Bayesian Additive Regression Trees For Regression Discontinuity Designs

Rafael Alcantara, Hedibert Lopes*

August 2, 2022

Abstract

The bayesian additive regression tree (BART) algorithm (Chipman et al., 2010) has recently received considerable interest because of its ability to outperform other machine learning methods in prediction tasks. One of BART's greatest qualities is its ability to deal with high-dimensional data. This characteristic can be very useful for estimating treatment effects in regression discontinuity designs (RDDs), where the inclusion of covariates might render the derivation of a fully-optimal bandwidth for the commonly employed local polynomial regression infeasible (Calonico et al., 2020). This paper investigates the application of BART to the context of regression discontinuity designs (RDDs). For this purpose, we first investigate how to obtain treatment effect estimates in an RDD with the original BART algorithm and how such estimates compare to the commonly used local linear regression approach. This analysis indicates that BART produces more reasonable results in scenarios with many 'pre-treatment' covariates. We then propose an extension of the BART algorithm where we consider a prior which incorporates the RDD assumptions and allows for direct estimation of the treatment effects.

Keywords: BART, regression discontinuity designs, high-dimensional data, causal inference

Introduction

Regression discontinuity (RD) designs, originally proposed by Thistlethwaite and Campbell (1960), have been widely used in economics and other social sciences as a substitute for randomized trials in observational settings. Such designs arise when some treatment assignment is based on whether the observed value of a given covariate – commonly known as the forcing variable – lies above or below a given cutoff value. The procedure consists of investigating if the discontinuity in the treatment probabilities, generated by the cutoff rule, leads to a discontinuity in some response variable at the cutoff. For example, if assignment to college education is determined by achieving a minimum score in a national test and interest lies in estimating the effect of college education in future earnings, one could study whether there is a discontinuity in the average earnings for people that scored slightly above and below the cutoff (and thus received college education or not, respectively). As this example suggests, such scenarios are quite common in administrative settings, in which eligibility to some program is based on clear rules and little

*rafaelca10@al.insper.edu.br, HedibertFL@insper.edu.br

discretion from administrators because of transparency constraints (Lee and Lemieux, 2010; Imbens and Lemieux, 2008).

The validity of an RDD as a substitute for a randomized treatment assignment depends crucially on two assumptions: first, individuals are unable to manipulate their realization of the forcing variable; second, the distribution of the response variable conditionally on the forcing variable must be smooth. The first assumption implies that, for observations near the cutoff, being above or below the cutoff is virtually a random assignment. For example, if the forcing variable is a test score and students can take the test only once, then students who score slightly above or below the cutoff are likely very similar except for their position relative to the cutoff. On the other hand, if students could retake the test infinitely, then a student who scored above the cutoff in their first try and a student who did so only in their tenth try are most likely not similar, but both would be eligible for treatment. The second assumption is necessary because otherwise it is not possible to determine if the discontinuity in the response variable arises from the discontinuity of the treatment probability at the cutoff or if it is a feature inherent to the data-generating process of the response (Lee and Lemieux, 2010; Imbens and Lemieux, 2008).

If the assumptions above hold, one can estimate the average treatment effect locally at the cutoff by measuring the discontinuity in the conditional expectation of the response given the forcing variable (Imbens and Lemieux, 2008). In principle, any method which estimates this conditional expectation at both sides of the cutoff well enough could be used for this task, the most common approach being the nonparametric local polynomial regression (Calonico et al., 2019). From a bayesian perspective, some examples include Karabatsos and Walker (2015), who propose approximating the conditional expectations by an infinite mixture of normals and Branson et al. (2019), who propose a Gaussian process prior for the expectations, in which observations are weighted by their distance to the cutoff[1].

Although the cutoff rule serves by assumption as a randomization device – and, therefore, we only need information about the forcing variable to estimate treatment effects –, additional covariates can be included in the estimation to increase precision as long as the treatment has no effect on them and their conditional expectation given the forcing variable is smooth (Calonico et al., 2019). However, as with any parametric model, covariate adjustment in the polynomial regression is subject to the curse of dimensionality, a problem that is aggravated by the reduction in data points when focusing on a window around the cutoff. In this case, the inclusion of covariates might even undermine the strategy of gaining precision in the estimation and produce less reliable results.

To overcome the issues described above, this paper extends the bayesian additive regression trees (BART) algorithm (Chipman et al., 2010) to the context of RD designs. Besides taking advantage of this method's predictive abilities[2], the nonparametric nature of the algorithm allows one to include many covariates in a straightforward way, which helps increasing precision in the estimation of the treatment effect.

The application of BART to causal inference contexts is part of a larger discussion about the possible gains from incorporating machine learning algorithms to estimate counterfactual outcomes[3]. Hill (2011)

[1] It is worth noting that these methods do not represent any parametric assumption about the original data generating process. Rather, they are just approximations of the true conditional expectation of the outcome given the forcing variable

[2] Dorie et al. (2019) discuss the results of a data analysis competition where BART performed better than other commonly used machine learning methods; Hill et al. (2020) present a detailed discussion of the algorithm and its applications in many contexts

[3] Perhaps the method that is most related to BART for causal inference in a frequentist context is the random causal forest of

discusses how BART can be used to obtain conditional average treatment effects for observational data. Hahn et al. (2020) extend this idea by proposing the Bayesian Causal Forest algorithm, which corrects for the fact that the regularization component of the BART prior might lead to bias in the treatment effect estimation. Hill and Su (2013) discuss how to deal with lack of common support when performing causal inference with BART. Finally, Dorie et al. (2016) discuss how to evaluate sensitivity to unmeasured confounding in treatment effect estimates obtained by BART. To the best of our knowledge, this is the first work to study the properties of BART in an RD design setting.

The remainder of this paper is organized as follows. Section 1 presents a brief overview of the RD design. Section 2 describes the BART algorithm and how it can be used for treatment effect estimation in an RDD setting. Section 3 presents a simulation study comparing BART and the local linear regression in terms of their performance in different scenarios regarding the dimension of the covariate set. Section 4 presents an extension of the BART prior that yields treatment effects directly and more closely resembles the specification of the covariate-adjusted RDD estimator of Calonico et al. (2019). Finally, section 5 concludes the paper with a discussion of the results presented here, the limitations of this study and the possible challenges in the implementation of our extension to BART.

1 Regression Discontinuity Designs

Following Imbens and Lemieux (2008), we frame the RD setting in a potential outcomes model. Let Z_i denote the treatment variable, W_i denote a M-vector of covariates and X_i denote the variable which defines treatment assignment, *i.e.* the forcing variable. That is, suppose that

$$Z_i = \begin{cases} 0, & \text{if} \quad X_i < c \\ 1, & \text{if} \quad X_i \geq c \end{cases}$$

for some cutoff value c. Let $Y_i(z_i)$ denote the potential outcome when $Z_i = z_i$. We observe only

$$Y_i = Y_i(1)Z_i + Y_i(0)(1 - Z_i). \tag{1}$$

The covariates, (X_i, W_i), are assumed to be unaffected by the treatment. Finally, the distribution of Y_i conditional on X_i is assumed to be smooth. Therefore, any discontinuity in this conditional distribution at the cutoff is assumed to come only from the treatment assignment and the effect of the treatment is measured by the size of this discontinuity.

The motivation behind the RDD is the assumption that individuals who lie just above or just below the cutoff must be very similar except for the treatment assignment. In this case, the cutoff rule acts as a randomization device for these units. Therefore, interest lies in treatment effects at the cutoff. That is, we consider some comparison between $\mathbb{E}[Y_i|Z_i = 0, X_i = c]$ and $\mathbb{E}[Y_i|Z_i = 1, X_i = c]$. For example, if one considers average treatment effects, the parameter of interest is

$$\tau_S := \mathbb{E}[Y_i|Z_i = 1, X_i = c, W_i] - \mathbb{E}[Y_i|Z_i = 0, X_i = c, W_i]. \tag{2}$$

In the context of a sharp RDD, *i.e.* when there is perfect compliance to the treatment assignment, (2)

Wager and Athey (2018)

captures the average treatment effect. However, in a fuzzy RD design, *i.e.* under imperfect compliance, this is not the case. In this scenario, three types of individuals might be observed: compliers, who take the treatment if $X_i \geq c$ and do not take the treatment if $X_i < c$, 'nevertakers', and 'alwaystakers'. In order to measure how much of the observed differences between the two conditional expectations above effectively comes from treatment assignment, one must weight (2) by the probability of being a complier:

$$\tau_F := \frac{\mathbb{E}[Y_i|Z_i = 1, X_i = c, W_i] - \mathbb{E}[Y_i|Z_i = 0, X_i = c, W_i]}{\mathbb{E}[Z_i = 1|X_i = c, W_i] - \mathbb{E}[Z_i = 0|X_i = c, W_i]}. \tag{3}$$

Note that the denominator in (3) measures the proportion of compliers in the data. In the case of a sharp RDD, the probability of treatment equals one for observations above and zero for observations below the cutoff, so the denominator above equals one and we obtain (2).

Because we cannot observe a given individual in both a treated and untreated state, the second expectation in (2) is not observable. However, under the assumption that the distribution of Y_i is smooth on X_i:

$$\mathbb{E}[Y_i|Z_i = 1, X_i = c, W_i] - \mathbb{E}[Y_i|Z_i = 0, X_i = c, W_i] = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x, W_i] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x, W_i]. \tag{4}$$

Thus, one can estimate $\mathbb{E}[Y_i|X_i, W_i]$ above and below the cutoff and extrapolate to obtain the predictions at the cutoff in order to estimate τ_S. The most common approach to this problem is to estimate a local polynomial regression of Y on X with a bandwidth choice that minimizes the mean-squared error (MSE) of the predictions (Hahn et al., 2001; Imbens and Kalyanaraman, 2012). More generally, Hahn et al. (2001) argue that any consistent method of estimating the limits above yields a consistent estimator of τ.

By assumption, the cutoff rule in the RD design acts as a randomization device so it is not technically necessary to adjust for covariates in the local regression to estimate the average treatment effects. However, Calonico et al. (2019) show how controlling for covariates can increase precision in the estimation. The authors further propose a covariate-adjusted RD estimator and discuss its asymptotic properties. The following linear-in-parameters specification is considered in that paper for the covariate-adjusted local regression:

$$Y_i = \alpha + Z_i\tau + X_i\beta_- + T_iX_i\beta_+ + W_i\gamma, \tag{5}$$

which corresponds to approximating the outcome distribution with different linear functions for observations at each side of the cutoff within the specified bandwidth.

One important limitation of parametric specifications like the one above is that the quality of the estimation decreases substantially in high-dimensional settings. This issue is aggravated by the fact that the regression is performed locally inside some bandwidth around the cutoff. The loss of sample points can increase the rate at which the quality of prediction decreases and even render estimation unfeasible for a sufficiently high number of covariates. This can act to undermine the strategy of adjusting for covariates to increase precision for high-dimensional data.

2 Bayesian Additive Regression Trees

The BART algorithm of Chipman et al. (2010) is based on a sum-of-trees model that essentially extends the ideas of the bayesian classification and regression tree (CART) model for a single tree proposed by Chipman et al. (1998). Suppose the response variable is a smooth function $f(\cdot)$ of the covariates:

$$Y_i = f(X_i, W_i, Z_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{6}$$

where ε_i is a random error term and the remaining notation is the same as in the previous section. The main idea behind BART is to approximate (6) by a sum of regression trees:

$$Y_i \approx h(X_i, W_i, Z_i) + \varepsilon_i, \tag{7}$$

where $h(\cdot) = \sum_{j=1}^{m} g_j(\cdot; T_j, \Theta_j)$, $g_j(\cdot)$ are individual regression trees, m is the number of trees in the sum, T_j is the set of interior nodes, decision rules and terminal nodes of tree j, $\Theta_j = \{\mu_{j1}, \ldots, \mu_{jb}\}$ is the set of terminal node parameters of tree j and b is the number of terminal nodes in tree j.

The parameters of the model are $(T_1, \Theta_1), \ldots, (T_j, \Theta_j)$ and σ. Chipman et al. (2010) consider priors of the form:

$$\begin{aligned} p((T_1, \Theta_1), \ldots, (T_m, \Theta_m), \sigma) &= \left[\prod_{k=1}^{m} p(T_k, \Theta_k)\right] p(\sigma^2) \\ &= \left[\prod_{k=1}^{m} p(\Theta_k | T_k) p(T_k)\right] p(\sigma^2) \end{aligned} \tag{8}$$

and

$$p(\Theta_k | T_k) = \prod_{j=1}^{b} p(\mu_{kj} | T_j). \tag{9}$$

In other words, the tree components (T_k, Θ_k) are assumed to be independent of each other and of σ^2, and the terminal node parameters $\mu_{k1}, \ldots, \mu_{kb}$ of a given tree k are assumed to be independent of each other. Furthermore, Chipman et al. (2010) consider the same priors for $p(T_k)$ and $p(\mu_{kj}|T_k)$ for all trees. The model thus consists of the specification of three priors: $p(T_k)$, $p(\sigma^2)$ and $p(\mu_{kj}|T_k)$.

The tree prior, $p(T_k)$, is defined by three components. First, the probability that a node d will split is determined by

$$\frac{\alpha}{(1+d)^{\beta}}, \quad \alpha \in (0, 1), \beta \in [0, \infty). \tag{10}$$

That is, the deeper the node (higher d), the higher the chance that it is a terminal node. This is essentially a regularization component of the tree prior to avoid overfitting.

The other components of the tree prior are the probability that a given variable will be chosen for the splitting rule at node d, and the probability that a given observed value of the chosen variable will be used for the splitting rule. The splitting variable is chosen uniformly among the set of covariates and then the

splitting value is chosen uniformly among the discrete set of observed values of that covariate.

For the error variance prior, Chipman et al. (2010) consider the following prior:

$$\sigma^2|T \sim IG(\nu/2, \nu\lambda/2), \tag{11}$$

where the hyperparameters ν and λ are calibrated in a way that the prior assigns substantial probability to the most likely range of values for σ^2.

For the prior on the terminal node parameters, $p(\mu_{kj}|T_k)$, Chipman et al. (2010) consider independent Gaussian distributions $\mu_{kj} \sim N(\mu_\mu, \sigma_\mu^2)$ for each node. Because this parameter dictates the prediction for the observations inside its corresponding terminal node, the Gaussian prior implies considering constant predictions at each node. In other words, each tree approximates the response by a piecewise constant function.

3 BART and the Local Linear Regression

In order to investigate how the inclusion of covariates impacts BART and the local linear regression when estimating treatment effects in an RD design, we performed the following simulation exercise. We generate samples of Y according to:

$$Y = \sum_{i=1}^{n} \sin(W_i)\cos(W_i) + \frac{1}{1+\exp(-5x)} + \tau Z + \varepsilon, \tag{12}$$

where $\varepsilon \sim N(0, 1)$ and n denotes the number of additional covariates included in the estimation. The covariates W are generated according to $W \sim U(0, 1)$ while the forcing variable is generated according to $X \sim U(-1, 1)$. We chose $c = 0$ as our cutoff. This specification mimics the standard practice in the RDD literature of normalizing the cutoff to zero before the analysis.

We consider variations of the following characteristics of the data-generating process:

1. Number of covariates: 10, 25 and 50
2. Size of τ: 0.5, 1 and 2
3. Sample size: 250, 500 and 700

Thus, we explore 27 variations of (12). Under this setting, our main goal is to evaluate how each method performs with a small, medium and large number of additional covariates under different sample sizes. It is reasonable to believe that increasing the dimension of the covariate set should harm precision in the estimation and that this effect should be larger for smaller samples. Additionally, we investigate whether this effect is affected by the size of the discontinuity at the cutoff. For example, it could be the case that when treatment effects are large, the decrease in precision is not so harmful for inference. For each of the 27 scenarios considered, we generated 100 samples of Y.

To evaluate the performance of each method, we compare 5 statistics: the root of the mean-squared error (RMSE), the variance of the point estimates, the length of the 95% credible/confidence interval[4],

[4] For the credible intervals of BART, we considered the interval $P(0.025 \geq \tau \geq 0.975)$ of the posterior of τ

interval coverage and the average statistical significance, measured as the proportion of times that the intervals contained zero. With this statistics, we are able to investigate how each method performs under each scenario in terms of precision in the point estimation and the quantification of uncertainty, which directly affects inference.

For the local linear regression, we applied the R package *rdrobust* of Calonico et al. (2015). For BART, we estimated the treatment effect as follows. We first estimate $E[Y|W, X, Z = 1]$ and $E[Y|W, X, Z = 0]$ with BART. That is, we estimate $E[Y|W, X, Z]$ for observations above and below the cutoff separately. Then, we use those results to predict $E[Y|W, X = 0, Z = 0]$ and $E[Y|W, X = 0, Z = 1]$ for the whole sample. That is, we predict $E[Y|W, X = 0, Z]$ for the whole sample using the curves fitted above and below the cutoff. Taking the difference between the two predictions then yields individual level treatment effects:

$$E[Y_i|W_i, X_i = 0, Z_i = 1] - E[Y_i|W_i, X_i = 0, Z_i = 0] = \tau_i. \tag{13}$$

Finally, for each draw of the MCMC, we average τ_i over the whole sample to obtain draws of the average treatment effect at $X = 0$. Tables 1, 2 and 3 present the results of our analysis for $\tau = 0.5$, $\tau = 1$ and $\tau = 2$ respectively.[5]

Table 1: Small treatment effects

		BART					Local Linear Regression				
n	k	RMSE	Variance	Interval Length	Coverage Rate	Includes Zero	RMSE	Variance	Interval Length	Coverage Rate	Includes Zero
250	10	0.36	0.05	1.17	0.91	0.22	0.74	0.54	2.18	0.86	0.76
250	25	0.45	0.04	1.05	0.71	0.04	1.29	1.68	1.91	0.62	0.53
250	50	0.56	0.06	1.04	0.53	0.06	62.22	3906.21	17.04	0.19	0.2
500	10	0.3	0.04	1.04	0.9	0.2	0.51	0.26	1.61	0.84	0.74
500	25	0.35	0.05	0.95	0.86	0.11	0.61	0.38	1.44	0.81	0.69
500	50	0.43	0.04	0.89	0.6	0.04	0.73	0.54	1.09	0.6	0.51
700	10	0.27	0.04	0.99	0.92	0.18	0.46	0.21	1.39	0.9	0.66
700	25	0.32	0.03	0.89	0.81	0.05	0.46	0.21	1.32	0.85	0.62
700	50	0.37	0.03	0.85	0.69	0.04	0.54	0.29	1.18	0.75	0.53

Table 2: Medium treatment effects

		BART					Local Linear Regression				
n	k	RMSE	Variance	Interval Length	Coverage Rate	Includes Zero	RMSE	Variance	Interval Length	Coverage Rate	Includes Zero
250	10	0.39	0.06	1.17	0.88	0	0.64	0.42	2.09	0.9	0.56
250	25	0.49	0.04	1.06	0.62	0	1.02	1.04	1.63	0.66	0.39
250	50	0.57	0.03	1.02	0.47	0	35.71	1283.8	21.52	0.24	0.24
500	10	0.32	0.04	1.05	0.88	0	0.53	0.27	1.57	0.89	0.31
500	25	0.35	0.03	0.94	0.81	0	0.56	0.31	1.48	0.85	0.29
500	50	0.44	0.02	0.89	0.62	0	0.78	0.62	1.14	0.54	0.29
700	10	0.29	0.04	0.99	0.89	0	0.4	0.16	1.41	0.89	0.22
700	25	0.34	0.02	0.88	0.85	0	0.38	0.14	1.3	0.91	0.12
700	50	0.39	0.03	0.85	0.62	0	0.52	0.27	1.16	0.73	0.22

When the dimension of the covariate set increases, we see an increase in the RMSE for both methods. For BART, the RMSE is generally less than two times larger when considering 10 versus 50 covariates. The same is true for the local linear regression except for one scenario: 250 observations with 50 covariates. In this case, the RMSE for that method can be up to 60 times as high when increasing the number of covariates. The same pattern is observed for the variance of the point estimate.

[5] n denotes the number of observations and k, the number of covariates W

Table 3: Large treatment effects

		BART					Local Linear Regression				
n	k	RMSE	Variance	Interval Length	Coverage Rate	Includes Zero	RMSE	Variance	Interval Length	Coverage Rate	Includes Zero
250	10	0.4	0.05	1.17	0.86	0	0.75	0.57	2.17	0.84	0.15
250	25	0.48	0.05	1.05	0.63	0	1.47	2.17	1.59	0.54	0.17
250	50	0.54	0.05	1.03	0.52	0	49.52	2440.3	24.25	0.2	0.22
500	10	0.32	0.03	1.04	0.91	0	0.43	0.19	1.63	0.95	0.01
500	25	0.36	0.04	0.94	0.79	0	0.58	0.35	1.47	0.84	0.02
500	50	0.4	0.03	0.92	0.73	0	0.91	0.82	1.13	0.58	0.01
700	10	0.26	0.03	0.98	0.93	0	0.33	0.11	1.4	0.96	0
700	25	0.34	0.04	0.88	0.82	0	0.46	0.22	1.35	0.88	0.01
700	50	0.37	0.03	0.84	0.72	0	0.57	0.33	1.16	0.68	0

The intervals generated by BART are tighter than those produced by the local linear regression in every scenario. For both methods, adding more covariates makes it less likely that the interval produced includes the real value of τ, although the coverage rate of the BART intervals is larger than that of the local linear regression in most scenarios. Additionally, the intervals produced by BART include zero less times than the intervals of the local linear regression in every scenario. Finally, the setting with small sample size and large number of covariates also makes uncertainty quantification much harder for the local linear regression.

The results of this section indicate that BART outperforms the local linear regression in most scenarios, although this difference is not too large when the sample size is big relative to the dimension of the covariate set. However, when the latter increases relative to the former, the performance of the local linear regression becomes poor very quickly, while the performance of BART, although mildly affected by the large number of covariates, is relatively stable. The main takeaway here is that BART is a valuable and more reliable alternative to the local linear regression in high-dimensional settings. One final interesting point worth noting from this exercise is that BART is more likely to produce statistically significant results even when treatment effects are small, while the local linear regression only performs similarly to BART in that regard when $\tau = 2$ and in medium to large sample sizes.

4 BART for Regression Discontinuity Designs

4.1 Model

We consider a model for each observation i of the form:

$$y_i = \sum_{k=1}^{m} g_k(\mathbf{x}_i, w_i, z_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{14}$$

where $\mathbf{x}_i$ is a vector of transformations of x_i (such as $\begin{bmatrix} 1 & x_i \end{bmatrix}$ or $\begin{bmatrix} 1 & x_i & x_i^2 \end{bmatrix}$), w_i is a vector of covariate values for i and z_i is the treatment assignment of i.

Inspired by equation (5), we consider an extension of the BART algorithm in which the 'pre-treatment' covariates are taken into consideration when fitting the tree, while the terminal node predictions account for the forcing and treatment variables[6]. More precisely, we propose using only the additional covariates

[6] A similar structure has been proposed by Starling et al. (2020), where the authors use the terminal node regressions to impose smoothness of the outcome over some target variable

when fitting the tree and then performing regressions in each terminal node of the form:

$$\mu_j(\mathbf{x_i}, z_i) = \mathbf{x}_i'\theta_j + \tau_j z_i, \tag{15}$$

Figure 1 presents an example of a single tree under this specification.

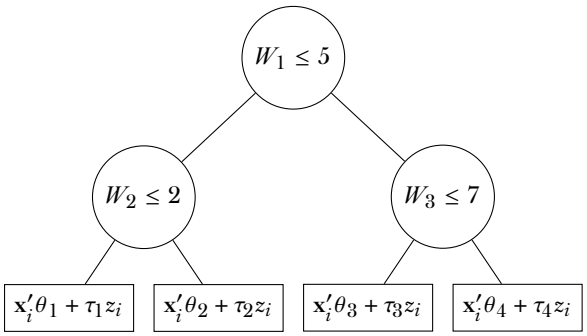

Figure 1: BART-RDD tree example

For each terminal node parameter vector $\Theta_j = \begin{bmatrix}\theta_j & \tau_j\end{bmatrix}$, we consider multivariate normal priors, which allow for great computational efficiency, as will be explored later. The priors take the form:

$$\Theta_j \sim MVN(0, \Lambda_0). \tag{16}$$

For the prior covariance matrix, we set $\Lambda_0 = \frac{1}{m}I$, which effectively shrinks the contribution of each individual tree to the overall fit to zero. A similar strategy has been considered for the same purpose by Chipman et al. (2010) and Starling et al. (2020).

For the error variance, we consider the same prior as Chipman et al. (2010):

$$\sigma^2 \sim IG\left(\frac{\nu}{2}; \frac{\nu\lambda}{2}\right), \tag{17}$$

where the hyperparameters ν and λ are chosen such that the prior assigns greater probability to more reasonable values of σ^2. Finally, we consider the same prior structure for the trees as Chipman et al. (2010).

The structure above makes it straightforward to consider many covariates in the estimation of treatment effects in RD designs without dimensionality issues for two reasons. First, the BART prior allows for regularization both through the size of individual trees and the number of trees in the additive model. Second, our approach of fitting the trees by first splitting the data on the covariates and then performing node-level regressions using only the forcing variable removes the problem of losing degrees of freedom in that regression because of the covariates.

Our model can also be written as a model for $E(Y|X, W, Z)$ of the form:

$$E(y_i|x_i, w_i, z_i) = \theta(w_i)\mathbf{x}_i + \tau(w_i)z_i. \tag{18}$$

In other words, we have an individual-level regression where the parameters for each individual are determined by the sum of the parameters assigned for that individual in each of the m trees. It is worth

noting that, when $\mathbf{x}_i = \begin{bmatrix} 1 & x_i & z_i x_i \end{bmatrix}$, our model can be seen as an individual-level analog of the local regression of Calonico et al. (2019) where the covariates are accounted for in the estimation of the parameters.

4.2 Posterior computation

We need to sample from:

$$P(\{T_1, \Theta_1\}, \ldots, \{T_m, \Theta_m\}, \sigma^2 | Y). \tag{19}$$

To sample from $\{T_1, \Theta_1\}, \ldots, \{T_m, \Theta_m\}|\sigma^2$ we proceed with the Bayesian backfitting algorithm proposed by Chipman et al. (2010). The idea is to sample $\{T_k, \Theta_k\}$ conditioned on the remaining trees and terminal node parameters $\{T_{-k}, \Theta_{-k}\}$. Note that T_k depends on the other trees only through the partial residual of those trees:

$$R_k = Y - \sum_{t \neq k} g(X, W, Z). \tag{20}$$

In other words, every tree depends on the remaining trees only through the residual of a model fitted with those trees. This means that sampling $\{T_k, \Theta_k\}$ conditional on $(\{T_{-k}, \Theta_{-k}\}, Y, \sigma^2)$ is equivalent to sampling $\{T_k, \Theta_k\}$ conditional on (R_k, σ^2) for every $k = 1, \ldots, m$. We can thus sample each $\{T_k, \Theta_k\}$ as a single tree model with R_k as our outcome.

To sample each tree, we take advantage of the fact that the priors described above are conjugate, which implies that we can integrate out Θ_k and obtain an analytical solution for $P(R_k|T_k, \sigma^2)$, which in turn allows us to obtain draws of T_k individually. More precisely:

$$\begin{aligned} P(R_k|T_k, \sigma^2) &= \int_\Theta P(R_k|\Theta_k, \sigma^2, T_k) P(\Theta_k|\sigma^2, T_k) d\Theta_k \\ &= c \prod_{j=1}^{b_k} \left(\frac{|\Lambda_{k,j,1}|}{|\Lambda_0|} \right)^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{b_k} \left(\frac{R'_{k,j} R_{k,j}}{\sigma^2} - \Theta'_{k,j,1} \Lambda^{-1}_{k,j,1} \Theta_{k,j,1} \right) \right\}, \end{aligned} \tag{21}$$

where b_k is the number of terminal nodes in T_k, c is a term that does not depend on the tree structure[7], $\bar{\Theta}_{k,j,1} = \Lambda_{k,j,1} \left(\frac{\mathrm{X}'_{k,j} R_{k,j}}{\sigma^2} \right)$ and $\Lambda_{k,j,1} = \left(\frac{\mathrm{X}'_{k,j} \mathrm{X}_{k,j}}{\sigma^2} + \Lambda_0^{-1} \right)^{-1}$.

Given $P(R_k|T_K, \sigma^2)$ and $P(T_K)$, described in the previous section, we obtain samples from the tree posterior $p(T_k|R_k, \sigma^2)$ implementing the Metropolis-Hastings algorithm proposed by Chipman et al. (2010), which is described in detail in appendix B. Then, given the posterior draw for tree T_k, we can obtain posterior draws of the parameter vector $\Theta_{k,j}$ for each terminal node j in T_k by sampling from a multivariate Normal with mean vector $\bar{\Theta}_{k,j,1}$ and covariance matrix $\Lambda_{k,j,1}$. We then proceed sequentially in the same manner for all m trees.

[7] This term is irrelevant for the Metropolis step for $P(T_k|R_k)$ since it cancels out in the Metropolis ratio because it is the same for any tree

Finally, after sampling the m pairs $\{T_m, \Theta_m\}$ with this procedure, we can obtain posterior samples for σ^2 from:

$$\sigma^2|\{T_1,\Theta_1\},\ldots,\{T_m,\Theta_m\},Y \sim IG\left(\frac{\nu+mn}{2};\frac{\nu\lambda+\sum_{k=1}^{m}\sum_{j=1}^{b_k}\left[(Y_{k,j}-X_{k,j}\Theta_{k,j})'(Y_{k,j}-X_{k,j}\Theta_{k,j})\right]}{2}\right). \tag{22}$$

Essentially, the backfitting algorithm amounts to a Gibbs sampler where we sample each T_k and Θ_k sequentially given the previous draws of T_{-k} and Θ_{-k} and finally draw σ^2 given all tree samples in the current draw.

4.2.1 Posterior inference

After d runs of the MCMC algorithm described above, we obtain d draws of the posterior distribution of $\Theta(w_i)$. Given these draws, we can recover the posterior distribution of the treatment effect – *i.e* the jump at the cutoff of the forcing variable – as follows. Note that (18) implies:

$$E(y_i|x_i=c,w_i,z_i=1)-E(y_i|x_i=c,w_i,z_i=0)=\tau(w_i). \tag{23}$$

Therefore, given one draw of the MCMC, we can average over all values of w_i to obtain one draw of the unconditional – relative to w_i – distribution of τ:

$$\hat{\tau}=\frac{\sum_{i=1}^{n}\tau(w_i)}{n}, \tag{24}$$

Similarly, to obtain draws of conditional treatment effects we take an average over some w_i of interest. Finally, once draws of the average treatment effects are obtained, we can construct credible intervals for inference as usual.[8]

The procedure above can also be used to test for discontinuities in other covariates. We simply run the same algorithm substituting x_i for some covariate of interest and obtaining draws of $\hat{\tau}$ for that new model. If the draws are sufficiently close to zero, we can be more certain of the validity of the hypotheses that any discontinuity in Y must come from X.

5 Conclusion

This paper investigated how the Bayesian Additive Regression Trees algorithm can improve the quality of estimation of treatment effects in an RD design relative to the commonly employed local linear regression in high-dimensional settings. Simulation exercises indicate that the performance gains of BART are sizable in scenarios where the covariate set is large relative to the sample size.

The main limitation of applying BART as we did in section 3 is a computational one: we need to estimate BART twice, which can be time consuming since the algorithm is already computationally intensive. With that in mind, we proposed an extension of the BART algorithm that allows for direct estimation

[8] This is essentially the same strategy of Hill (2011) to obtain ATE and CATE draws from the BART posterior

of the treatment effects, besides allowing for more control on how to model the relationship between the outcome and the forcing variable.

Even with our extension, computational intensity is still a limitation. For example, in their original application, Chipman et al. (2010) consider a model with 200 trees, which means that for d draws of the sum of trees, we need $200d$ iterations of the MCMC. It is easy to see how computation time can increase very rapidly with the sample size even if the algorithm is efficiently written. Our model has the additional difficulty of considering operations with X for each terminal node in each tree, which makes the process even more intensive. Besides efficient programming, one possible solution to this problem is to consider a smaller sample of the data to perform inference on, as proposed by Starling et al. (2020). Additional samples of the data can then be chosen for analysis as a form of robustness check.

In order to produce an efficient implementation of our proposed algorithm, we are currently working on an R package using the *Rcpp* library, which allows for the integration of R and C++ code. This is the common practice for implementations of BART and its extensions and can lead to great computational benefits[9].

Despite the computational complexity inherent to this sort of model, we believe that the performance gains relative to the local linear regression still justify our method as a valid alternative to treatment effect estimation in RD designs, especially in high-dimensional settings. Inclusion of additional covariates in RD models is a reasonable safeguard against possible violations of the independence assumptions of the RD designs, besides allowing for more precise estimation. The fact that this strategy might be undermined in the local linear regression because of this method's inability to handle many covariates makes it reasonable to believe that the computational cost of BART is more than worth it.

[9] See the packages proposed by Chipman et al. (2010); Starling et al. (2020); Hahn et al. (2020) for example

References

Branson, Z., Rischard, M., Bornn, L., and Miratrix, L. W. (2019). A nonparametric bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, 202:14–30.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, 23(2):192–210.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, 101(3):442–451.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2015). rdrobust: An r package for robust nonparametric inference in regression-discontinuity designs. *R J.*, 7(1):38.

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

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

Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470.

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., et al. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Hahn, P. R., Murray, J. S., Carvalho, C. M., et al. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*.

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.

Hill, J. and Su, Y.-S. (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, 79(3):933–959.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635.

Karabatsos, G. and Walker, S. G. (2015). A bayesian nonparametric causal model for regression discontinuity designs. In *Nonparametric Bayesian Inference in Biostatistics*, pages 403–421. Springer.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., Scott, J. G., et al. (2020). Bart with targeted smoothing: An analysis of patient-specific stillbirth risk. *Annals of Applied Statistics*, 14(1):28–50.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Appendix

A Posterior probabilities

A.1 Posterior draws for T_k and Θ_k

We derive the likelihood of each tree and the posterior for Θ for a general prior mean vector Θ_0 and covariance matrix Λ_0. Given our assumptions about independence of terminal nodes within a tree, and between different trees, we obtain:

$$P(Y|\{T_1,\Theta_1\},\ldots,\{T_m,\Theta_m\},\sigma^2)P(\{T_1,\Theta_1\},\ldots,\{T_m,\Theta_m\}|\sigma^2)P(\sigma^2) \tag{25}$$

$$=\prod_{k=1}^{m}[P(Y|\{T_k,\Theta_k\},\sigma^2)P(\{T_k,\Theta_k\}|\sigma^2)]P(\sigma^2). \tag{26}$$

Let us now focus on some tree T_k. The term inside brackets is:

$$P(Y|\{T_k,\Theta_k\},\sigma^2)P(\{T_k,\Theta_k\}|\sigma^2) \tag{27}$$

$$=(2\pi)^{-\frac{n+db_k}{2}}|\Lambda_0|^{-\frac{b_k}{2}}\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{b_k}\left[\frac{(Y_{k,j}-X_{k,j}\Theta_{k,j})'(Y_{k,j}-X_{k,j}\Theta_{k,j})}{\sigma^2}+(\Theta_{k,j}-\bar{\Theta}_0)'\Lambda_0^{-1}(\Theta_{k,j}-\bar{\Theta}_0)\right]\right\}, \tag{28}$$

where $\Theta_{k,j}$ is the parameter vector for bottom node j of T_k and d is the dimension of Θ. The term in the summation can be rewritten as:

$$\frac{(Y_{k,j}-X_{k,j}\Theta_{k,j})'(Y_{k,j}-X_{k,j}\Theta_{k,j})}{\sigma^2}+(\Theta_{k,j}-\bar{\Theta}_0)'\Lambda_0^{-1}(\Theta_{k,j}-\bar{\Theta}_0) \tag{29}$$

$$=\frac{Y_{k,j}'Y_{k,j}}{\sigma^2}-\frac{Y_{k,j}'X_{k,j}\Theta_{k,j}}{\sigma^2}-\frac{\Theta_{k,j}'X_{k,j}'Y_{k,j}}{\sigma^2}+\frac{\Theta_{k,j}'X_{k,j}'X_{k,j}\Theta_{k,j}}{\sigma^2}+\Theta_{k,j}'\Lambda_0^{-1}\Theta_{k,j}-\Theta_{k,j}'\Lambda_0^{-1}\bar{\Theta}_0-\bar{\Theta}_0'\Lambda_0^{-1}\Theta_{k,j}+\bar{\Theta}_0'\Lambda_0^{-1}\bar{\Theta}_0 \tag{30}$$

$$=\Theta_{k,j}'\left(\frac{X_{k,j}'X_{k,j}}{\sigma^2}+\Lambda_0^{-1}\right)\Theta_{k,j}-\Theta_{k,j}'\left(\frac{X_{k,j}'Y_{k,j}}{\sigma^2}+\Lambda_0^{-1}\bar{\Theta}_0\right)-\left(\frac{Y_{k,j}'X_{k,j}}{\sigma^2}\bar{\Theta}_0'\Lambda_0^{-1}\right)\Theta_{k,j}+\frac{Y_{k,j}'Y_{k,j}}{\sigma^2}+\bar{\Theta_0}'\Lambda_0^{-1}\bar{\Theta}_0. \tag{31}$$

Let $\Theta_{k,j,1}=\Lambda_{k,j,1}\left(\frac{X_{k,j}'Y_{k,j}}{\sigma^2}+\Lambda_0^{-1}\bar{\Theta}_0\right)$ and $\Lambda_{k,j,1}=\left(\frac{X_{k,j}'X_{k,j}}{\sigma^2}+\Lambda_0^{-1}\right)^{-1}$. So (31) can be further rewritten as:

$$\Theta'_{k,j}\Lambda^{-1}_{k,j,1}\Theta_{k,j} - \Theta'_{k,j}\Lambda^{-1}_{k,j,1}\Theta_{k,j,1} - \Theta'_{k,j,1}\Lambda^{-1}_{k,j,1}\Theta_{k,j} + \frac{Y'_{k,j}Y_{k,j}}{\sigma^2} + \bar{\Theta}'_0\Lambda_0^{-1}\bar{\Theta}_0 \tag{32}$$

$$=(\Theta_{k,j} - \Theta_{k,j,1})'\Lambda^{-1}_{k,j,1}(\Theta_{k,j} - \Theta_{k,j,1}) - \Theta'_{k,j,1}\Lambda^{-1}_{k,j,1}\Theta_{k,j,1} + \frac{Y'_{k,j}Y_{k,j}}{\sigma^2} + \bar{\Theta}'_0\Lambda_0^{-1}\bar{\Theta}_0. \tag{33}$$

It follows that $\Theta_{k,j}$ has a multivariate Normal posterior distribution with mean $\Theta_{k,j,1}$ and covariance $\Lambda_{k,j,1}$. Integrating out Θ_k thus yields:

$$P(Y|T_k, \sigma^2) = \int_{\Theta_k} P(Y|T_k, \Theta_k\, \sigma^2)P(\Theta_k|T_k, \sigma^2)d\Theta_k \tag{34}$$

$$= (2\pi)^{-\frac{n}{2}} \prod_{j=1}^{b_k} \left(\frac{|\Lambda_{k,j,1}|}{|\Lambda_0|}\right)^{\frac{1}{2}} \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{b_k}\left(\frac{Y'_{k,j}Y_{k,j}}{\sigma^2} + \bar{\Theta}'_0\Lambda_0^{-1}\bar{\Theta}_0 - \Theta'_{k,j,1}\Lambda^{-1}_{k,j,1}\Theta_{k,j,1}\right)\right\} \tag{35}$$

$$= c\prod_{j=1}^{b_k} \left(\frac{|\Lambda_{k,j,1}|}{|\Lambda_0|}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{b_k}\left(\frac{Y'_{k,j}Y_{k,j}}{\sigma^2} + \bar{\Theta}'_0\Lambda_0^{-1}\bar{\Theta}_0 - \Theta'_{k,j,1}\Lambda^{-1}_{k,j,1}\Theta_{k,j,1}\right)\right\}, \tag{36}$$

where c are the terms which do not depend on the tree structure. Substituting for the partial residuals for T_k and values considered for Θ_0 and Λ_0 in the paper yields the results of section 4.

A.2 Posterior for σ^2

Given the posterior draws for each tree and terminal node parameter vector, we can obtain draws from the posterior of $\sigma^2|\{T_1, \Theta_k\}, \ldots, \{T_m, \Theta_m\}, Y$. Note that:

$$P(Y|\{T_1, \Theta_1\}, \ldots, \{T_m, \Theta_m\}, \sigma^2)P(\{T_1, \Theta_1\}, \ldots, \{T_m, \Theta_m\}|\sigma^2)P(\sigma^2) \tag{37}$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{mn}{2}} \exp\left\{-\frac{1}{2}\sum_{k=1}^{m}\sum_{j=1}^{b_k}\left[\frac{(Y_{k,j} - X_{k,j}\Theta_{k,j})'(Y_{k,j} - X_{k,j}\Theta_{k,j})}{\sigma^2}\right]\right\}\left(\frac{1}{\sigma^2}\right)^{\frac{\nu}{2}-1} \exp\left\{-\frac{\frac{\nu\lambda}{2}}{\sigma^2}\right\} \tag{38}$$

$$= \left(\frac{1}{\sigma^2}\right)^{\frac{\nu+mn}{2}-1} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{\nu\lambda + \sum_{k=1}^{m}\sum_{j=1}^{b_k}\left[(Y_{k,j} - X_{k,j}\Theta_{k,j})'(Y_{k,j} - X_{k,j}\Theta_{k,j})\right]}{2}\right)\right\}. \tag{39}$$

It follows that $\sigma^2|\{T_1, \Theta_k\}, \ldots, \{T_m, \Theta_m\}, Y \sim IG\left(\frac{\nu+mn}{2}; \frac{\nu\lambda + \sum_{k=1}^{m}\sum_{j=1}^{b_k}\left[(Y_{k,j} - X_{k,j}\Theta_{k,j})'(Y_{k,j} - X_{k,j}\Theta_{k,j})\right]}{2}\right)$.

B Algorithm to sample trees

To sample new trees, Chipman et al. (1998) propose a Metropolis-Hastings algorithm which has a proposal distribution $q(T_k^i, T_k^*)$ that generates T_k^* from T_k^i by randomly selecting between the following steps:

1. **Grow:** randomly select a terminal node and split it according to the mechanism defined for the tree prior

2. **Prune:** randomly select a parent of two terminal nodes and make it a terminal node by collapsing its children

3. **Change**: randomly select an internal node and assign it a new splitting rule according to the mechanism defined for the tree prior

4. **Swap:** randomly select a pair of parent-child internal nodes and swap their splitting rules

We accept T_k^* with probability

$$\alpha(T_k^i, T_k^*) = \min\left\{\frac{q(T_k^*, T_k^i)}{q(T_k^i, T_k^*)}\frac{P(R_k|T_k^*, \sigma^2)P(T_k^*)}{P(R_k|T_k^i, \sigma^2)P(T_k^i)}, 1\right\}. \tag{40}$$

B.1 Grow step

When the **grow** step is chosen, there is substantial cancellation between $q(T_k^i, T_k^*)$ and $P(T_k^*)$ in (40). To see this, note first that:

$$q(T_k^i, T_k^*) = \frac{1}{b_k^i} \times \frac{1}{n_{\text{var}}} \times \frac{1}{n_{\text{val}}}, \tag{41}$$

where b_k^i is the number of terminal nodes in T_k^i, n_{var} is the number of variables available for splitting and n_{val} is the number of values available for splitting at the chosen variable.

Similarly, let η denote the terminal node chosen for a split from T_k^i to T_k^* and note that T_k^* is equal to T_k^i except for η. This implies:

$$P(T_k^*) = P(T_k^i) \times \frac{\frac{\alpha}{(1+d_\eta)^\beta}}{1 - \frac{\alpha}{(1+d_\eta)^\beta}} \times \frac{1}{n_{\text{var}}} \times \frac{1}{n_{\text{val}}}. \tag{42}$$

Finally, note that $q(T_k^*, T_k^i)$ is the probability of choosing node η in T_k^* and applying the **prune** step to it. That is:

$$q(T_k^*, T_k^i) = \frac{1}{nog_k^*}, \tag{43}$$

where nog_k^* is the number of no-grandchildren nodes in T_k^*, that is, nodes that are parents of terminal nodes. Therefore, when the move chosen is **grow**, we have that the ratio in (40) reduces to:

$$\frac{q(T_k^*, T_k^i)}{q(T_k^i, T_k^*)}\frac{P(R_k|T_k^*, \sigma^2)P(T_k^*)}{P(R_k|T_k^i, \sigma^2)P(T_k^i)} = \frac{b_k^i}{nog_k^*} \times \frac{\frac{\alpha}{(1+d_\eta)^\beta}}{1 - \frac{\alpha}{(1+d_\eta)^\beta}} \times \frac{P(R_k|T_k^*, \sigma^2)}{P(R_k|T_k^i, \sigma^2)}. \tag{44}$$

B.2 Prune step

Similar cancellation can be achieved if the prune step is selected. Let η now denote the parent of two terminal nodes in T_k^i that was chosen to be pruned. Note that now we have:

$$q(T_k^*, T_k^i) = \frac{1}{b_k^*} \times \frac{1}{n_{\text{var}}} \times \frac{1}{n_{\text{val}}}. \tag{45}$$

That is, $q(T_k^*, T_k^i)$ denotes the probability of growing from T_k^* back to T_k^i. Similarly, $q(T_k^i, T_k^*)$ is the probability of choosing η in T_k^i for pruning:

$$q(T_k^i, T_k^*) = \frac{1}{nog_k^i}. \tag{46}$$

Finally, note that T_k^* is the same as T_k^i except that node η splits in T_k^i but not in T_k^*. This implies:

$$P(T_k^i) = P(T_k^*) \times \frac{\frac{\alpha}{(1+d_\eta)^\beta}}{1 - \frac{\alpha}{(1+d_\eta)^\beta}} \times \frac{1}{n_{\text{var}}} \times \frac{1}{n_{\text{val}}}. \tag{47}$$

So the ratio in (40) reduces to:

$$\frac{q(T_k^*, T_k^i)}{q(T_k^i, T_k^*)} \frac{P(R_k|T_k^*, \sigma^2)P(T_k^*)}{P(R_k|T_k^i, \sigma^2)P(T_k^i)} = \frac{nog_k^i}{b_k^*} \times \frac{1 - \frac{\alpha}{(1+d_\eta)^\beta}}{\frac{\alpha}{(1+d_\eta)^\beta}} \times \frac{P(R_k|T_k^*, \sigma^2)}{P(R_k|T_k^i, \sigma^2)}. \tag{48}$$

B.3 Change and swap step

For the change and swap step, note that $q(T_k^i, T_k^*)$ and $q(T_k^*, T_k^i)$ are the same, so the ratio in (40) reduces to:

$$\frac{P(R_k|T_k^*, \sigma^2)P(T_k^*)}{P(R_k|T_k^i, \sigma^2)P(T_k^i)}. \tag{49}$$