

# BART MCMC updates

Andrew C. Parnell

March 13, 2018

## **Overall algorithm details**

The equations referred to below are displayed later in this document.

**Algorithm 1:** BART Markov chain Monte Carlo

**Data:** Target variables  $y$  (length  $n$ ; standardised), feature matrix  $X$  ( $n$  rows and  $p$  columns)

**Result:** Posterior list of trees  $T$ , values of  $\tau$ , fitted values  $\hat{y}$

**Initialisation;**

Hyper-parameter values of  $\alpha, \beta, \tau_\mu, \nu, \lambda$ ;

Number of trees  $M$ ;

Number of iterations  $N$ ;

Initial value  $\tau = 1$ ;

Set trees  $T_j$ ;  $j = 1, \dots, M$  to stumps;

Set values of  $\mu$  to 0;

**for** iterations  $i$  from 1 to  $N$  **do**

**for** trees  $j$  from 1 to  $M$  **do**

        Compute partial residuals from  $y$  minus predictions of all trees except tree  $j$ ;

        Grow a new tree  $T_j^{new}$  based on grow/prune/change/swap;

        Set  $l_{new} = \log$  full conditional of new tree  $T_j^{new}$  based on Equation 1 plus Equation 4;

        Set  $l_{old} = \log$  full conditional of old tree  $T_j$  based on same equations;

        Set  $a = \exp(l_{new} - l_{old})$ ;

        Generate  $u \sim U(0, 1)$ ;

**if**  $a > u$  **then**

            Set  $T_j = T_j^{new}$ ;

**end**

        Simulate  $\mu$  values using Equation 3;

**end**

    Get predictions  $\hat{y}$  from all trees;

    Update  $\tau$  using Equation 4;

**end**

# 1 Updating trees

Suppose there are  $n$  observations in a terminal node and suppose that the (partial) residuals in this terminal node are denoted  $R_1, \dots, R_n$ . The prior distribution for these residuals is:

$$R_1, \dots, R_n | \mu, \tau \sim N(\mu, \tau^{-1})$$

Furthermore the prior on  $\mu$  is:

$$\mu \sim N(0, \tau_\mu^{-1})$$

Using  $\pi$  to denote a probability distribution, we want to find:

$$\begin{aligned} \pi(R_1, \dots, R_n | \sigma) &= \int \pi(R_1, \dots, R_n | \mu, \tau) \pi(\mu) \partial \mu \\ &\propto \prod_{i=1}^n \tau^{1/2} e^{-\frac{\tau}{2}(R_i - \mu)^2} \tau_\mu^{1/2} e^{-\frac{\tau_\mu}{2}\mu^2} \partial \mu \\ &= \int \tau^{n/2} e^{-\frac{\tau}{2} \sum (R_i - \mu)^2} \tau_\mu^{1/2} e^{-\frac{\tau_\mu}{2}\mu^2} \partial \mu \\ &= \int \tau^{n/2} \tau_\mu^{1/2} e^{-\frac{1}{2}[\tau \{\sum R_i^2 + n\mu^2 - 2\mu n \bar{R}\} + \tau_\mu \mu^2]} \partial \mu \\ &= \tau^{n/2} \tau_\mu^{1/2} e^{-\frac{1}{2}[\tau \sum R_i^2]} \int e^{-\frac{1}{2}Q} \partial \mu \end{aligned}$$

where

$$\begin{aligned} Q &= \tau n \mu^2 - 2\tau n \mu \bar{R} + \tau_\mu \mu^2 \\ &= (\tau_\mu + n\tau) \mu^2 - 2\tau n \mu \bar{R} \\ &= (\tau_\mu + n\tau) \left[ \mu^2 - \frac{2\tau n \mu \bar{R}}{\tau_\mu + n\tau} \right] \\ &= (\tau_\mu + n\tau) \left[ \left( \mu - \frac{2\tau n \bar{R}}{\tau_\mu + n\tau} \right)^2 - \left( \frac{\tau n \bar{R}}{\tau_\mu + n\tau} \right)^2 \right] \\ &= (\tau_\mu + n\tau) \left( \mu - \frac{2\tau n \bar{R}}{\tau_\mu + n\tau} \right)^2 - \frac{(n\tau \bar{R})^2}{\tau_\mu + n\tau} \end{aligned}$$

so therefore:

$$\begin{aligned} \int e^{-\frac{1}{2}Q} \partial\mu &= \int \exp \left[ -\frac{\tau_\mu + n\tau}{2} \left( \mu - \frac{2\tau n \bar{R}}{\tau_\mu + n\tau} \right)^2 + \frac{(n\tau \bar{R})^2}{2(\tau_\mu + n\tau)} \right] \partial\mu \\ &\propto \exp \left[ \frac{1}{2} \frac{(\tau n \bar{R})^2}{\tau_\mu + n\tau} \right] (\tau_\mu + n\tau)^{-1/2} \end{aligned}$$

And finally:

$$\begin{aligned} \pi(R_1, \dots, R_n | \tau) &\propto (\tau_\mu + n\tau)^{-1/2} \tau^{n/2} \tau_\mu^{1/2} \exp \left[ \frac{1}{2} \frac{(\tau n \bar{R})^2}{\tau_\mu + n\tau} \right] \exp \left[ -\frac{\tau}{2} \sum R_i^2 \right] \\ &= \tau^{n/2} \left( \frac{\tau_\mu}{\tau_\mu + n\tau} \right)^{1/2} \exp \left[ -\frac{\tau}{2} \left\{ \sum R_i^2 - \frac{\tau(n\bar{R})^2}{\tau_\mu + n\tau} \right\} \right] \end{aligned}$$

### Including multiple terminal nodes

When we put back in terminal nodes we write  $R_{ji}$  where  $j$  is the terminal node and  $i$  is still the observation, so in terminal node  $j$  we have partial residuals  $R_{j1}, \dots, R_{jn_j}$ . When we have  $j = 1, \dots, b$  terminal nodes the full conditional distribution is then:

$$\prod_{j=1}^b \pi(R_{j1}, \dots, R_{jn_j} | \tau) \propto \prod_{j=1}^b \left\{ \tau^{n_j/2} \left( \frac{\tau_\mu}{\tau_\mu + n_j\tau} \right)^{1/2} \exp \left[ -\frac{\tau}{2} \left\{ \sum_{i=1}^{n_j} R_{ji}^2 - \frac{\tau(n_j \bar{R}_j)^2}{\tau_\mu + n_j\tau} \right\} \right] \right\}$$

which on the log scale gives:

$$\sum_{j=1}^b \left\{ \frac{n_j}{2} \log(\tau) + \frac{1}{2} \log \left( \frac{\tau_\mu}{\tau_\mu + n_j\tau} \right) - \frac{\tau}{2} \left[ \sum_{i=1}^{n_j} R_{ji}^2 - \frac{\tau(n_j \bar{R}_j)^2}{\tau_\mu + n_j\tau} \right] \right\}$$

This can be simplified further to give:

$$\frac{n}{2} \log(\tau) + \frac{1}{2} \sum_{j=1}^b \log \left( \frac{\tau_\mu}{\tau_\mu + n_j\tau} \right) - \frac{\tau}{2} \sum_{j=1}^b \sum_{i=1}^{n_j} R_{ji}^2 + \frac{\tau^2}{2} \sum_{j=1}^b \frac{S_j^2}{\tau_\mu + n_j\tau} \quad (1)$$

where  $S_j = \sum_{i=1}^{n_j} R_{ji}$

## Updating $\mu$

The full conditional for  $\mu_j$  (the terminal node parameters for node  $j$ ) is similar to the above but without the integration:

$$\begin{aligned}
\pi(\mu_j | \dots) &\propto \prod_{i=1}^{n_j} \tau^{1/2} e^{-\frac{\tau}{2}(R_{ji} - \mu_j)^2} \tau_\mu^{1/2} e^{-\frac{\tau_\mu}{2} \mu_j^2} \\
&\propto e^{-\frac{\tau}{2} \sum_{i=1}^{n_j} (R_{ji} - \mu_j)^2} e^{-\frac{\tau_\mu}{2} \mu_j^2} \\
&\propto e^{-\frac{\tau}{2} [n_j \mu_j^2 - 2\mu_j \sum_{i=1}^{n_j} R_{ji}] - \frac{\tau_\mu}{2} \mu_j^2} \\
&\propto e^{-\frac{Q}{2}}
\end{aligned}$$

Now:

$$\begin{aligned}
Q &= n_j \tau \mu_j^2 - 2\mu_j \tau S_j + \tau_\mu \mu_j^2 \\
&= (n_j \tau + \tau_\mu) \mu_j^2 - 2\tau \mu_j S_j \\
&= (n_j \tau + \tau_\mu) \left[ \mu_j^2 - \frac{2\tau \mu_j S_j}{n_j \tau + \tau_\mu} \right] \\
&\propto (n_j \tau + \tau_\mu) \left[ \mu_j - \frac{\tau \mu_j S_j}{n_j \tau + \tau_\mu} \right]^2
\end{aligned}$$

so therefore:

$$\mu_j | \dots \sim N \left( \frac{\tau S_j}{n_j \tau + \tau_\mu}, (n_j \tau + \tau_\mu)^{-1} \right) \quad (2)$$

## Update for $\tau$

I am using the shape/rate parameterisation of the gamma with prior  $\tau \sim Ga(\nu/2, \nu\lambda/2)$ . Letting  $\mu_i$  be the prediction of the  $i$ th observation we get:

$$\pi(\tau | \dots) \propto \prod_{i=1}^n \tau^{1/2} e^{-\frac{\tau}{2}(y_i - \mu_i)^2} \tau^{\nu/2} e^{-\tau \nu \lambda / 2}$$

Letting  $S = \sum_{i=1}^n (y_i - \mu_i)^2$  we get:

$$\begin{aligned}\pi(\tau | \dots) &\propto \tau^{n/2} e^{-\frac{\tau}{2} S} \tau^{\nu/2} e^{-\tau \nu \lambda / 2} \\ &= \tau^{(n+\nu)/2} e^{-\frac{\tau}{2} (S + \nu \lambda)}\end{aligned}$$

so

$$\tau | \dots \sim Ga\left(\frac{n+\nu}{2}, \frac{S+\nu\lambda}{2}\right) \quad (3)$$

## Tree prior

The tree prior used by BARTMachine says that the probability of a node being non-terminal is:

$$P(\text{node is non-terminal}) = \alpha(1 + d)^{-\beta}$$

So the probability of a node being terminal is 1 minus this. A stump just has probability 1 -  $\alpha$ . For Bart Machine  $\alpha = 0.95$  and  $\beta = 2$

Thus for a tree with  $k$  non-terminal nodes and  $b$  terminal nodes we have:

$$P = \prod_{i=1}^b [1 - \alpha(1 + d_i^t)^{-\beta}] \prod_{i=1}^k [\alpha(1 + d_i^{nt})^{-\beta}]$$

where  $d_i^t$  is the depth of the  $i$ th terminal node and  $d_i^{nt}$  is the depth of the  $i$ th non-terminal node. On the log scale this gives:

$$\log P = \sum_{i=1}^b [\log(1 - \alpha(1 + d_i^t)^{-\beta})] + \sum_{i=1}^k [\log(\alpha) - \beta \log(1 + d_i^{nt})] \quad (4)$$