Electronic Journal of Statistics
Vol. 19 (2025) 3041–3067
ISSN: 1935-7524
https://doi.org/10.1214/25-EJS2397

On mixing rates for Bayesian CART

Jungeum Kim and Veronika Ročková

The Booth School of Business, University of Chicago,
e-mail: jungeum.kim@chicagobooth.edu; Veronika.Rockova@chicagobooth.edu

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 superpolynomial 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.

Keywords and phrases: Bayesian CART, non-parametric regression, MCMC, mixing rates.

Received July 2024.

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 [44]. 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 [61]). This paper aims to create similar reasons for optimism (as well as caution) in the context for Bayesian tree-based regression.

Bayesian tree-based regression (Bayesian CART of [15, 12] and BART of [13]) is one of the most popular machine learning tools in practice today. A host of frequentist theory now exists to certify their inferential validity [9, 52, 31, 40]. 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 [15, 12] 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 [51]) have been unavailable.

There is an apparent disconnect between theory for optimization and sampling [44] and between theory for posterior distributions and their MCMC approximations. Bayesian CART implementations [15, 12] are instantiations of the Metropolis-Hastings (MH) algorithm [46] 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 [12, 60, 49, 39, 28]. Several enhancements have been proposed such as modifications of the proposal [60, 49], "warm start" initializations [27], or running multiple chains [8, 20]. 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 [42, 21, 45]. Notably, [45] 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 [33, 16, 22]. Perhaps the first systematic approach to handling spectral bounds was developed in [37] using the conductance concept due to [10]. Conductance is a measure of edge expansion of the Markov chain, see [41] 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, [61] show the rapid mixing of Bayesian variable selection. This bound is improved in [64] in the context of informed MCMC (that uses posterior information in the proposal) using the drift condition of [32] rooted in the coupling inequality [48, 38]. 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 [9]. 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 [61] 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 [9]) where the MH proposal distribution consists of a simple attachment of a terminal node (GROW) or a detachment of two sister bottom nodes (PRUNE) [15, 12]. This algorithm is a simplified version of the BART method and we analyze it here to identify computational bottlenecks that will likely carry forward to the full-blown BART version. We reconstruct the regression function using Haar wavelets with a tree-shaped sparsity structure on the wavelet coefficients. This prior design was studied thoroughly by [9] and [52]. It turns out that, for a given tree, independent product wavelet

coefficient priors correspond to the construction proposed originally by [11]. While we use wavelets more as a technical framework to articulate these computational bottlenecks, the wavelet method alone can be of independent interest. For example, in the context of uncertainty quantification for non-parametric regression with spatially varying smoothness, this algorithm can be used to compute locally adaptive confidence band for inference [52]. The context of univariate nonparametric regression with wavelets is still of practical interest [30, 6, 52, 58]. A referee pointed out a connection to vertical blocking [7] which jointly thesholds out wavelet coefficients that are related according to a family tree. We do acknowledge that, for purely univariate wavelet analysis, the Bayesian CART prior is inferior to the Spike-and-Slab prior [29] both in terms of the rates of posterior convergence (missing a logarithmic factor [9]) and computation (exact posterior analysis available). Our goal was to understand computational bottlenecks of Bayesian CART and BART on this one-dimensional version.

Rapid mixing rate bounds in [61] and [64] 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 (a superpolynomial 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. There is a conceptual connection to the vertical blocking approach [7] where adding a twig corresponds to adding a vertical block of wavelet coefficients. 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 [61]. 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 [64], we show linear mixing of Markov chains under locally informed proposals.

Recently, independently from our work, [51] 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 [51] 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. Extending [51], [57] conducts hitting time analysis in additive tree scenarios, where high posterior density regions are identified using BIC, yielding lower bounds that are at most polynomial in n. Even after disregarding differences in the assumptions of the data-generating models, we do not see a contradiction between their results and ours.

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 [19, 9], where the underlying tree provides a skeleton for signal coefficients. This regression re-interpretation of Bayesian CART yields easy implementation of the Bayesian CART algorithm through closed-form tree posterior probabilities. Tree-shaped wavelets were shown to have favorable theoretical properties by [9, 52] and their prior development dates back to at least [11].

2.1. Trees as wavelets

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

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

where $\mathcal{X} = \{x_i = i/n : 1 \le i \le n\}$ are fixed observations on a regular grid. The assumption $x_i = i/n$ could be avoided using either unbalanced Haar wavelets [23] or regularity relaxations [52]. 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.

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

From $n = 2^{L_{max}+1}$, we limit the resolution of Haar wavelets up to $\log_2 n - 1$. This ensures that both the positive and negative supports of all the wavelets considered have at least one observation matched to each of them. We assume that the columns of $\boldsymbol{X}$ have been ordered according to the index $2^l + k$ (increasing ordering). We denote with $F_0 = (f_0(x_1), \ldots, f_0(x_n))'$ the vector of realized values of the true regression function at design points. The nonparametric 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 [9], as we re-iterate in Section 2.1.1 below.

Definition 2.1 (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 $j = 0, \ldots, l-1$. Given l, the range of k is $0 \le k \le 2^l - 1$. We distinguish between two types of nodes: *internal* ones $\mathcal{T}_{int} = \{(l,k) \in \mathcal{T} : \{(l+1, 2k), (l+1, 2k+1)\} \in \mathcal{T}\} \cup (-1, 0)$ 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 with maximal depth L is defined as $\mathcal{T}^L_{full} = \{(l,k) : l \le L\}$. [1]

[1] The maximal depth L includes the external nodes. That is, for external nodes, levels $l \le L$, and for internal nodes, $l \le L - 1$. Therefore, $\mathcal{T}^L_{full}$ has 2^L external nodes, which are at the level L, and 2^L internal nodes.

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 (there are $|\mathcal{T}_{ext}|$ of those) and with $\boldsymbol{\beta}_{\setminus\mathcal{T}}$ 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 $\boldsymbol{X}_{\setminus\mathcal{T}}$.

2.1.1. The Bayesian CART posterior

The distinguishing feature of Bayesian CART, compared to selective wavelet reconstructions such as *RiskShrink* of [18], 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 [9]

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

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

Similarly as in [52], 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. This prior on $\boldsymbol{\beta}_{\mathcal{T}}$ differs from the usual wavelet priors [29] because it assumes that the active wavelet coefficients are a-priori correlated, yielding an independent product prior on the bottom tree node coefficients (see [9] for more discussion). This prior corresponds to what is done in practice [12, 13]. Alternatively, we could have considered a-priori independent wavelet coefficients which would have given rise to correlated step heights in the tree reconstruction of the regression function (as is done in [11]). We expect that our theoretical results can be extended to this prior setting, as our analysis primarily focuses on the posterior distribution of trees $\mathcal{T}$ after integrating out $\boldsymbol{\beta}_{\mathcal{T}}$.

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 [12] uses the heterogeneous Galton-Watson (GW) process (see Section 2.1 in [9] and [53]) 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 [12] 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

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

The Bayesian CART posterior has many favorable properties, such as near-minimax rate adaptation under the supremum loss [9] 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 [15, 12], 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 [54, 14]. 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 [52, 9], 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, 43, 61, 64]. While our theory has been derived for the regular fixed design, similar theoretical conclusions can be obtained also for fixed irregular design as in [52] using the unit information g-prior.

Theorem 2.2 (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 S2.

Remark 3. In the context of Bayesian inference with phylogenetic trees, [47] 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. Theorem 2.2 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 2.2 is different from posterior concentration rate results in [9] and [52] 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 Theorem 2.2 because we cannot leverage the decaying property of wavelet coefficients.

Unfortunately, unlike for certain independent product priors [29], Bayesian wavelet analysis with the CART tree prior does not admit exact posterior computation due to the combinatorial intractability of the normalizing constant. Much of the value of the optimality properties of the Bayesian CART posterior (e.g. adaptation to local smoothness [52] and frequentist validity of inference about certain f_0 [9]) thereby 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 [15, 12] 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 [26, 49], to better explore the tree space.

The Bayesian CART algorithm generates a chain of trees $\mathcal{T}^0, \mathcal{T}^1, \ldots$ 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

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

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

Under weak conditions (Section 7.4 of [50]), the sequence obtained by this algorithm will be an irreducible and aperiodic Markov chain with a limiting distribution $\Pi(\mathcal{T} \mid Y)$. Below, we will describe a *dyadic* one-dimensional version of Bayesian CART [12] 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 in Section S1. 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

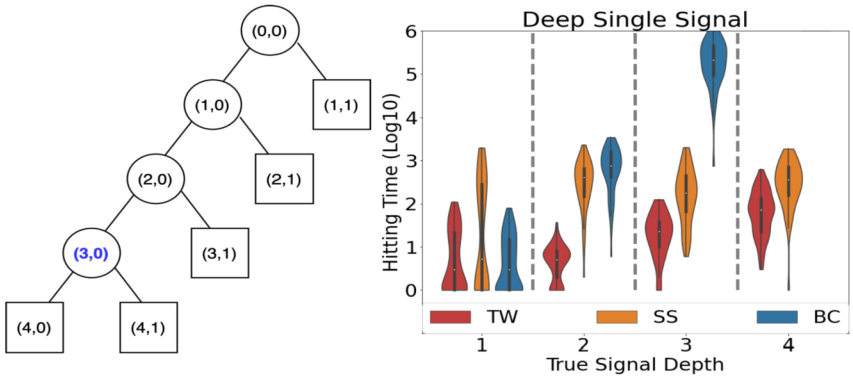
$$S_{GROW}(\mathcal{T}^i \to \widetilde{\mathcal{T}}) = \frac{1}{|\mathcal{T}^i_{ext}|}. \tag{10}$$

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

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



(a) (b)

Fig 1. *(a) An example of the minimal spanning tree $\mathcal{T}^*$ from Example 3 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.*

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 [52]. 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 [12, 60, 49, 39, 28], limited theoretical studies of the mixing times have been available [51].

3. Bayesian CART with a twist

Local Metropolis-Hastings proposals are known to induce poor mixing [60, 49] which may result in misleading under-representations of uncertainty. In the context of trees, [26] remediate this issue by applying a rotation algorithm [56] while [49] proposes various elaborate moves for radical restructuring (see also [60]). 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 [13]). Alternatives to MH samplers have also been recently explored, see [35] for Sequential Monte Carlo approach and [36] 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 (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, we measure the hitting time

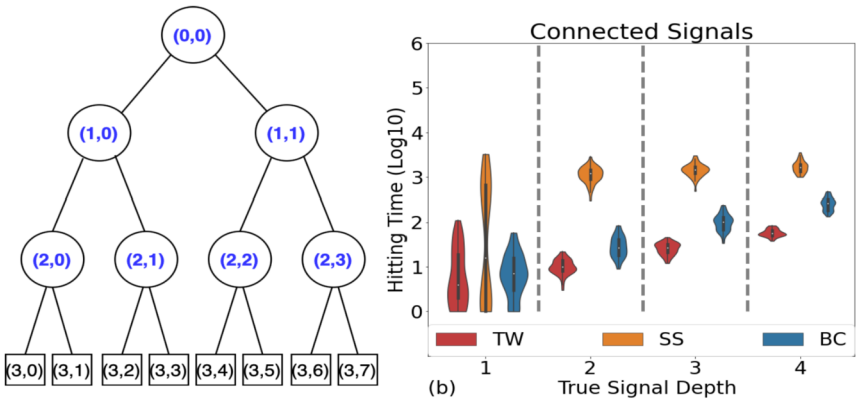


Fig 2. *(a) An example of the minimal spanning tree $\mathcal{T}^*$ from Example 3 with oval internal and square external nodes. (b) The hitting time of 50 chains initialized at* $(0, 0)$. *TW: Twiggy Bayesian CART, SS: Spike-and-Slab, BC: Bayesian CART.*

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 3 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 (The Benefits of Bayesian CART)**.** In contrast with Example 3, 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 3. 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 3, 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.

Definition 3.1 (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 $A_{lk}(\mathcal{T}) = \{(l',k') \in \mathcal{T}_{int} : \exists j \in \{0,1,\ldots,L-1\}\ s.t.\ (l',k') = (l-j, \lfloor k/2^j \rfloor)\}$. Descendants of (l,k) inside $\mathcal{T}$ are defined as $D_{lk}(\mathcal{T}) = \{(l',k') \in \mathcal{T}_{int} : (l,k) \in A_{l'k'}(\mathcal{T})\}$.

Definition 3.2 (Twig)**.** For two distinct nodes (l,k) and (l',k') where (l,k) is an ancestor of (l',k') and $l \le l' < L$, 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 $[(l,k) \leftrightarrow (l,k)] = \{(l,k)\}$.

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

Remark 5. Growing a twig to reach deeper signals is conceptually related to vertical blocking proposed by [7] in the context of wavelet thresholding. The wavelet coefficients within a group are jointly thresholded based on the magnitude of the sum of squares for this entire block. Likewise, for our twiggy grow movement, the proposed newly added twiggy is accepted with a higher probability if the joint posterior gain is large enough.

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 proposal within the chosen layer, i.e.

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

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

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

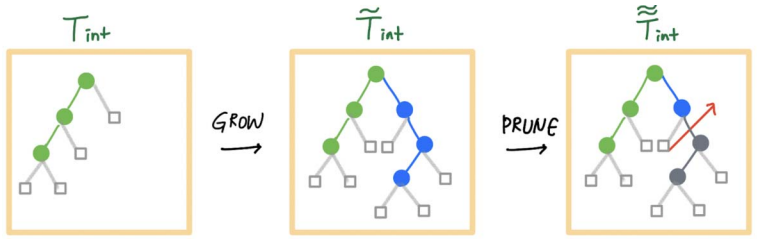


Fig 3. *The Twiggy GROW and PRUNE.*

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

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

In actual implementation, computing S_{PRUNE} is easy due to the tree data structure used for coding the binary tree (each node is connected to others through pointers to its parent and children). We can propagate from each pre-terminal node to the root until we encounter the first ancestor having two children. A cartoon of the twig proposals is depicted in Figure 3. It can be easily verified that the Twiggy Bayesian CART also yields a Markov Chain that is both irreducible (i.e., all states communicate; see Section 6.3.1 in [50]) and reversible. 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, [62] 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, [64] 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, [64] 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

$$S_{GROW}(\mathcal{T}^i \to \widetilde{\mathcal{T}}) = \frac{w_g(\widetilde{\mathcal{T}} \mid \mathcal{T}^i)}{Z_g(\mathcal{T}^i)} \mathbb{I}_{\mathcal{N}_g(\mathcal{T}^i)}(\widetilde{\mathcal{T}}),$$

$$S_{PRUNE}(\mathcal{T}^i \to \widetilde{\mathcal{T}}) = \frac{w_p(\widetilde{\mathcal{T}} \mid \mathcal{T}^i)}{Z_p(\mathcal{T}^i)} \mathbb{I}_{\mathcal{N}_p(\mathcal{T}^i)}(\widetilde{\mathcal{T}}), \tag{15}$$

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 4.1. 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 $\Pi(\cdot \,|\, Y)$ 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 ([55], equation 2.9 in [59])

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

For our theoretical study, we will work with a modified transition matrix (as suggested in [55]) 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 [33]. In particular, Theorem 2 in [55] shows that in a reversible Markov chain

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

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* [55].

Definition 4.2 (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 an edge e in G, define $\mathcal{S}(e) = \{(\mathcal{T}, \mathcal{T}') : e \in T_{\mathcal{T},\mathcal{T}'}\}$ the set of tree pairs whose canonical path in $\mathcal{E}$ contains the edge e. With a slight abuse of notation, we say $e \in \mathcal{E}$ if $\mathcal{S}(e) \neq \emptyset$.

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 [55])

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

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

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

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 S4.1 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 3 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 3, we show that the mixing lower bound increases exponentially in n.

Theorem 5.1. *Assume the model* (1) *with the Bayesian CART prior from Section 2.1.1 with* $L \geq 2$ *and* $p_{1k} = n^{-c}$ *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* $C > 1$

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

which is exponential in the signal depth L *and superpolynomial in* n *when* $L = L_{max} \sim \log(n/2)$.

Proof. Section S3.

Remark 6. In work independent from ours, [51] provided a lower bound result showing that a simplified version of BART [13] mixes poorly (at least exponentially in n). In particular, [51] 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.

Note that the mixing time in Theorem 5.1 is the worst-case scenario over all initializations. A referee suggested the possibility of computing a tree point estimate using the dynamic programming approach in [17] which connects the classical CART algorithm with best orthogonal basis search. Such a point estimate could be a good initialization for which faster mixing rate could be achieved. We have used the CART algorithm for the design of the canonical paths but it could be also used as an initialization.

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 5.2. *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 S4.2.

Remark 7. 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 $\mathbb{P}_{f_0}(\mathbb{T}^c_L) = o(1)$.

Remark 8 (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 (MH) proposal. [61] 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 attempted to rephrase their result in the context of our wavelet regression matrix where $p = n/2$ (Theorem S1 in the Supplement Section S7 [34]). The point-mass Spike-and-Slab prior in [61] 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. The mixing upper bound is slower for this Spike-and-Slab prior when the signal has a tree structure. This comparison should not be overinterpreted because it does not necessarily imply inferior performance for which a lower bound would be needed.

In addition, adaptive Bayesian wavelet analysis with Spike-and-Slab priors involves prior inclusion probabilities decaying with the depth of the wavelet coefficient [29, 52]. Our Spike-and-Slab practical comparisons use this decaying prior while our theory (namely Theorem S1) uses the prior in [61]. Our independent product Spike-and-Slab prior does not require MCMC due to the availability of the posterior as an independent product of mixtures. We use the one-site MH purely to assess how the tree structure obstructs MH convergence. While the one-site MH is not needed for orthogonal regression with independent product Spike-and-Slab priors, for hierarchical priors with a prior on the level of sparsity, the MH algorithm can be used to approximate the posterior which corresponds to a global mixture of a product of local mixtures.

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 5.1). 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 [61] when signal is obscured by layers of noise.

Theorem 5.3. *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* $1 - 4/n$

$$\tau_\epsilon \leq \frac{(D^L - 1)}{D - 1} \times 2^{2L+3} \left\{ n \left[\left(c + \frac{1}{2} \right) \log(1 + n) + |\mathcal{T}^*_{int}| C^2_{f_0} + 1 \right] + 4\,|\mathcal{T}^*_{int}| \log n + \log\left(\frac{2}{\epsilon} \right) \right\} \tag{23}$$

Proof. See Section S5.

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 5.4. *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 S6.

Remark 9. Theorem 5.4 provides at most linear in n mixing for Bayesian CART *only* under Assumption 1 (a). The superpolynomial mixing rate lower bound in Theorem 5.1 still applies to the informed Bayesian CART[2]. Therefore, the proposal informativeness alone does not solve the myopic problem of Bayesian CART.

Remark 10. It is worthwhile to point out that the linear mixing in Theorem 5.4 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. However, this improvement is limited to asymptotic settings, while the bound (22) is exact. For more explanation, see the discussion in Remark S2 in Supplementary Material [34]. The proof is in Section S6.4.

6. Performance evaluation

We compare Bayesian CART, Twiggy Bayesian CART (with $D = 2$) and their informed versions[3] on simulated data as well as a real dataset. To appreciate the effect of tree-shaped regularization, we compare Bayesian CART with the Metropolis-Hastings one-site sampler for Spike-and-Slab priors.

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 S6 in Section S8, 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 [52] with which a Spike-and-Slab algorithm can achieve locally adaptive minimax rate.

[2]See that S16 in the proof is valid as long as the proposal neighbor is the same as the Bayesian CART.

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

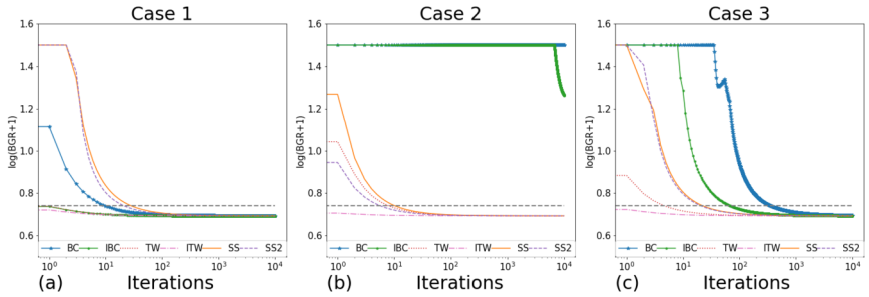


Fig 4. *The local BGRs (log-transformed) when* $n = 2^7$. *The local BGR values tend to decrease during the course of MCMC sampling. We have capped the values at a threshold* $log(Y + 1) = 1.5$ *for clearer visualization. The horizontal grey dotted line corresponds to the local BGR value 1.1. (Legend) 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.*

Performance measure We first define a proxy for the mixing time defined in (18) using BGR (Gelman-Rubin diagnostic) [24]. BGR measures the difference between in-chain and across-chain variability when considering multiple initializations. Formally, consider a collection of K chains $C = \{C_1, ..., C_K\}$. Denoting the L tree samples from each C_k by $\{\mathcal{T}_1^{C_k}, ..., \mathcal{T}_L^{C_k}\}$ and the j-th coefficient sample from tree $\mathcal{T}_i^{C_k}$ by $\beta_j(\mathcal{T}_i^{C_k})$, the BGR for β_j is

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

where $W_j = \frac{1}{K}\sum_{k=1}^K \frac{1}{L-1}\sum_{i=1}^L (\beta_j(\mathcal{T}_i^{C_k}) - \bar{\beta}_{jk})^2$ 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 $\bar{\beta}_{jk} = \frac{1}{L}\sum_{i=1}^L \beta_j(\mathcal{T}_i^{C_k})$ 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 $\hat{\beta}_j(\mathcal{T}_i^{C_k})$ instead of the sampled $\beta_j(\mathcal{T}_i^{C_k})$. 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 C_k by $\mathcal{T}_t^{C_k}$. Given $K = 10$ chains, a local BGR of the j^{th} indicator I_j at time t is defined as

$$\text{BGR}(j, C|t) = \text{BGR}(I_j|\{\mathcal{T}_t^{C_k}, \mathcal{T}_{t-1}^{C_k}, ..., \mathcal{T}_{t-99}^{C_k}\}_{k=1}^{10}). \tag{24}$$

As shown in Figure 4, as the iteration proceeds, the $\text{BGR}(j, C|t)$ 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

$$\tau_\alpha^{\text{BGR}} = 10^6 \wedge \arg\min_{t\geq 0}\{\max_{j\in S} \text{BGR}(j, C|t) \leq \alpha\}, \tag{25}$$

where S is the index set of true signals in the data generating mechanism. The inner maximum $\max_{j\in S}\text{BGR}(j, C|t)$ 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 [24].

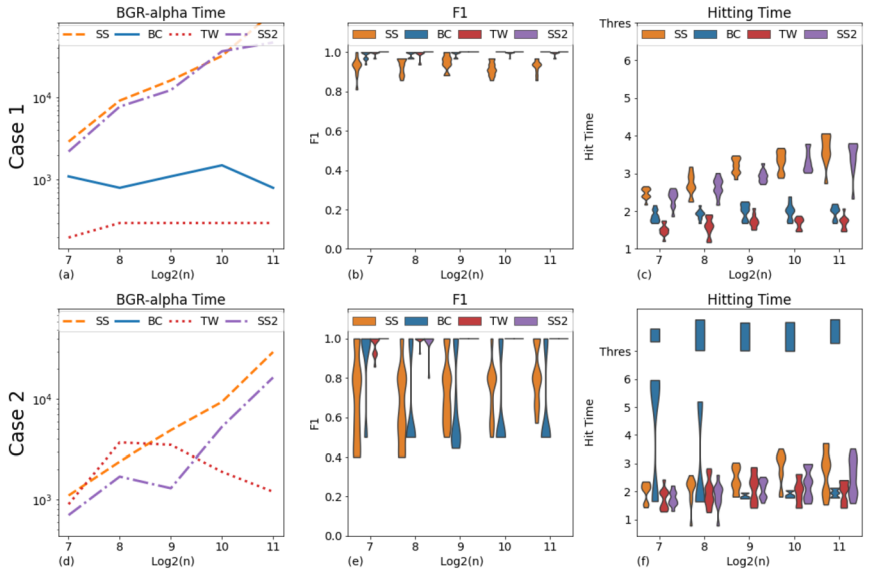


Fig 5. *Plots (a) and (d) depict time for achieving the worst case local BGR below 1.1. Plots (b) and (e) show F1 scores. The small F1 value of Spike-and-Slab with* p_1^{ss} *is due to low precision (overfit). Plots (c) and (f) show hitting times. (Legend) BC: Bayesian CART, TW: Twiggy Bayesian CART,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

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 (25) 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 $(l,k) \in \mathcal{T}_{int} \setminus \mathcal{T}^*_{int}$ as a *false positive* and $(l,k) \in \mathcal{T}^*_{int} \setminus \mathcal{T}_{int}$ as a *false negative*. 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 *not* 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.

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 5.2, 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 τ_α^{BGR} for $\alpha = 1.1$ in (25) 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 τ_α^{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 superpolynomial lower bound of Bayesian CART in Theorem 5.1 in the context of deep isolate signals as in Example 3. On the other hand, Theorem 5.3 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 τ_α^{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 (25) 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 S8 we additionally display the smallest local BGRs in Figure S2 (e). This value is obtained by running all the chains for 10^6 iterations and by taking the minimum over local BGRs defined in (24) 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 S5 in Section S8, we see that informed versions overall improve on their non-informed counterparts. The BGR-α (Figure 6 (a) and (d), and Figure S5 (a)) and hitting times (Figure 6 (c) and (f), and Figure S5 (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 S5 (b)). It might be understood as a trade-off of having a higher acceptance rate than Twiggy Bayesian CART (Figure S4). 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)).

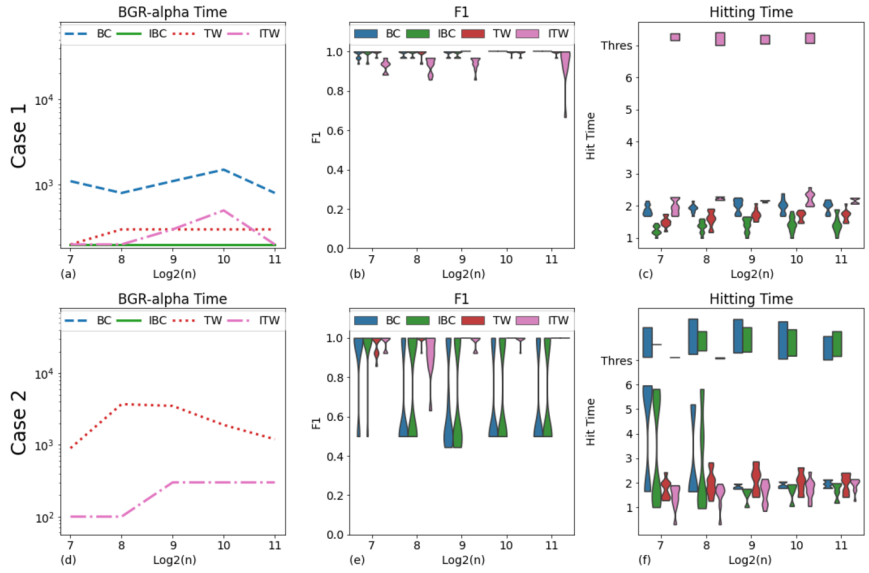


Fig 6. *The improvement of the informed variants. (a) and (d) Time for achieving the worst case local BGR below 1.1. (b) and (e) Informed Twiggy Bayesian CART tends to overfit than its non-informed version. (c) and (f) Hitting times. (Legend) BC and IBC: original and informed Bayesian CART, TW and ITW: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

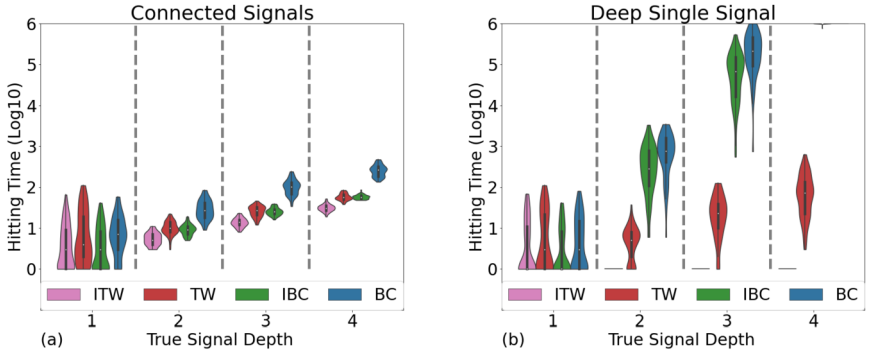


Fig 7. *Hitting time when the true tree gets deeper. Informed Bayesian CART hits the true signals faster than Bayesian CART. Likewise, the informed Twiggy Bayesian CART is faster than Twiggy Bayesian CART. (b) However, for an isolated deep signal, informed Bayesian CART does not hit faster than Twiggy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

When the signal depth deepens, in the same settings of Example 3 and Example 3, 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 9, 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 S1 (b). On the other hand, the informed Twiggy Bayesian CART includes the signal in a single step.

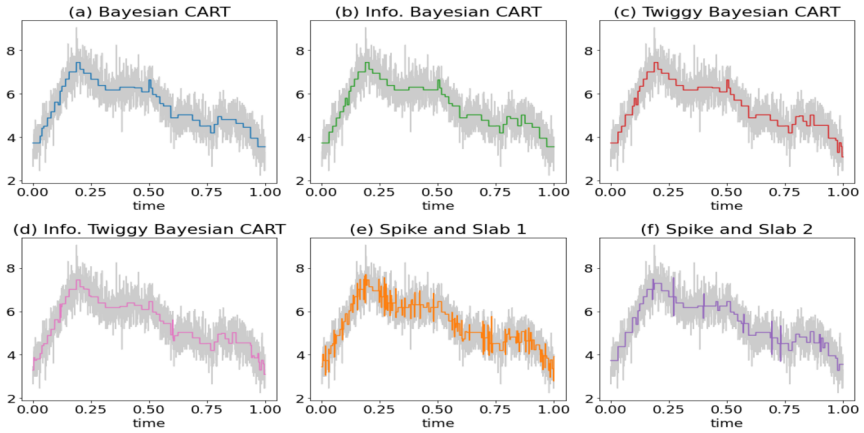


Fig 8. *The visualization of MCMC chains on Call Center Data. The colored lines are the median tree fit obtained from 1000 samples after 10,000 burn-in and the gray lines are the data. (a) Bayesian CART (b) informed Bayesian CART (c) Twiggy Bayesian CART (d) informed Twiggy Bayesian CART (e) Spike-and-Slab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

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, 52] 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 [52] where the response is $Y_i = \sqrt{N_i + 1/4}$ where N_i 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 $\max_j \text{BGR}(\beta_j | \{\mathcal{T}_t^{C_k}, \mathcal{T}_{t-1}^{C_k}, \ldots, \mathcal{T}_{t-99}^{C_k}\}_{k=1}^{10})$ 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, 52], 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

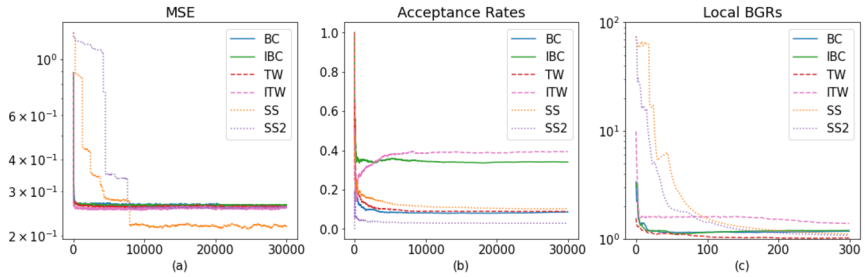


Fig 9. *The performance measures on the Call Center data. (a) The MSE (log transformed) over the MCMC iterations (b) Acceptance rates. (c) The local BGRs for every 100 iterations on the Call center data. The minimum local BGRs achieved were Bayesian CART: 3.58, informed Bayesian CART: 1.21, Twiggy Bayesian 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.*

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 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, where trees can represent the same partition of the space, composing an equivalence class sharing the same likelihood. The non-identifiability prevents us from guaranteeing posterior consistency (e.g., Theorem 2.2), which is crucial in the canonical path argument. If the posterior is concentrated on the equivalence class of the true tree, the chain at convergence should sample the trees that belong to that equivalence class. However, standard Bayesian CART does not naturally incorporate movements within an equivalence class. One possible approach to promote mixing would be to explicitly incorporate movements between equivalence classes [63]. We leave this extension for future research. 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.

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

The second author was supported by NSF Grant DMS-1944740.

Supplementary Material

Supplementary Material for "On Mixing Rates for Bayesian CART"
(doi: 10.1214/25-EJS2397SUPP; .pdf). This supplementary material contains the description of the Bayesian CART Algorithm as well as proofs of all the theorems in the main text.

References

[1] Applegate, D. and Kannan, R. Sampling and integration of near log-concave functions. In *Proceedings of the Twenty-Third Annual ACM Symposium on Theory of Computing*.

[2] Baraniuk, R. G., Cevher, V., Duarte, M. F. and Hegde, C. (2010). Model-based compressive sensing. *IEEE Transactions on Information Theory* **56** 1982–2001. MR2654489

[3] Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics* **37** 2011–2055. MR2533478

[4] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group. MR0726392

[5] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* **100** 36–50. MR2166068

[6] Cai, T. T., Low, M. and Ma, Z. (2014). Adaptive confidence bands for nonparametric regression functions. *Journal of the American Statistical Association* **109** 1054–1070. MR3265680

[7] Cai, T. T. and Low, M. G. (2005). Nonparametric estimation over shrinking neighborhoods: Superefficiency and adaptation. *The Annals of Statistics* **33** 184–213. MR2157801

[8] Carnegie, N. B. (2019). Comment: Contributions of Model Features to BART Causal Inference Performance Using ACIC 2016 Competition Data. *Statistical Science* **34** 90–93. MR3938969

[9] Castillo, I. and Ročková, V. (2021). Uncertainty quantification for Bayesian CART. *The Annals of Statistics* **49** 3482–3509. MR4352538

[10] Cheeger, J. (2015). A lower bound for the smallest eigenvalue of the Laplacian. In *Problems in analysis* 195–200. Princeton University Press. MR0402831

[11] Chipman, H. and McCulloch, R. E. (2000). Hierarchical priors for Bayesian CART shrinkage. *Statistics and Computing* **10** 17–24.

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

[13] Chipman, H. A., George, E. I. and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4** 266–298. MR2758172

[14] Crouse, M. S., Nowak, R. D. and Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on signal processing* **46** 886–902. MR1665651

[15] Denison, D. G., Mallick, B. K. and Smith, A. F. (1998). A Bayesian CART algorithm. *Biometrika* **85** 363–377. MR1649118

[16] Diaconis, P. and Stroock, D. (1991). Geometric Bounds for Eigenvalues of Markov Chains. *The Annals of Applied Probability* **1** 36–61. MR1097463

[17] Donoho, D. L. (1997). CART and best-ortho-basis: A connection. *The Annals of Statistics* **25** 1870–1911. MR1474073

[18] Donoho, D. L. and Johnstone, I. M. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika* **81** 425–455. MR1311089

[19] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90** 1200–1224. MR1379464

[20] DORIE, V., HILL, J., SHALIT, U., SCOTT, M. and CERVONE, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science* **34** 43–68. MR3938963

[21] FRIEZE, A., KANNAN, R. and POLSON, N. (1994). Sampling from Log-Concave Distributions. *The Annals of Applied Probability* **4** 812–837. MR1284987

[22] FRIGESSI, A., DI STEFANO, P., HWANG, C.-R. and SHEU, S.-J. (1993). 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** 205–219. MR1210432

[23] FRYZLEWICZ, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *Journal of the American Statistical Association* **102** 1318–1327. MR2412552

[24] GELMAN, A. and RUBIN, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7** 457–472. MR1294072

[25] GHOSH, M. (2021). Exponential tail bounds for chisquared random variables. *Journal of Statistical Theory and Practice* **15** 35. MR4228660

[26] GRAMACY, R. B. and LEE, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* **103** 1119–1130. MR2528830

[27] HE, J. and HAHN, P. R. (2023). Stochastic tree ensembles for regularized nonlinear regression. *Journal of the American Statistical Association* **118** 551–57. MR4571141

[28] HILL, J., LINERO, A. and MURRAY, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application* **7** 251–278. MR4104193

[29] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On Adaptive Posterior Concentration Rates. *The Annals of Statistics* **43** 2259–2295. MR3396985

[30] HUANG, M., LI, R. and WANG, S. (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association* **108** 929–941. MR3174674

[31] JEONG, S. and ROČKOVÁ, V. (2023). The Art of BART: Minimax Optimality over Nonhomogeneous Smoothness in High Dimension. *Journal of Machine Learning Research* **24** 1–65. MR4690286

[32] JERISON, D. (2016). *The drift and minorization method for reversible Markov chains*. Ph.D. Thesis, Department of Mathematics, Stanford University. MR4172228

[33] JERRUM, M. and SINCLAIR, A. (1988). 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* 235–244.

[34] KIM, J. and ROČKOVÁ, V. (2025). Supplement to "On mixing rates for Bayesian CART". DOI: https://doi.org/10.1214/25-EJS2397SUPP

[35] LAKSHMINARAYANAN, B., ROY, D. and TEH, Y. W. (2013). Top-down particle filtering for Bayesian decision trees. In *International Conference on Machine Learning* 280–288. PMLR.

[36] LAKSHMINARAYANAN, B., ROY, D. and TEH, Y. W. (2015). Particle Gibbs for Bayesian additive regression trees. In *Artificial Intelligence and Statistics* 553–561. PMLR.

[37] Lawler, G. F. and Sokal, A. D. (1988). 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** 557–580. MR0930082

[38] Lindvall, T. (2002). *Lectures on the coupling method*. Courier Corporation. MR1924231

[39] Linero, A. R. (2017). A review of tree-based Bayesian methods. *Communications for Statistical Applications and Methods* **24** 543–559.

[40] Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **80** 1087–1110. MR3874311

[41] Lovász, L. and Simonovits, M. (1990). The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science* 346–354. IEEE. MR1150706

[42] Lovász, L. and Simonovits, M. (1993). Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms* **4** 359–412. MR1238906

[43] Lovász, L. and Vempala, S. (2004). Hit-and-run from a corner. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing* 310–314. MR2121613

[44] Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N. and Jordan, M. I. (2019). Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences* **116** 20881–20885. MR4025861

[45] Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics* **24** 101–121. MR1389882

[46] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21** 1087–1092.

[47] Mossel, E. and Vigoda, E. (2005). Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* **309** 2207–2209.

[48] Pitman, J. (1976). On coupling of Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **35** 315–322. MR0415775

[49] Pratola, M. T. (2016). Efficient Metropolis–Hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Analysis* **11** 885–911. MR3543912

[50] Robert, C. P., Casella, G. and Casella, G. (1999). *Monte Carlo statistical methods* **2**. Springer. MR1707311

[51] Ronen, O., Saarinen, T., Tan, Y. S., Duncan, J. and Yu, B. (2022). A Mixing Time Lower Bound for a Simplified Version of BART. *arXiv preprint arXiv:2210.09352*.

[52] Ročková, V. and Rousseau, J. (2024). Ideal Bayesian Spatial Adaptation. *Journal of the American Statistical Association* **119** 2078–2091. MR4797924

[53] Ročková, V. and Saha, E. (2019). On theory for BART. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics* 2839–2848. PMLR.

[54] Shapiro, J. M. (1993). Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing* **41** 3445–3462.

[55] Sinclair, A. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, Probability and Computing* **1** 351–370. MR1211324

[56] Sleator, D. D., Tarjan, R. E. and Thurston, W. P. (1988). Rotation distance, triangulations, and hyperbolic geometry. *Journal of the American Mathematical Society* **1** 647–681. MR0928904

[57] TAN, Y. S., RONEN, O., SAARINEN, T. and YU, B. (2024). The Computational Curse of Big Data for Bayesian Additive Regression Trees: A Hitting Time Analysis. *arXiv preprint arXiv:2406.19958*.

[58] VAN DER PAS, S. and ROČKOVÁ, V. (2017). Bayesian Dyadic Trees and Histograms for Regression. In *Advances in Neural Information Processing Systems* **30**.

[59] WOODARD, D. B. and ROSENTHAL, J. S. (2013). Convergence rate of Markov chain methods for genomic motif discovery. *The Annals of Statistics* **41** 91–124. MR3059411

[60] WU, Y., TJELMELAND, H. and WEST, M. (2007). Bayesian CART: Prior specification and posterior simulation. *Journal of Computational and Graphical Statistics* **16** 44–66. MR2345747

[61] YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics* **44** 2497–2532. MR3576552

[62] ZANELLA, G. (2020). Informed proposals for local MCMC in discrete spaces. *Journal of the American Statistical Association* **115** 852–865. MR4107684

[63] ZHOU, Q. and CHANG, H. (2023). Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes. *The Annals of Statistics* **51** 1058–1085. MR4630940

[64] ZHOU, Q., YANG, J., VATS, D., ROBERTS, G. O. and ROSENTHAL, J. S. (2022). Dimension-Free Mixing for High-Dimensional Bayesian Variable Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **84** 1751–1784. MR4515557