp\left\{-\frac{|X'_{[j]}Y|^2}{2(n+1)}\right\},\end{aligned} \tag{39}$$

where $P_j = X_{\mathcal{T}^j}X'_{\mathcal{T}^j} = P_{j-1} + X_{[j]}X'_{[j]}$ and where $X_{[j]}$ is the column added at the j^{th} step of branch growing. Let $X_{[K]}$ be the <u>last</u> column to be added to $X_{\mathcal{T}^+}$, i.e. the <u>signal</u> 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 $F_0 - X\boldsymbol{\beta}^* = 0$ 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

$$\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)} \leq \exp\left\{\frac{K}{2}\log(1+n) - \frac{nA^2\log^2 n}{4(n+1)} + \frac{4n\log n}{2(n+1)}\right\} \tag{40}$$

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

$$\frac{\Pi(\mathcal{T})}{\Pi(\mathcal{T}^+)} = \frac{1-p_{l_0}}{p_{l_0}} \times \left(\prod_{j=1}^{K-1}\frac{1}{p_{l_j}(1-p_{l_j})}\right) \times \frac{1}{(1-p_{l_K})^2}. \tag{41}$$

Defining

$$b(n) := \frac{\Pi(\mathcal{T}\,|\,Y)}{\Pi(\mathcal{T}^+\,|\,Y)}$$

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 $K \leq L \leq L_{max} = \log_2[n/2]$)

$$b(n) \leq 2^K \exp\left\{c\,K\log n + \frac{K}{2}\log(1+n) - \frac{nA^2\log^2 n}{4(n+1)} + \frac{4n\log n}{2(n+1)}\right\}. \tag{42}$$

Following the proof technique in Lemma 2 in [8] we conclude that for some sufficiently large $A > 0$

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

Thereby,

$$\Pi[\mathcal{B}(A) \not\subseteq \mathcal{T}_{int}\,|\,Y] \leq \sum_{(l_S,k_S)\in\mathcal{B}(A)} \Pi[(l_S,k_S) \notin \mathcal{T}_{int}\,|\,Y] \leq \mathrm{e}^{-A^2/4\log^2 n}2^L \leq \mathrm{e}^{-A^2/8\log^2 n} \to 0.$$

This concludes the proof of (36). Because $\mathcal{T}^*$ is the minimal tree that contains $\mathcal{B}(A)$ as its internal nodes, this implies $\Pi[\mathcal{T} \in \mathbb{T} : \mathcal{T}^* \not\subseteq \mathcal{T}\,|\,Y] = \Pi[\mathbb{T}_U\,|\,Y] = o(1)$.

10 Proof of Theorem 2 (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 [49]

$$Gap(P) \leq 2\Phi, \quad \text{where} \quad \Phi = \min_{\substack{A\subset\mathbb{T} \\ 0<\Pi[A\,|\,Y]\leq 1/2}} \frac{\sum_{\mathcal{T}\in A,\mathcal{T}'\in\mathbb{T}\setminus A}\Pi(\mathcal{T}\,|\,Y)P(\mathcal{T},\mathcal{T}')}{\Pi[A\,|\,Y]} \tag{43}$$

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 (43), 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.

$$A_{\backslash(l^*-1,0)} \equiv \{\mathcal{T} \in \mathbb{T} \mid (l^* - 1, 0) \notin \mathcal{T}_{int}\} \tag{44}$$

to bound the conductance. Note that $\Pi[A_{\backslash(l^*-1,0)} \mid Y] < 1/2$ since the posterior is concentrated to the true tree with $(l^*, 0)$ (See, Lemma 1)[7]. A tree $\mathcal{T} \in A_{\backslash(l^*-1,0)}$ must contain $(l^* - 2, 0)$ to have a non-zero transition probability $P(\mathcal{T}, \mathcal{T}')$ for $\mathcal{T}' \in A^c_{\backslash(l^*-1,0)}$. Therefore, denoting

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

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

$$\Phi \leq \frac{\sum_{\mathcal{T} \in B_{l^*-1}} \Pi(\mathcal{T} \mid Y) P(\mathcal{T}, A^c_{\backslash(l^*-1,0)})}{\Pi[A_{\backslash(l^*-1,0)} \mid Y]} \leq \frac{\Pi[B_{l^*-1} \mid Y]}{\Pi[A_{\backslash(l^*-1,0)} \mid Y]}. \tag{45}$$

We now show that the tightest upper bound in (45) is obtained when $l^* = L - 1$. Namely, we first derive a bound w.r.t. a general l^*, and show that the bound in (45) becomes smaller as l^* increases. We will work conditionally on the set $\mathcal{A}_n$ defined in (30). 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 (45), we decompose $A_{\backslash(l^*-1,0)}$ 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 $A_{\backslash(l^*-1,0)} = \bigcup_{i=0}^{l^*-1} B_i$, so that

$$\Pi[A_{\backslash(l^*-1,0)} \mid Y] = \sum_{i=0}^{l^*-1} \Pi[B_i \mid Y]. \tag{46}$$

Therefore, the bound in (45) can be rewritten as

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

[7] By consistency established in Lemma 1, 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 $\mathcal{T}^* \notin A_{\backslash(l^*-1,0)}$, we have $\Pi[A_{\backslash(l^*-1,0)} \mid Y] \leq 1/2$.

To see how small $\Pi[B_{l^*-1}\,|\,Y]$ is compared to $\sum_{i=0}^{l^*-2}\Pi[B_i\,|\,Y]$, we will scrutinize the posterior ratio $\Pi[B_{i-1}\,|\,Y]/\Pi[B_i\,|\,Y]$ 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 $\mathcal{T}'=M(\mathcal{T})$

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

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

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

Each tree $\mathcal{T}\in\mathcal{N}(\mathcal{T}')$ differs from $\mathcal{T}'$ by addition of at least one node without signal. We decompose $\mathcal{N}(\mathcal{T}')=\cup_{K=1}^{2^{L_{max}}}\mathcal{N}(\mathcal{T}',K)$ 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 (33) for nested models to conclude that for any $\mathcal{T}\in\mathcal{N}(\mathcal{T}',K)$ we have

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

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 [8], satisfies $\mathbb{C}_K\asymp 4^K/K^{3/2}$. Then

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

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

$$\begin{aligned}\Pi[A_{\backslash(l^*-1,0)}\,|\,Y]&=\sum_{i=0}^{l^*-1}\Pi[B_i\,|\,Y]\geq\Pi[B_{l^*-1}\,|\,Y]\sum_{i=0}^{l^*-1}\left(\frac{1}{\gamma_n}\right)^i\\&=\Pi[B_{l^*-1}\,|\,Y]\frac{\gamma_n}{1-\gamma_n}\left[\left(\frac{1}{\gamma_n}\right)^{l^*}-1\right].\end{aligned}\tag{47}$$

Therefore, it follows from (47) and (45) that

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

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

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

11 Proof of Theorem 3 (Bayesian CART Mixing Upper Bound)

11.1 Proof of Lemma 1

We want to upper bound the length of the longest canonical path constructed in Section 5.2.1. 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 $\max\limits_{\mathcal{T}:\mathcal{T}\supset\mathcal{T}^*}\{|T_{\mathcal{T},\mathcal{T}^*}|\} \leq (2^L - |\mathcal{T}^*_{int}|)$. 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 $\max\limits_{\mathcal{T}:\mathcal{T}\subset\mathcal{T}^*} |T_{\mathcal{T},\mathcal{T}^*}| \leq |\mathcal{T}^*_{int}|$. 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 $\max_{\mathcal{T},\mathcal{T}'\in\mathbb{T}} |T_{\mathcal{T},\mathcal{T}'}| \leq 2^{L+1}$.

11.2 Proof of Lemma 2

We will work conditionally on the set $\mathcal{A}_n$ defined in (30), where $p = 2^{L_{max}} = n/2$. We know that the complement of this set has a vanishing probability $\mathbb{P}(\mathcal{A}_n^c) \leq 2/p \to 0$. 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

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

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

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

First, we denote with

$$\Delta(\mathcal{T}') = \{\mathcal{T} : \mathcal{T}' \in T_{\mathcal{T},\mathcal{T}^*}\} \tag{48}$$

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

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

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

$$\rho(\mathcal{E}) \leq \max_{e_{\mathcal{T},\mathcal{T}'}\in\mathcal{E}} \frac{\Pi[\Delta(\mathcal{T})]}{Q(e_{\mathcal{T},\mathcal{T}'})} \leq \max_{(\mathcal{T},\mathcal{T}')\in\Gamma^*} \frac{\Pi[\Delta(\mathcal{T})]}{Q(\mathcal{T},\mathcal{T}')}, \tag{49}$$

where

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

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 5.2.1. 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

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

Plugging this into (49) we obtain

$$\rho(\mathcal{E})\leq 2\max_{(\mathcal{T},\mathcal{T}')\in\Gamma^*}\left\{\frac{\Pi[\Delta(\mathcal{T})]}{\Pi(\mathcal{T}\,|\,Y)S(\mathcal{T}\to\mathcal{T}')}\times\max\left[1,\frac{\Pi(\mathcal{T}\,|\,Y)S(\mathcal{T}\to\mathcal{T}')}{\Pi(\mathcal{T}'\,|\,Y)S(\mathcal{T}'\to\mathcal{T})}\right]\right\}. \tag{50}$$

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 $\mathbb{T}_U=\{\mathcal{T}:\mathcal{T}\not\supset\mathcal{T}^*\}$.

11.2.1 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 (32) and (33) for overfitted trees with $K=1$ we obtain (for $j=2^{l_S}+k_S+1$)

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

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 (50) satisfies

$$\frac{\Pi(\mathcal{T}\,|\,Y)S(\mathcal{T}\to\mathcal{T}^-)}{\Pi(\mathcal{T}^-\,|\,Y)S(\mathcal{T}^-\to\mathcal{T})}\leq 4\mathrm{e}^{-(c-3/2)\log n}=o(1)\quad\text{for } c>3/2. \tag{51}$$

We now focus on the second ratio in the product in (50). When $\mathcal{T}^*\subset\mathcal{T}$, all precedents $\mathcal{T}'\in\Delta(\mathcal{T})$ (recall the definition of $\Delta(\mathcal{T})$ in (48)) 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

$$\Delta(\mathcal{T},K)=\{\mathcal{T}'\in\Delta(\mathcal{T}):|T_{\mathcal{T}',\mathcal{T}}|=K\} \tag{52}$$

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 (32) and (33), 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 $\text{card}[\Delta(\mathcal{T}, K))] \leq \prod_{j=1}^{K}(|\mathcal{T}_{ext}|+j-1)$ and $\text{card}[\Delta(\mathcal{T}, 0))] = 1$. Thereby

$$\begin{aligned}\frac{\Pi[\Delta(\mathcal{T}) \mid Y]}{\Pi(\mathcal{T} \mid Y)} &= 1 + \sum_{K=1}^{2^L} \sum_{\mathcal{T}' \in \Delta(\mathcal{T},K)} \frac{\Pi(\mathcal{T}' \mid Y)}{\Pi(\mathcal{T} \mid Y)} \leq 1 + \sum_{K=1}^{2^L} \mathrm{e}^{K \log[(|\mathcal{T}_{ext}|+K-1)]} \mathrm{e}^{-(c-3/2)\, K \log n} \\ &< 1 + \frac{1}{n^{(c-5/2)} - 1}. \end{aligned} \tag{53}$$

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

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

11.2.2 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 (39) and (42) 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 (48) 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 <u>not</u> 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 $\text{card}[\Delta(\mathcal{T}, K)] \leq 2^{LK}$.

Because under the Assumption 1 (a), all internal nodes in $\mathcal{T}^*$ are signals, we can modify the expressions in (39) to include all K signals (not just one) to obtain for large enough A

$$\frac{\Pi[\mathcal{U}_1(\mathcal{T})\,|\,Y]}{\Pi(\mathcal{T}\,|\,Y)} \leq \sum_{K=1}^{|\mathcal{T}^*_{int}|} \sum_{\mathcal{T}'\in\Delta(\mathcal{T},K)\cap\mathcal{U}_1(\mathcal{T})} \frac{\Pi(\mathcal{T}'\,|\,Y)}{\Pi(\mathcal{T}\,|\,Y)} < \sum_{K=1}^{|\mathcal{T}^*_{int}|} \mathrm{e}^{-KA^2/8\log^2 n} < \frac{1}{n^{A^2/8\log n}-1}. \quad (54)$$

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 (34) and (35) we find that

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

and thereby using (54)

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

Putting it all together, we have

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

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 (50) yields $\rho\leq 2^{L+1}(1+o(1))$ for $c>5/2$.

12 Proof of Theorem 3

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

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

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 9.0.1 that (using the shorthand notation $p_l=p_{lk}$)

$$\frac{\Pi(\mathcal{T}\,|\,Y)}{\Pi(\mathcal{T}^*\,|\,Y)}=(1+n)^{-K/2}\prod_{j=1}^{K}\left(\frac{p_{l_j}}{1-p_{l_j}}\times(1-p_{l_j+1})^2\right)\times\exp\left\{\frac{|X'_{[j]}Y|^2}{2(n+1)}\right\}. \tag{55}$$

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

$$\min_{\mathcal{T}\in\mathbb{T}_O}\Pi(\mathcal{T}\,|\,Y)>\frac{1}{2}\min_{\mathcal{T}\in\mathbb{T}_O}\frac{\Pi(\mathcal{T}\,|\,Y)}{\Pi(\mathcal{T}^*\,|\,Y)}>\frac{1}{2}\left(\frac{1}{2n^c\sqrt{1+n}}\right)^K>\frac{\mathrm{e}^{-\frac{n}{2}[\log 2+(c+1/2)\log(1+n)]}}{2}. \tag{56}$$

Similarly as in Section 11.2.2, 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 $\mathcal{T}^0=\mathcal{T}\to\mathcal{T}^1\to\dots\to\mathcal{T}^K=\mathcal{T}^+$ the sequence of nested trees obtained by adding one additional internal node (l_j,k_j) towards (l_S,k_S). As in Section 9.0.2, we find that

$$\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)}=(1+n)^{K/2}\prod_{j=1}^{K}\exp\left\{-\frac{|X'_{[j]}Y|^2}{2(n+1)}\right\}, \tag{57}$$

We have

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

and thereby

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

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 $\mathcal{T}^{+\dots+}=\mathcal{T}^*$ We then bound

$$\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^*)}=\frac{N_Y(\mathcal{T})}{N_Y(\mathcal{T}^+)}\times\frac{N_Y(\mathcal{T}^+)}{N_Y(\mathcal{T}^{++})}\times\dots\times\frac{N_Y(\mathcal{T}^{+\dots+})}{N_Y(\mathcal{T}^*)} \tag{58}$$

$$>(1+n)^{1/2}\exp\left\{-\frac{n^2\,\|\boldsymbol{\beta}^*_{\mathcal{T}^*}\|^2}{n+1}-\frac{4n\,|\mathcal{T}^*_{int}|\log n}{n+1}\right\}. \tag{59}$$

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

$$\frac{\Pi(\mathcal{T})}{\Pi(\mathcal{T}^+)} = \frac{1-p_{l_1}}{p_{l_1}} \times \left(\prod_{j=2}^{K-1} \frac{1}{p_{l_j}(1-p_{l_j})} \right) \times \frac{1}{(1-p_{l_K})^2} > n^{c(K-1)} \tag{60}$$

This yields

$$\min_{\mathcal{T}\in\mathbb{T}_U:\mathcal{T}\subset\mathcal{T}^*} \Pi(\mathcal{T}\,|\,Y) > (1+n)^{1/2} \exp\left\{ -\frac{n^2\,\|\boldsymbol{\beta}^*_{\mathcal{T}^*}\|^2}{n+1} - \frac{4n\,|\mathcal{T}^*_{int}|\log n}{n+1} \right\}. \tag{61}$$

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

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

and combining the expression (56) with (61) we find

$$\begin{aligned}\min_{\mathcal{T}\in\mathbb{T}_U:\mathcal{T}\not\subset\mathcal{T}^*}\frac{\Pi(\mathcal{T}\,|\,Y)}{\Pi(\mathcal{T}^*\,|\,Y)} &> \min_{\mathcal{T}\in\mathbb{T}_U:\mathcal{T}\not\subset\mathcal{T}^*}\frac{\Pi(\mathcal{T}\,|\,Y)}{\Pi(\mathcal{T}_U\,|\,Y)} \times \min_{\mathcal{T}\in\mathbb{T}_U:\mathcal{T}\subset\mathcal{T}^*}\frac{\Pi(\mathcal{T}\,|\,Y)}{\Pi(\mathcal{T}^*\,|\,Y)}\\ &> \frac{1}{2}\exp\left\{ -n\left[1+\left(c+\frac{1}{2}\right)\log(1+n)\right] - \frac{n^2\,|\mathcal{T}^*_{int}|C^2_{f_0}}{n+1} - \frac{4n\,|\mathcal{T}^*_{int}|\log n}{n+1}\right\}\end{aligned}$$

Now, by Lemma 1 we have $l(\mathcal{E}) \leq 2^{L+1}$ and by Lemma 2 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

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

13 Proof of Theorem 4 (Twiggy Bayesian CART Mixing Upper Bound)

We follow the same recipe as in the proof of Theorem 3. We first need to show Lemma 1 and Lemma 2 for the canonical path ensemble for Twiggy Bayesian CART constructed in Section 13.0.1 below.

13.0.1 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

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

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

$$\mathcal{T}^-_{int} = \mathcal{T}_{int} \setminus [(\widetilde{l},\widetilde{k}) \leftrightarrow (l^*,k^*)] \quad \text{where} \quad (\widetilde{l},\widetilde{k}) = \arg \min_{(l,k)\in S(\mathcal{T})} (2^l + k). \tag{62}$$

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

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

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 1 and Lemma 2 for the unstructured signal Assumption 1 (b) (see proof of Theorem 4 in Section 13).

13.1 Version of Lemma 1 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 $|\mathcal{P}(\mathcal{T})\backslash\mathcal{T}^*_{int}|$ 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 $\max\limits_{\mathcal{T}:\mathcal{T}\subset\mathcal{T}^*}|T_{\mathcal{T},\mathcal{T}^*}|\leq|\mathcal{P}(\mathcal{T}^*)|$. 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.

13.2 Version of Lemma 2 for Twiggy Bayesian CART

We will again work on the event space $\mathcal{A}_n$ in (30), which has probability at least $1-4/n$. The strategy is the same as in the proof of Lemma 2. We again split the considerations into overfitted and underfitted trees.

13.2.1 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

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

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 (50). 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 (32) and (33) for overfitted trees we obtain

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

This means that the second maximum quantity in (50) 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 (50). When $\mathcal{T}^* \subset \mathcal{T}$, all precedents $\mathcal{T}' \in \Delta(\mathcal{T})$ (recall the definition of $\Delta(\mathcal{T})$ in (48)) 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 2 in Section 11.2, we decompose $\Delta(\mathcal{T}) = \cup_{K=1}^{2^L}\Delta(\mathcal{T},K)$ into shells $\Delta(\mathcal{T},K) = \{\mathcal{T}' \in \Delta(\mathcal{T}) : |T_{\mathcal{T}',\mathcal{T}}| = K\}$ defined as in (52). The difference now is that each tree $\mathcal{T}' \in \Delta(\mathcal{T},K)$ can have more than K redundant nodes. We denote with $\boldsymbol{\kappa} = (k(1),\dots,k(K))' \in (\mathbb{N}\backslash\{0\})^K$ 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 (32) and (33), we now obtain for $\mathcal{T}' \in \Delta(\mathcal{T},K)$

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

Moreover, we define

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

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 [8], 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

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

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

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

and thereby

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

13.2.2 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 (39) and (42) we again find that

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

We now proceed similarly as in Section 11.2.2. The precedents $\Delta(\mathcal{T})$ in (48) of an underfitted model $\mathcal{T}$ are again divided into mutually exclusive categories defined in Section 11.2.2, 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 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 [8], we have $\mathbb{C}_K \asymp 4^K/K^{3/2}$ and using the expressions in (39) we obtain for large enough $A > 0$ and $|\mathcal{T}^*_{int}| \lesssim \log^2 n$

$$\frac{\Pi[\mathcal{U}_1(\mathcal{T})\,|\,Y]}{\Pi(\mathcal{T}\,|\,Y)} \leq \sum_{K=1}^{|\mathcal{P}(\mathcal{T}^*)|} \sum_{\mathcal{T}' \in \Delta(\mathcal{T},K)\cap\mathcal{U}_1(\mathcal{T})} \frac{\Pi(\mathcal{T}'\,|\,Y)}{\Pi(\mathcal{T}\,|\,Y)} \lesssim |\mathcal{T}^*_{int}| \times 4^{|\mathcal{T}^*_{int}|} \times \mathrm{e}^{-A^2/8\log^2 n} = o(1). \tag{63}$$

For the second type of underfitting precedents, we follow the same arguments as in Section 11.2.2 to conclude

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

and thereby

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

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 (50) 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 (25) for Twiggy Bayesian CART is then obtained by the sandwich relation (19) with (21) in the same way as in the proof of Theorem 3.

14 Proof of Theorem 5 (Mixing Upper Bound for Locally Informed Versions)

14.1 General two-stage drift condition

We will use a similar strategy as in [55]. We first state the general two-stage drift condition theorem and its corollary, which are a slight modification from [55].

Theorem 6. *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) $PV_1 \leq \lambda_1 V_1$ on A^c,

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

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

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

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

(v) For any $x \in A$, $P(x, A^c) \leq q$ 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

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

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 1. *Recall the definition of the ϵ-mixing time in* (18). *In the setting of Theorem 6, assume that $\lambda_1, \lambda_2 \to 1$ and $q \leq \min\{1-\lambda_1, (1-\lambda_2)/C_2K_2\}$ 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*

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

14.1.1 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 [55] can be applied in the context of regression trees. First, we consider the case of Bayesian CART. We will check the conditions in Theorem 6. 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 1 by a factor of 1/2[8]. Therefore, the bounds in Proposition 1 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 $\mathcal{T} \neq \mathcal{T}^*$,

$$\frac{(P_{lazy}V_2)(\mathcal{T})}{V_2(\mathcal{T})} \leq 1 - \frac{1}{2^{L+3}}\frac{1}{(1+n^{5/2-c})} + \frac{M}{2n^{c-3/2}} + \frac{1}{2}n^{1-(A^2\log n)/8}, \tag{64}$$

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

$$\begin{aligned}\lambda_1 &= 1 - \frac{A^2}{72(C_{f_0}+2)^2}\frac{\log^2 n}{2^L\, n},\\ \lambda_2 &= 1 - \frac{1}{2^{L+7/2}}.\end{aligned} \tag{65}$$

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 6 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 3 (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 10). To check condition (v), we set $C_2 = 2\mathrm{e}$ as a universal constant in Corollary 1. 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 (67) and by the definition of P_{lazy}, we

[8] 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 (69), we have

$$\begin{aligned}\frac{B(\mathcal{T},A^c)}{2} &= \sum_{\widetilde{\mathcal{T}}\in A^c\cap N_p(\mathcal{T})} \frac{B(\mathcal{T},\widetilde{\mathcal{T}})}{2} \le \sum_{\widetilde{\mathcal{T}}\in A^c\cap N_p(\mathcal{T})} \frac{1}{2n^{(A^2\log n)/8}} \\ &\le \frac{1}{2n^{(A^2\log n)/8-1}} \le q = \min(\frac{A^2}{72(C_{f_0}+2)^2}\frac{\log^2 n}{2^L\, n}, \frac{1}{\mathrm{e}^2\, 2^{L+9/2}}), \end{aligned} \tag{66}$$

for large enough A and n. Therefore, condition (v) is satisfied. Now, by applying Corollary 1, we have

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

Lastly, when it comes to the Twiggy Bayesian CART, the only difference is that we have $M = 2L$ in (64) instead of $M = 1$. This change does not affect the above proof because λ_2 in (65) is still valid. Therefore, we finish our proof.

14.2 Proof of Proposition 1

The proof is based on the key decomposition characterized in [55] as (for $i = 1, 2$)

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

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

$$P(\mathcal{T},\widetilde{\mathcal{T}}) = \min\{S(\mathcal{T}\to\widetilde{\mathcal{T}}), B(\mathcal{T},\widetilde{\mathcal{T}})S(\widetilde{\mathcal{T}}\to\mathcal{T})\} \le B(\mathcal{T},\widetilde{\mathcal{T}}). \tag{67}$$

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 (30), we have two prior ratios. First, similar to (33) for any overfitted trees $\mathcal{T} \subset \widetilde{\mathcal{T}} \in \mathbb{T}_L$ such that $\mathcal{T} \supseteq \mathcal{T}^*$ and $|\widetilde{\mathcal{T}} \backslash \mathcal{T}| = K$,

$$\frac{\Pi(\widetilde{\mathcal{T}} \mid Y)}{\Pi(\mathcal{T} \mid Y)} \leq n^{-K(c-3/2)}. \tag{68}$$

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 $k \geq 1$ extra nodes for the Twiggy Bayesian CART. For any such $\widetilde{\mathcal{T}}$, from (42) we have

$$\frac{\Pi(\widetilde{\mathcal{T}} \mid Y)}{\Pi(\mathcal{T} \mid Y)} \geq n^{(A^2 \log n)/8}. \tag{69}$$

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

Lemma 3. *Under the same assumptions of Theorem 5, 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) $1 \leq V_1(\mathcal{T}) \leq \mathrm{e}$ *and* $1 \leq V_2(\mathcal{T}) \leq \mathrm{e}$.

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

$$R_1(\mathcal{T}, \widetilde{\mathcal{T}}) \leq 0, \quad R_1(\widetilde{\mathcal{T}}, \mathcal{T}) \geq 0.$$

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

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

Proof. For part (i), we first show the upper bound of $V_1(\mathcal{T})$. We will work on $\mathcal{A}'_n = \mathcal{A}_n \cap \{\varepsilon : \|\varepsilon\|_2^2 \leq 2n\}$, where $\mathcal{A}_n$ is defined in (30). Since $\|\varepsilon\|_2^2 \sim \chi^2(n)$, by applying the tail bound in [22] (Theorem 1), we have $\mathbb{P}(\|\varepsilon\|_2^2 > 2n) \leq \mathrm{e}^{-n/8}$. Therefore, $\mathbb{P}(\mathcal{A}'_n) \geq 1 - 4/n - \mathrm{e}^{-n/8}$. As $\boldsymbol{\nu} = \varepsilon$, we obtain the bound by observing that on the event $\mathcal{A}'_n$ with $p := 2^L$, the following holds.

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

where we use the assumption that $|\beta^*_{lk}| \leq C_{f_0}$. The other upper bound in part (i) is trivial since for any tree $\mathcal{T} \in \mathbb{T}_L$ we have $\mathcal{T}_{int} \leq 2^L$. 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,

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

For part (iii), we have $|\widetilde{\mathcal{T}}\backslash\mathcal{T}^*|-|\mathcal{T}\backslash\mathcal{T}^*|=k\leq 2^L$, and $V_2(\widetilde{\mathcal{T}})/V_2(\mathcal{T})=e^{k/2^L}$. The result follows by using the two inequalities as in [55]

$$\mathrm{e}^x\leq 1+2x,\quad \mathrm{e}^{-x}\leq 1-\frac{x}{2},\quad \forall x\in[0,1]. \tag{70}$$

□

14.2.1 Drift condition for overfitted models (R_2)

Lemma 4. *Recall the definition of w_p and w_g in (16). Under the same assumptions of Theorem 5, for any overfitted tree $\mathcal{T}\in\mathbb{T}_L$,*

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

(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) $Z_p(\mathcal{T})\leq|a_{\mathcal{T}}|+|b_{\mathcal{T}}|\,n^{c-3/2}$, 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 (68),

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

(ii) When $\widetilde{\mathcal{T}}\supset\mathcal{T}^*$, by applying (68), 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 (69), 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 5. *Under the same assumptions of Theorem 5, 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 $a_{\mathcal{T}}$ and $b_{\mathcal{T}}$ in Lemma 4 (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 (68), 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 4 (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 (67), 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 3 (iii), and then Lemma 4 (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 $c > 5/2$,

$$-\sum_{\widetilde{\mathcal{T}} \in b_{\mathcal{T}}} R_2(\mathcal{T}, \widetilde{\mathcal{T}})P(\mathcal{T}, \widetilde{\mathcal{T}}) \geq \frac{|b_{\mathcal{T}}| n^{c-3/2}}{2^{L+2}(n + |b_{\mathcal{T}}|\, n^{c-3/2})} \geq \frac{1}{2^{L+2}}\frac{1}{1 + n^{5/2-c}}. \tag{71}$$

Note that from (67) and (69), 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 (67) and (68),

$$P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq B(\mathcal{T}, \widetilde{\mathcal{T}}) \leq \frac{1}{n^{c-3/2}}. \tag{72}$$

With Lemma 3 (iii) with $k = |\widetilde{\mathcal{T}}_{int}| - |\mathcal{T}_{int}|$, we obtain that

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

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. □

14.2.2 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 14.2.

Lemma 6. *Under the same assumptions of Theorem 5, 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 (69), 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 7. *Suppose $B(\mathcal{T}, \widetilde{\mathcal{T}}) \geq n^a$ for some $a \in \mathbb{R}$ and define*

$$b = \frac{1}{2^{L-1}(C_{f_0}+2)^2 n}\left(a \log n - \log\left(\frac{\Pi(\widetilde{\mathcal{T}})}{\Pi(\mathcal{T})}\right) - \frac{|\mathcal{T}_{ext}| - |\widetilde{\mathcal{T}}_{ext}|}{2}\log(1+n)\right). \tag{73}$$

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 (70) to get $-R_1(\mathcal{T}, \widetilde{\mathcal{T}}) \geq b/2$. □

Lemma 8. *Under the same assumptions of Theorem 5, for any underfitted tree $\mathcal{T} \in \mathbb{T}_L$ i.e., for any $\mathcal{T} \not\supset \mathcal{T}^*$,*

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

The bounds are for both the Bayesian and Twiggy CART.

Proof. **The GROW movement.** By (69), 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 $B(\mathcal{T}, \mathcal{G}(\mathcal{T})) \geq n^{(A^2 \log n)/8}$. By Lemma 7 with $a = (A^2 \log n)/8$ and with large enough n so that b in (73) is less than 1[9], we find that

$$\begin{aligned}
-R_1(\mathcal{T},\mathcal{G}(\mathcal{T})) &\geq \frac{1}{2n\,2^{L-1}(C_{f_0}+2)^2}\left(a\log n - \log\left(\frac{\Pi(\mathcal{G}(\mathcal{T}))}{\Pi(\mathcal{T})}\right) - \frac{|\mathcal{T}_{ext}|-|\mathcal{G}(\mathcal{T})_{ext}|}{2}\log(1+n)\right) \\
&\geq \frac{1}{2^L\, n(C_{f_0}+2)^2}\left(a\log n - k\log\left(n^{-c}(1-n^{-c})\right) + \frac{k}{2}\log(1+n)\right) \\
&\geq \frac{1}{2^L\, n(C_{f_0}+2)^2}\, a\log n, \qquad (74)
\end{aligned}$$

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

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

Again using Lemma 7, we have for all $\widetilde{\mathcal{T}} \in \mathcal{D}(\mathcal{T})$,

$$\begin{aligned}
-R_1(\mathcal{T},\widetilde{\mathcal{T}}) &\geq \frac{1}{2n\,2^{L-1}(C_{f_0}+2)^2}\left((a-V)\log n - \log\left(\frac{\Pi(\widetilde{\mathcal{T}})}{\Pi(\mathcal{T})}\right) - \frac{k}{2}\log(1+n)\right) \\
&= \frac{1}{2^L\, n(C_{f_0}+2)^2}\left((a-V)\log n - k\log\left(n^{-c}(1-n^{-c})\right) + \frac{k}{2}\log(1+n)\right) \\
&\geq \frac{1}{2^L\, n(C_{f_0}+2)^2}\,(a-V)\log n, \qquad (75)
\end{aligned}$$

where $k = |\widetilde{\mathcal{T}}_{ext}| - |\mathcal{T}_{ext}| \geq 1$. 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 $\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})$,

$$S(\widetilde{\mathcal{T}} \to \mathcal{T}) \geq S_{PRUNE}(\widetilde{\mathcal{T}} \to \mathcal{T})/2 = \frac{w_p(\mathcal{T}|\widetilde{\mathcal{T}})}{2Z_p(\widetilde{\mathcal{T}})} \geq \frac{1}{2Z_p(\widetilde{\mathcal{T}})}. \qquad (76)$$

This lower bound of $S(\widetilde{\mathcal{T}} \to \mathcal{T})$ in (76) 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 $\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})$,

$$\begin{aligned}
P(\mathcal{T},\widetilde{\mathcal{T}}) &= \min\{S(\mathcal{T} \to \widetilde{\mathcal{T}}), B(\mathcal{T},\widetilde{\mathcal{T}})S(\widetilde{\mathcal{T}} \to \mathcal{T})\} \\
&\geq \min\{\frac{w_g(\widetilde{\mathcal{T}}|\mathcal{T})}{2Z_g(\mathcal{T})}, \frac{B(\mathcal{T},\widetilde{\mathcal{T}})}{2Z_p(\widetilde{\mathcal{T}})}\} \\
&\geq w_g(\widetilde{\mathcal{T}}|\mathcal{T})\min\{\frac{1}{2Z_g(\mathcal{T})}, \frac{1}{2Z_p(\widetilde{\mathcal{T}})}\} \\
&\geq \frac{w_g(\widetilde{\mathcal{T}}|\mathcal{T})}{2Z_g(\mathcal{T})}. \qquad (77)
\end{aligned}$$

In the last inequality, we used Lemma 6 (i) and (ii), for a large enough A,

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

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

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

$$\begin{aligned}
Z_g(\mathcal{T}) &= \sum_{\widetilde{\mathcal{T}}\in\mathcal{N}_g(\mathcal{T})} w_g(\widetilde{\mathcal{T}}|\mathcal{T}) \\
&= w_g(\mathcal{G}(\mathcal{T})|\mathcal{T}) + \sum_{\widetilde{\mathcal{T}}\in\mathcal{D}'} w_g(\widetilde{\mathcal{T}}|\mathcal{T}) + \sum_{\widetilde{\mathcal{T}}\in\mathcal{N}_g(\mathcal{T})\backslash\mathcal{D}} w_g(\widetilde{\mathcal{T}}|\mathcal{T}) \\
&= n^{(A^2\log n)/8} + W + n^{(A^2\log n)/8-V+1} \\
&\leq W + 2n^{(A^2\log n)/8}. \qquad (78)
\end{aligned}$$

Now, putting all things together using Lemma 3 (ii), (74), (75), (77), and (78), and recalling $a = (A^2\log n)/8$, we get

$$\begin{aligned}
-\sum_{\widetilde{\mathcal{T}}\in\mathcal{N}_g(\mathcal{T})} R_1(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}}) &\geq -\sum_{\widetilde{\mathcal{T}}\in\mathcal{D}(\mathcal{T})} R_1(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}}) \\
&\geq \frac{\log n}{2^L\,n(C_{f_0}+2)^2}\left(a\frac{n^{(A^2\log n)/8}}{2Z_g(\mathcal{T})} + (a-V)\sum_{\widetilde{\mathcal{T}}\in\mathcal{D}'(\mathcal{T})}\frac{w_g(\widetilde{\mathcal{T}}|\mathcal{T})}{2Z_g(\mathcal{T})}\right) \\
&\geq \frac{a\log n}{2^{L+2}\,n(C_{f_0}+2)^2}\frac{n^{(A^2\log n)/8} + (1-V/a)W}{n^{(A^2\log n)/8} + W/2}. \qquad (79)
\end{aligned}$$

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 6 (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 (67), 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 3 (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,

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

□

14.3 Proof of Remark 9

Here, we present the non-informed counterpart of Proposition 1. To achieve this, we modify V_1 as

$$V_1(\mathcal{T}) = \exp\left\{\frac{1}{2^L\, C_{f_0}^2 n}\left((\boldsymbol{X}\boldsymbol{\beta}^*)'(I - P_{\mathcal{T}}/n)\boldsymbol{X}\boldsymbol{\beta}^*\right)\right\}, \tag{80}$$

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 3 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 3 (i), we obtain the bound by

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

Proposition 2. *Under the same assumptions of Theorem 5, 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 $c > 3/2$,

$$\frac{(PV_2)(\mathcal{T})}{V_2(\mathcal{T})} \leq 1 - \frac{1}{2^{2L+2}} + \frac{M}{n^{c-3/2}} + n^{1-(A^2 \log n)/8}, \tag{81}$$

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 (81) 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 14.1.1 (the application of the two-drift condition) to this case, obtaining the bound in Remark 9. Proposition 2 is derived from the non-informed counterpart of Lemma 5 and Lemma 8 presented below.

Lemma 9. *Under the same assumptions of Theorem 5, 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 $\mathcal{T} \neq \mathcal{T}^*$,*

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_p(\mathcal{T})} R_2(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq -\frac{1}{2^{2L+2}} + n^{1-(A^2 \log n)/8}, \tag{82}$$

$$\sum_{\widetilde{\mathcal{T}} \in \mathcal{N}_g(\mathcal{T})} R_2(\mathcal{T}, \widetilde{\mathcal{T}}) P(\mathcal{T}, \widetilde{\mathcal{T}}) \leq \frac{M}{n^{c-3/2}}, \tag{83}$$

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 5, 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 (68), 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 3 (iii), and then Lemma 4 (iii), we find that

$$-R_2(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}})\geq\frac{(|\mathcal{T}|-|\widetilde{\mathcal{T}}|)}{2^{L+1}}S(\mathcal{T}\to\widetilde{\mathcal{T}})\geq\frac{(|\mathcal{T}|-|\widetilde{\mathcal{T}}|)}{2^{L+1}\times 2^{L+1}}\geq\frac{1}{2^{2L+2}}. \tag{84}$$

The other parts of the proof in Lemma 5 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), (67), 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 (83) is the same as the informed case in Lemma 5. The proof of Lemma 5 does not depend on a specific choice of $S(\cdot\to\cdot)$ but uses only (72), which is obtained by (67). Since the non-informed version shares all the same movement neighbor and posterior ratios, the proof also results in (83) in the current lemma.

Lemma 10. *Under the same assumptions of Theorem 5, 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 $\mathcal{T}\not\supset\mathcal{T}^*$,*

$$\begin{aligned}\sum_{\widetilde{\mathcal{T}}\in\mathcal{N}_g(\mathcal{T})}R_1(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}})&\leq-\frac{\delta_1A^2\log^2 n}{2^{2L+2}C_{f_0}^2\,n},\\\sum_{\widetilde{\mathcal{T}}\in\mathcal{N}_p(\mathcal{T})}R_1(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}})&\leq\frac{\mathrm{e}-1}{2n^{(A^2/8\log n-1)}},\end{aligned}$$

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

Proof. **The GROW movement.** Consider first the case of the Bayesian CART. As in Lemma 8, by (69) 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 (70), and with a decomposition $P_{\mathcal{G}(\mathcal{T})} = P_{\mathcal{T}} + P_{\mathcal{G}(\mathcal{T})\setminus\mathcal{T}}$,

$$
\begin{aligned}
-R_1(\mathcal{T},\mathcal{G}(\mathcal{T})) &\geq \frac{1}{2}\frac{1}{C_{f0}^2 n\, 2^L}\left((\boldsymbol{X}\boldsymbol{\beta}^*)'(P_{\mathcal{G}(\mathcal{T})}/n - P_{\mathcal{T}}/n)\boldsymbol{X}\boldsymbol{\beta}^*\right) \\
&\geq \frac{1}{2^{L+1}}\frac{1}{C_{f0}^2 n}\left((\boldsymbol{X}\boldsymbol{\beta}^*)'(P_{\mathcal{G}(\mathcal{T})\setminus\mathcal{T}}/n)\boldsymbol{X}\boldsymbol{\beta}^*\right) \\
&\geq \frac{1}{2^{L+1}\, n\, C_{f0}^2} n \frac{A^2\log^2 n}{n} = \frac{A^2\log^2 n}{2^{L+1}\, n\, C_{f0}^2}.
\end{aligned}
$$

Besides $R_1(\mathcal{T},\widetilde{\mathcal{T}}) \leq 0$ for all $\widetilde{\mathcal{T}} \in N_G(\mathcal{T})$ by Lemma 3 (ii). Therefore, by considering this movement to $\mathcal{G}(\mathcal{T})$, we obtain

$$
\begin{aligned}
-\sum_{\widetilde{\mathcal{T}}\in\mathcal{N}_g(\mathcal{T})} R_1(\mathcal{T},\widetilde{\mathcal{T}})P(\mathcal{T},\widetilde{\mathcal{T}}) &\geq -R_1(\mathcal{T},\mathcal{G}(\mathcal{T}))P(\mathcal{T},\mathcal{G}(\mathcal{T})) \qquad (85) \\
&\geq \frac{A^2\log^2 n}{2^{L+1}\, n\, C_{f0}^2} P(\mathcal{T},\mathcal{G}(\mathcal{T})) \geq \frac{A^2\log^2 n}{2^{2L+2}\, n\, C_{f0}^2}.
\end{aligned}
$$

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 (67),

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

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 (80), 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 (69), $B(\mathcal{T},\widetilde{\mathcal{T}}) \leq n^{-(A^2\log n)/8}$, and by Lemma 3 (i), $R_1(\mathcal{T},\widetilde{\mathcal{T}}) \leq \mathrm{e}-1$. From (67), 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 $\mathcal{T}\setminus\widetilde{\mathcal{T}}$ are non-signals, and second when $\mathcal{T}\setminus\widetilde{\mathcal{T}}$ contains at least one signal. In these cases, the above reasoning applies in the same way.

Remark 11. 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 upper bounded by $S(\mathcal{T} \to \widetilde{\mathcal{T}}) \leq 1/2^L$ in the non-informed algorithms. This contrast is highlighted by comparing the large proposal probability bound in (71) with the small lower bound (of a uniform proposal) in (84). 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 (79) 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, (79) 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 9, we considered only a single best movement when handling underfitted tree cases (GROW) as in (85). Note that this consideration was unavoidable. In the non-informed setting, like in the informed setting, guaranteeing multiple 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) is $\leq 1/2^{L+1}$. Therefore, the upper bound in Remark 9 slower than that of the informed algorithms is not because only a single movement was considered in (85). Rather, this is due to the difference in the proposal distributions.

15 Comparison to [53]

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

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

where $X \in \mathbb{R}^{n\times p}$ 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 [53] and compare them with our settings. Denote the design matrix of the selected columns by $X_{\boldsymbol{\gamma}} \in \mathbb{R}^{n \times |\boldsymbol{\gamma}|}$. The Bayesian hierarchical model considered in [53] is

$$\begin{aligned} \boldsymbol{\omega} &\sim \mathcal{N}(0, \phi^{-1} I_n) \\ \pi(\phi) &\propto \frac{1}{\phi}, \\ \boldsymbol{\beta}|\boldsymbol{\gamma} &\sim \mathcal{N}(0, g\phi^{-1}(X'_{\boldsymbol{\gamma}} X_{\boldsymbol{\gamma}})^{-1}), \\ \Pi(\boldsymbol{\gamma}) &\propto \left(\frac{1}{p}\right)^{\kappa|\boldsymbol{\gamma}|} \mathbb{I}[|\boldsymbol{\gamma}| \leq s_0], \end{aligned} \tag{86}$$

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 [53] ((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 $X'X = nI_p$ and Assumption 1, these hyperparameter settings meet their regularity conditions (Assumption A to D in [53]) 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 $\|\frac{1}{\sqrt{n}} X\boldsymbol{\beta}^*\|_2^2 \leq \log n$. 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 $\frac{1}{n} X'X = I_p$. Due to the orthogonality of X as discussed in [53], 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 [53], 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 $\omega := \max_{\mathcal{T}} \|(X'_{\mathcal{T}} X_{\mathcal{T}})^{-1} X'_{\mathcal{T}} X_{\mathcal{T}^* \setminus \mathcal{T}}\|_{\text{op}}^2$. 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 $L \leq L_{max} - \log_2 L_{max} - 4$.

Therefore, the rapid mixing guarantee (Theorem 2 in [53]) is translated as follows.

Theorem 7 ([53] Theorem 2)**.** *Assume the model* (3) *with Assumption 1 and the spike-and-slab prior in* (86) *with* $\kappa = c \geq 17 + 1/2$*. Consider the Spike-and-Slab MH algorithm in [53] without a tree structure restriction (*γ *is the vectorized* $\mathcal{T} \in \mathbb{T}_L$*). Assume* $C_{f_0}^2 2^L \leq \log(n/2)$ *and* $1 \leq L \leq L_{max} - \log_2 L_{max} - 4$*. With a large enough constant* $A > 0$*, with probability at least* $1 - c_3 p^{-c_4}$*,*

$$\tau_\epsilon \leq 3 \times 2^{2L} n \left[n \log(n/2) + (1 + 4c) 2^L \log(n/2)) + \log(2/\epsilon) \right], \tag{87}$$

for some c_3*, and* c_4*.*

Now, for the comparison purpose, we match the settings by applying the sparsity prior in (86) to our result instead of the classical Bayesian CART prior in 2.1.1. That is, for the comparison, we use the prior $\Pi(\mathcal{T}) \propto \widetilde{p}_{lk}^{(-|\mathcal{T}_{int}|)} \mathbb{I}[|\mathcal{T}_{int}| \leq 2^L]$, 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 1 and Theorem 3 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 (87).

16 Additional Visualizations

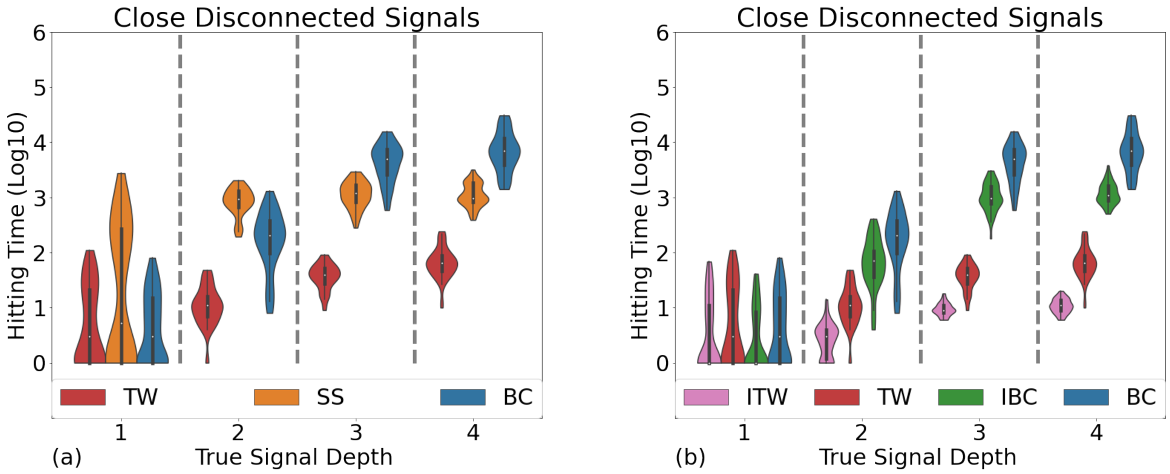


Figure 10: Hitting time $\tau = \min_{t \geq 0}\{\mathcal{B} \subset \mathcal{T}^t_{int}\}$ when true tree gets deeper of Case (3) (by gradually making the deeper part of the tree). (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: original and informed Twiggy Bayesian CART, ss: Spike-and-Slab 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.

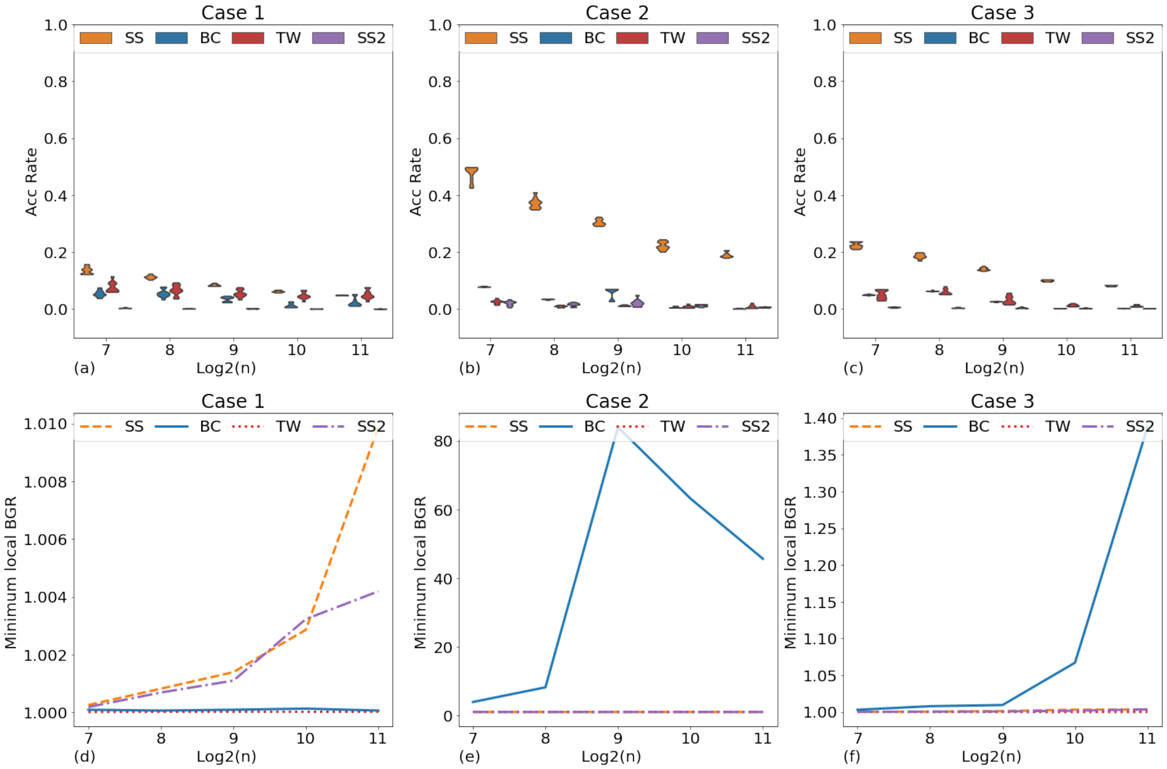


Figure 11: The acceptance rates and minimum local BGRs. (Legend) BC: Bayesian CART, TW: Twiggy Bayesian CART, SS and SS2: Spike-and-Slab with prior p_1^{ss} and p_2^{ss} respectively.

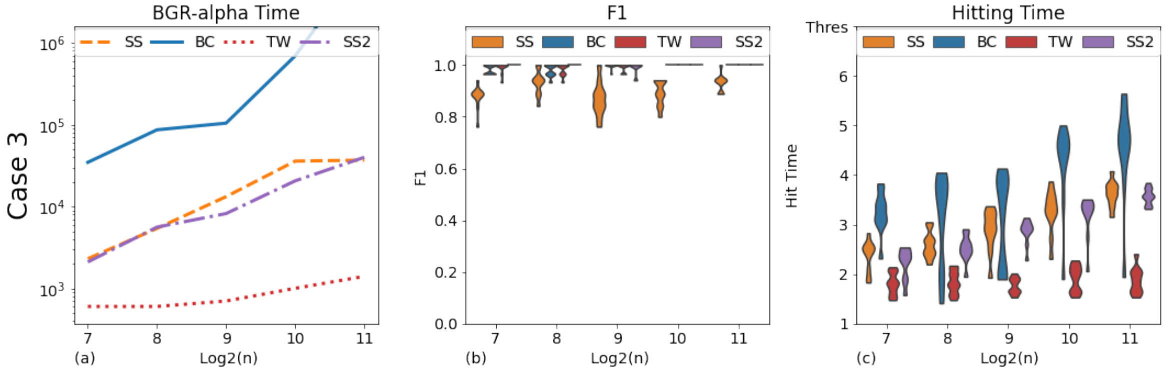


Figure 12: The MCMC performance measures for Case (3). (Legend) BC: Bayesian CART, TW: Twiggy Bayesian CART, SS and SS2: Spike-and-Slab with prior p_1^{ss} and p_2^{ss} respectively.

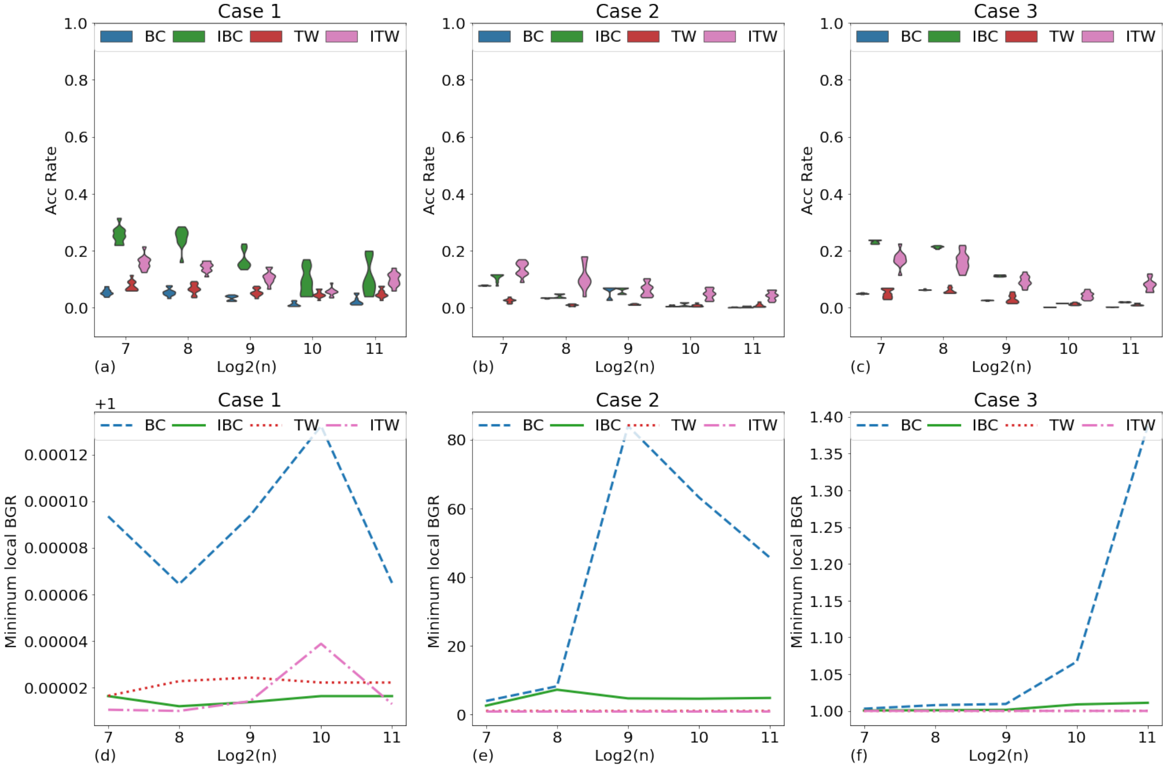


Figure 13: The acceptance rates, hit time and minimum local BGRs for Case (1) and (2). (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: original and informed Twiggy Bayesian CART.

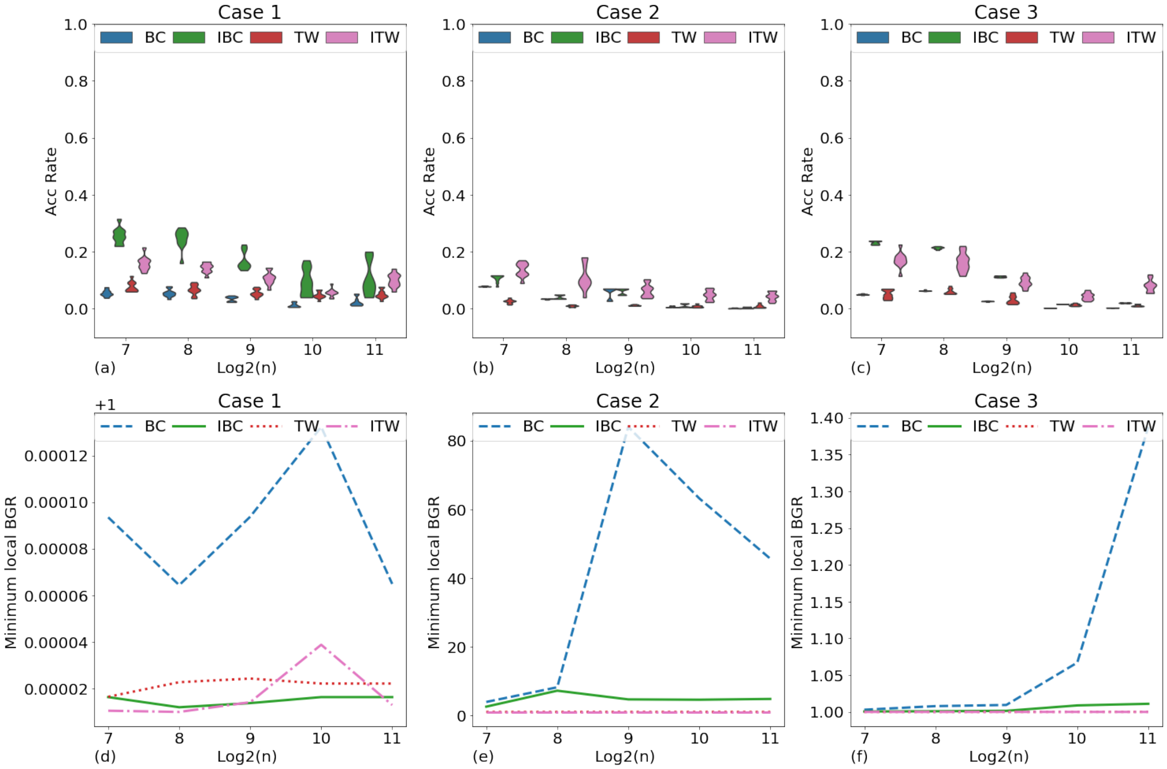


Figure 14: The MCMC performance measures for Case (3). (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: original and informed Twiggy Bayesian CART.

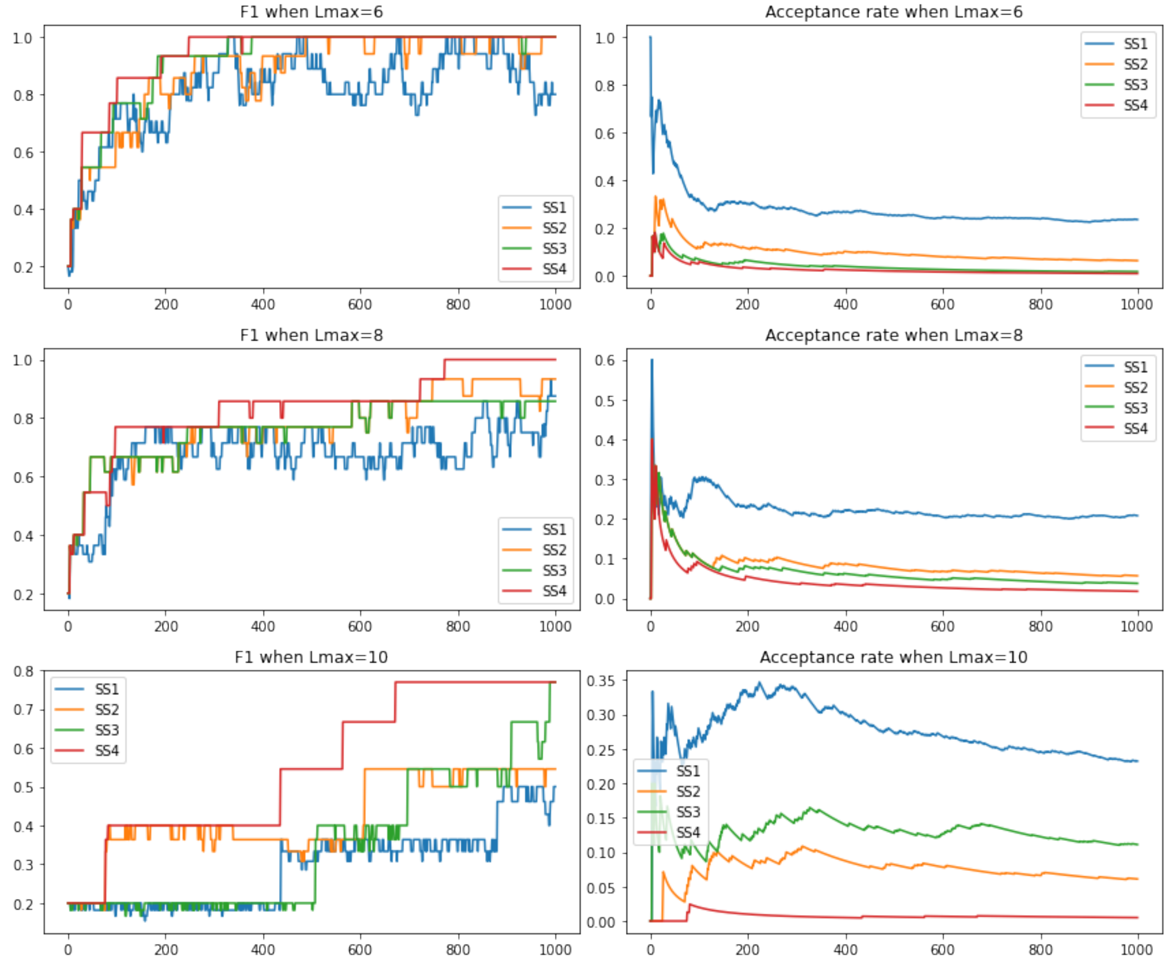


Figure 15: The behavior of Spike-and-Slab for Case (3) for different node inclusion priors for increasing data size (L_{max}). The x-axis in all plots are the number of iterations. SS1: $p_{lk} = 0.25/2^{L_{max}-6}$. SS2: $p_{lk} = 0.05/2^{L_{max}-6}$. SS3: $0.01\,n^{1/4}2^{-l/2}$. SS4: $0.01\,n^{1/4}6^{-l/2}$. 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.

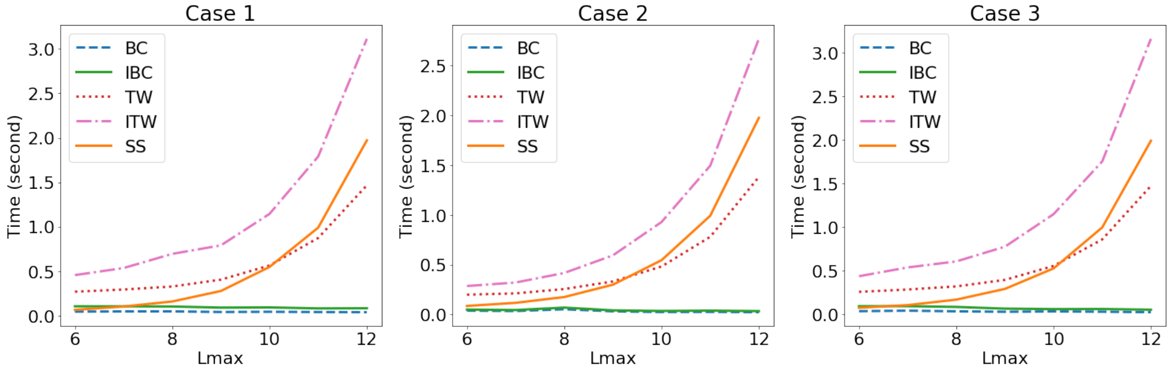


Figure 16: The computational times. Informed Bayesian CART is generally slower than Bayesian CART due to the time calculating the proposal probabilities (e.g., (16)).

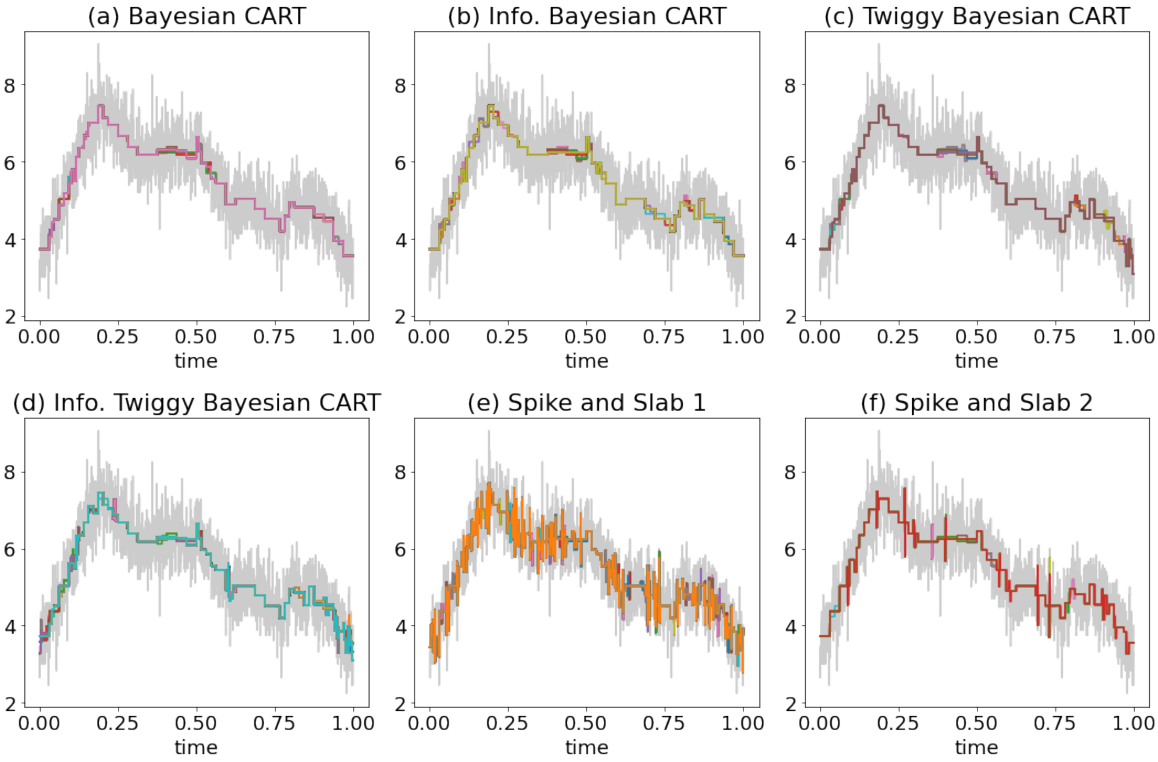


Figure 17: The visualization of 1000 samples after 10,000 burn-in of the MCMC chains on Call Center Data. The gray lines are the data. (a) Bayesian CART (b) informed Bayesian CART (c) Twiggy Bayesian CART (d) informed Twiggy Bayesian CART (e) Spike-and-Slab (prior: $p_{lk}^{ss,1} = 0.01$) (f) Spike-and-Slab (prior: $p_{lk}^{ss,2} = 0.01 \times 6^{-l/2}$)

References

[1] D. Applegate and R. Kannan. Sampling and integration of near log-concave functions. In Proceedings of the twenty-third annual ACM symposium on Theory of computing, pages 156–163, 1991.

[2] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. IEEE Transactions on information theory, 56(4):1982–2001, 2010.

[3] A. Belloni and V. Chernozhukov. On the computational complexity of mcmc-based estimators in large samples. The Annals of Statistics, 37(4):2011–2055, 2009.

[4] L. Breiman. Classification and regression trees. Routledge, 2017.

[5] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. Journal of the American statistical association, 100(469):36–50, 2005.

[6] T. T. Cai, M. Low, and Z. Ma. Adaptive confidence bands for nonparametric regression functions. Journal of the American Statistical Association, 109(507):1054–1070, 2014.

[7] N. B. Carnegie. Comment: Contributions of model features to bart causal inference performance using acic 2016 competition data. Statistical Science, 2019.

[8] I. Castillo and V. Ročková. Uncertainty quantification for bayesian cart. The Annals of Statistics, 49(6):3482–3509, 2021.

[9] J. Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In Problems in analysis, pages 195–200. Princeton University Press, 2015.

[10] H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian cart model search. Journal of the American Statistical Association, 93(443):935–948, 1998.

[11] H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1):266–298, 2010.

[12] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden markov models. IEEE Transactions on signal processing, 46(4):886–902, 1998.

[13] D. G. Denison, B. K. Mallick, and A. F. Smith. A bayesian cart algorithm. Biometrika, 85(2):363–377, 1998.

[14] P. Diaconis and D. Stroock. Geometric bounds for eigenvalues of markov chains. The annals of applied probability, pages 36–61, 1991.

[15] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. biometrika, 81(3):425–455, 1994.

[16] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. Journal of the american statistical association, 90(432):1200–1224, 1995.

[17] V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. Statistical Science, 2019.

[18] A. Frieze, R. Kannan, and N. Polson. Sampling from log-concave distributions. The Annals of Applied Probability, pages 812–837, 1994.

[19] A. Frigessi, P. Di Stefano, C.-R. Hwang, and S.-J. Sheu. Convergence rates of the gibbs sampler, the metropolis algorithm and other single-site updating dynamics. Journal of the Royal Statistical Society: Series B (Methodological), 55(1):205–219, 1993.

[20] P. Fryzlewicz. Unbalanced haar technique for nonparametric function estimation. Journal of the American Statistical Association, 102(480):1318–1327, 2007.

[21] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. Statistical science, pages 457–472, 1992.

[22] M. Ghosh. Exponential tail bounds for chisquared random variables. Journal of Statistical Theory and Practice, 15(2):35, 2021.

[23] R. B. Gramacy and H. K. H. Lee. Bayesian treed gaussian process models with an application to computer modeling. Journal of the American Statistical Association, 103 (483):1119–1130, 2008.

[24] J. He and P. R. Hahn. Stochastic tree ensembles for regularized nonlinear regression. Journal of the American Statistical Association, pages 1–20, 2021.

[25] J. Hill, A. Linero, and J. Murray. Bayesian additive regression trees: A review and look forward. Annual Review of Statistics and Its Application, 7:251–278, 2020.

[26] S. Jeong and V. Rockova. The art of bart: On flexibility of bayesian forests. arXiv preprint arXiv:2008.06620, 2020.

[27] D. Jerison. The drift and minorization method for reversible Markov chains. Stanford University, 2016.

[28] M. Jerrum and A. Sinclair. Conductance and the rapid mixing property for markov chains: the approximation of permanent resolved. In Proceedings of the twentieth annual ACM symposium on Theory of computing, pages 235–244, 1988.

[29] B. Lakshminarayanan, D. Roy, and Y. W. Teh. Top-down particle filtering for bayesian decision trees. In International Conference on Machine Learning, pages 280–288. PMLR, 2013.

[30] B. Lakshminarayanan, D. Roy, and Y. W. Teh. Particle gibbs for bayesian additive regression trees. In Artificial Intelligence and Statistics, pages 553–561. PMLR, 2015.

[31] G. F. Lawler and A. D. Sokal. Bounds on the l^2 spectrum for markov chains and markov processes: a generalization of cheeger's inequality. Transactions of the American mathematical society, 309(2):557–580, 1988.

[32] T. Lindvall. Lectures on the coupling method. Courier Corporation, 2002.

[33] A. R. Linero. A review of tree-based bayesian methods. Communications for Statistical Applications and Methods, 24(6):543–559, 2017.

[34] A. R. Linero and Y. Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 80(5):1087–1110, 2018.

[35] L. Lovász and M. Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In Proceedings [1990] 31st annual symposium on foundations of computer science, pages 346–354. IEEE, 1990.

[36] L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. Random structures & algorithms, 4(4):359–412, 1993.

[37] L. Lovász and S. Vempala. Hit-and-run from a corner. In Proceedings of the thirty-sixth annual ACM symposium on Theory of computing, pages 310–314, 2004.

[38] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. Proceedings of the National Academy of Sciences, 116(42):20881–20885, 2019.

[39] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the hastings and metropolis algorithms. The annals of Statistics, 24(1):101–121, 1996.

[40] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. The journal of chemical physics, 21(6):1087–1092, 1953.

[41] E. Mossel and E. Vigoda. Phylogenetic mcmc algorithms are misleading on mixtures of trees. Science, 309(5744):2207–2209, 2005.

[42] J. Pitman. On coupling of markov chains. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 35(4):315–322, 1976.

[43] M. T. Pratola. Efficient metropolis–hastings proposal mechanisms for bayesian regression tree models. Bayesian analysis, 11(3):885–911, 2016.

[44] C. P. Robert, G. Casella, and G. Casella. Monte Carlo statistical methods, volume 2. Springer, 1999.

[45] V. Rockova and J. Rousseau. Ideal bayesian spatial adaptation. Journal of the American Statistical Association (minor revision), 2023.

[46] V. Ročková and E. Saha. On theory for bart. In The 22nd international conference on artificial intelligence and statistics, pages 2839–2848. PMLR, 2019.

[47] O. Ronen, T. Saarinen, Y. S. Tan, J. Duncan, and B. Yu. A mixing time lower bound for a simplified version of bart. arXiv preprint arXiv:2210.09352, 2022.

[48] J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. IEEE Transactions on signal processing, 41(12):3445–3462, 1993.

[49] A. Sinclair. Improved bounds for mixing rates of markov chains and multicommodity flow. Combinatorics, probability and Computing, 1(4):351–370, 1992.

[50] D. D. Sleator, R. E. Tarjan, and W. P. Thurston. Rotation distance, triangulations, and hyperbolic geometry. Journal of the American Mathematical Society, 1(3):647–681, 1988.

[51] D. B. Woodard and J. S. Rosenthal. Convergence rate of markov chain methods for genomic motif discovery. The Annals of Statistics, 41(1):91–124, 2013.

[52] Y. Wu, H. Tjelmeland, and M. West. Bayesian cart: Prior specification and posterior simulation. Journal of Computational and Graphical Statistics, 16(1):44–66, 2007.

[53] Y. Yang, M. J. Wainwright, and M. I. Jordan. On the computational complexity of high-dimensional bayesian variable selection. The Annals of Statistics, 44(6):2497–2532, 2016.

[54] G. Zanella. Informed proposals for local mcmc in discrete spaces. Journal of the American Statistical Association, 115(530):852–865, 2020.

[55] Q. Zhou, J. Yang, D. Vats, G. O. Roberts, and J. S. Rosenthal. Dimension-free mixing for high-dimensional bayesian variable selection. arXiv preprint arXiv:2105.05719, 2021.