Some notes on the Metropolis-Hastings Implementation

MCMC for BART

According to Gelman p.291, the Metropolis-Hastings algorithm differs from the Metropolis algorithm because you need to consider the ratio of ratios:

$$r = \frac{\frac{\mathbb{P}\left(\theta^* \,\middle|\, Y\right)}{J_t(\theta^* \mid \theta^{t-1})}}{\frac{\mathbb{P}\left(\theta^{t-1} \,\middle|\, Y\right)}{J_t(\theta^{t-1} \mid \theta^*)}} = \frac{J_t(\theta^{t-1} \mid \theta^*)}{J_t(\theta^* \mid \theta^{t-1})} \frac{\mathbb{P}\left(\theta^* \mid Y\right)}{\mathbb{P}\left(\theta^{t-1} \mid Y\right)}$$

Where we accept if a random draw from a uniform is less than the ratio above *i.e.* $X \sim \mathrm{U}\left(0,\, 1\right) \rightsquigarrow x < r$.

The parameters of interest in our case are the new tree (which is created from one of three types of proposals), which we denote T^*, and the original tree, denoted by T. For jump notation which is J, we denote this just using the regular probability symbol (to be boring). We denote the residual left over from Y, the vector of random variables, as $\boldsymbol{R} := [R_1, \ldots, R_n]^\top$. We denote σ^2 as the random variable estimating the homoskedastic noise in the model which is sampled after all the trees are sampled. We do not notate all the hyperparameters to avoid notational messiness. The M-H ratio then becomes:

$$r = \frac{\mathbb{P}\left(T^* \to T\right)}{\mathbb{P}\left(T \to T^*\right)} \frac{\mathbb{P}\left(T^* \,\middle|\, \boldsymbol{R}, \sigma^2\right)}{\mathbb{P}\left(T \,\middle|\, \boldsymbol{R}, \sigma^2\right)}$$

My goal is to come up with an exact way of calculating r for all possible tree proposals.

Throughout this document, we use the following notation:

- H — the collection of *all* nodes in tree T, not just the terminal nodes
- η — a node $\in H$.
- d_η — the depth of the ηth node. The root node is defined as having depth 0, its first child has depth 1, etc.
- b — the number of terminal nodes / leaves in the tree T. These are the nodes that can potentially be "grown." For example, the number of terminal nodes in the T^* tree is $b + 1$ if T was grown and $b - 1$ if T was pruned.

- w_2 — the number of 2nd generation internal nodes in tree T *i.e.* the number of nodes who only have two children. These are the nodes that can potentially be "pruned."
- w_2^* — the number of 2nd generation internal nodes in tree T^*. This can be equal to w or $w+1$ so it has to be recalculated. Draw a few pictures of trees and grow steps and you'll see why this unfortunate fact is so.
- ℓ — the index of the terminal node which we've "picked" to grow from it is a number $\in \{1, \ldots, b\}$.
- ℓ_L and ℓ_R — represent the new left and right node indices in tree T^* which are "grown" from the ℓth node of tree T

Likelihood Calculation

It is imperative that we can calculate $\mathbb{P}\left(T \,\middle|\, \boldsymbol{R}, \sigma^2\right)$ for the calculation of r. Let's analyze carefully what it means to get the likelihood of a tree.

First, note that the likelihood for T given the data is not defined in our model, so we use Bayes Rule to obtain something that is tractable in our model:

$$\mathbb{P}\left(T \,\middle|\, \boldsymbol{R}, \sigma^2\right) = \frac{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, T, \sigma^2\right) \mathbb{P}\left(T \,\middle|\, \sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, \sigma^2\right)}$$

What's in the numerator? The likelihood of the data given the tree and the variance times the probability of the tree given the variance. The probability of the tree is based on probabilities of splits and rules and is not dependent on the variance, $\mathbb{P}\left(T \,\middle|\, \sigma^2\right) = \mathbb{P}\left(T\right)$, so we can already simplify to:

$$\mathbb{P}\left(T \,\middle|\, \boldsymbol{R}, \sigma^2\right) = \frac{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, T, \sigma^2\right) \mathbb{P}\left(T\right)}{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, \sigma^2\right)}$$

What about the likelihood given the tree? The only reason you need the T information is so we know which R values fall in which of the leaves:

$$\mathbb{P}\left(R_1, \ldots, R_n \,\middle|\, T, \sigma^2\right) = \prod_{\ell=1}^{b} \mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \,\middle|\, \sigma^2\right)$$

The r.h.s is the likelihoods of each leaf separately which is *not* dependent on T anymore. We can multiply the likelihoods of each leaf because we assume the leaves are independent. The R_ℓ's are the data in the ℓth leaf and there is n_ℓ of them, the portion of n in the leaf where $n = \sum_{\ell=1}^{b} n_\ell$.

Let's look at the likelihood of a single leaf more carefully. We know that if we knew the mean at the leaf, which we denote μ_ℓ, we would have:

$$R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} | \mu_\ell, \sigma^2 \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$$

This means that if we can margin out μ, we're in good shape. Recall that we put a prior on the average value of $\mu_\ell \sim \mathcal{N}\left(0,\, \sigma_\mu^2\right)$ and thus:

$$\mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \,\middle|\, \sigma^2\right) = \int_{\mathbb{R}} \mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \,\middle|\, \mu_\ell, \sigma^2\right) \mathbb{P}\left(\mu_\ell; \sigma_\mu^2\right) d\mu_\ell$$

Let's take a more careful look at the non-margined leaf-likelihood expression and recall from basic mathematical statistics the result that the sample average, $\bar{R}_\ell$, is a sufficient statistic for μ_ℓ via the factorization theorem:

$$\begin{aligned}
\mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} | \mu_\ell, \sigma^2\right) &= \frac{1}{\left(2\pi\sigma^2\right)^{n_\ell/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n_\ell} \left(R_{\ell_i} - \mu_\ell\right)^2\right) \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{n_\ell/2}} \exp\left(-\frac{1}{2\sigma^2}\left(n_\ell \left(\bar{R}_\ell - \mu_\ell\right)^2 + \sum_{i=1}^{n_\ell} \left(R_{\ell_i} - \bar{R}_\ell\right)^2\right)\right) \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{n_\ell/2}} \exp\left(-\frac{1}{2\frac{\sigma^2}{n_\ell}}\left(\left(\bar{R}_\ell - \mu_\ell\right)^2 + \frac{1}{n_\ell}\sum_{i=1}^{n_\ell} \left(R_{\ell_i} - \bar{R}_\ell\right)^2\right)\right) \\
&= \underbrace{\frac{1}{\left(2\pi\sigma^2\right)^{n_\ell/2}} \exp\left(\frac{1}{n_\ell}\sum_{i=1}^{n_\ell} \left(R_{\ell_i} - \bar{R}_\ell\right)^2\right)}_{h} \underbrace{\exp\left(-\frac{1}{2\frac{\sigma^2}{n_\ell}}\left(\left(\bar{R}_\ell - \mu_\ell\right)^2\right)\right)}_{g}
\end{aligned}$$

So now we margin and notice that the integral is the definition of a convolution:

$$\begin{aligned}
\mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \,\middle|\, \sigma^2\right) &= \int_{\mathbb{R}} \mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} | \mu_\ell, \sigma^2\right) \mathbb{P}\left(\mu_\ell; \sigma_\mu^2\right) d\mu_\ell \\
&= h\sqrt{2\pi\frac{\sigma^2}{n_\ell}} \int_{\mathbb{R}} \sqrt{\frac{1}{2\pi\frac{\sigma^2}{n_\ell}}} \exp\left(-\frac{1}{2\frac{\sigma^2}{n_\ell}}\left(\left(\bar{R}_\ell - \mu_\ell\right)^2\right)\right) \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right) d\mu_\ell \\
&= h\sqrt{2\pi\frac{\sigma^2}{n_\ell}} \mathcal{N}\left(0,\, \frac{\sigma^2}{n_\ell}\right) \star \mathcal{N}\left(0,\, \sigma_\mu^2\right) \\
&= h\sqrt{2\pi\frac{\sigma^2}{n_\ell}} \mathcal{N}\left(0,\, \frac{\sigma^2}{n_\ell} + \sigma_\mu^2\right) \\
&= h\sqrt{2\pi\frac{\sigma^2}{n_\ell}} \frac{1}{\sqrt{2\pi\left(\frac{\sigma^2}{n_\ell} + \sigma_\mu^2\right)}} \exp\left(-\frac{1}{2\left(\frac{\sigma^2}{n_\ell} + \sigma_\mu^2\right)} \bar{R}_\ell^2\right) \\
&= h\sqrt{\frac{\sigma^2}{\sigma^2 n_\ell + \sigma_\mu^2}} \exp\left(-\frac{1}{2\left(\frac{\sigma^2}{n_\ell} + \sigma_\mu^2\right)} \bar{R}_\ell^2\right)
\end{aligned}$$

Let's finish this section off by returning to the probability of the tree which was:

$$\mathbb{P}\left(T \,\middle|\, \boldsymbol{R}, \sigma^2\right) = \frac{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, T, \sigma^2\right) \mathbb{P}\left(T\right)}{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, \sigma^2\right)}$$

What about the denominator — the probability of the data? The probability of the data is weighted over every possible tree configuration:

$$\mathbb{P}\left(\boldsymbol{R} \,\middle|\, \sigma^2\right) = \int_{T \in \mathcal{T}} \mathbb{P}\left(\boldsymbol{R} \,\middle|\, T, \sigma^2\right) \mathbb{P}\left(T\right) dT$$

and removing the dependency on T becomes:

$$\mathbb{P}\left(\boldsymbol{R} \,\middle|\, \sigma^2\right) = \int_{T \in \mathcal{T}} \prod_{\ell=1}^{b_T} \mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \,\middle|\, \sigma^2\right) \mathbb{P}\left(T\right) dT$$

which of course is arrived at via the margining out of the means of each leaf:

$$\mathbb{P}\left(\boldsymbol{R} \,\middle|\, \sigma^2\right) = \int_{T \in \mathcal{T}} \left(\prod_{\ell=1}^{b_T} \int_{\mathbb{R}} \mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \,\middle|\, \mu, \sigma^2\right) \mathbb{P}\left(\mu; \sigma_\mu^2\right) d\mu \right) \mathbb{P}\left(T\right) dT$$

The point being is that this quantity is the same for all data $\boldsymbol{R}$. This is useful since we're creating ratios where we're using the same data, and this quantity will cancel.

Now let's look at the ratio of the likelihoods which is what we care about for the calculation of r. I identify three pieces which we will use for the next couple of sections:

$$\begin{aligned} r &= \frac{\mathbb{P}\left(T^* \to T\right)}{\mathbb{P}\left(T \to T^*\right)} \underbrace{\frac{\mathbb{P}\left(T^* \,\middle|\, \boldsymbol{R}, \sigma^2\right)}{\mathbb{P}\left(T \,\middle|\, \boldsymbol{R}, \sigma^2\right)}} \\ &\qquad \frac{\dfrac{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, T^*, \sigma^2\right) \mathbb{P}\left(T^*\right)}{\cancel{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, \sigma^2\right)}}}{\dfrac{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, T, \sigma^2\right) \mathbb{P}\left(T\right)}{\cancel{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, \sigma^2\right)}}} \\ &\qquad \text{proportional likelihood ratio} \\ &= \underbrace{\frac{\mathbb{P}\left(T^* \to T\right)}{\mathbb{P}\left(T \to T^*\right)}}_{\text{transition ratio}} \times \overbrace{\frac{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, T^*, \sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R} \,\middle|\, T, \sigma^2\right)}} \times \underbrace{\frac{\mathbb{P}\left(T^*\right)}{\mathbb{P}\left(T\right)}}_{\text{tree structure ratio}} \end{aligned}$$

Grow Proposal

Let's pretend we're transitioning from $T \to T^*$ using a GROW step and let's analyze each of the above expressions one-by-one.

Transition Ratio

So $\mathbb{P}(T \to T^*)$ means the probability of transitioning from T into the new tree proposal T^*. This would have to be equal to the following:

$$\begin{aligned}\mathbb{P}(T \to T^*) &= \mathbb{P}(\text{GROW})\,\mathbb{P}(\text{selecting the } \ell\text{th node to grow from}) \times \\ &\quad \mathbb{P}(\text{selecting the } j\text{th attribute to split on})\,\mathbb{P}(\text{selecting the } i\text{th value to split on})\end{aligned}$$

We're picking from on of the terminal nodes, and then we're picking an attribute and split point, this becomes:

$$\mathbb{P}\left(T^* \,\middle|\, T\right) = \mathbb{P}(\text{GROW})\,\frac{1}{b}\frac{1}{p_{adj}}\frac{1}{n_{adj}}$$

Now, p_{adj} is the number of predictors left available to split on. This is *from the perspective* of the ℓth node in tree T. Why would this be less than p? Because if you look up into the node's lineage, you may have already used all available split values for some attributes. Those would no longer be available to split from.

Then, n_{adj} is the number of split values left available considering we picked attribute j. We can obtain this by looking at the node's lineage for any splits on j and then taking the minimum of those split values, and then finding the subset of the design matrix whose values are less than that minimum for column j.

So now $\mathbb{P}(T^* \to T)$ is the probability of transitioning from the new tree back to the old tree which would be:

$$\begin{aligned}\mathbb{P}(T^* \to T) &= \mathbb{P}(\text{PRUNE})\,\mathbb{P}(\text{selecting the } \ell\text{th node to prune from}) \\ &= \mathbb{P}(\text{PRUNE})\,\frac{1}{w_2^*}\end{aligned}$$

Thus, the transition ratio will be:

$$\frac{\mathbb{P}(T^* \to T)}{\mathbb{P}(T \to T^*)} = \frac{\mathbb{P}(\text{PRUNE})\,\frac{1}{w_2^*}}{\mathbb{P}(\text{GROW})\,\frac{1}{b}\frac{1}{p_{adj}}\frac{1}{n_{adj}}} = \frac{\mathbb{P}(\text{PRUNE})}{\mathbb{P}(\text{GROW})}\frac{b\;p_{adj}\;n_{adj}}{w_2^*}$$

Why don't the probabilities of prune and grow cancel? Well, under the case where we cannot grow anymore (if we use all split variables), the probability of growth will be 0. Thus, those steps *cannot* be considered since the ratio would be undefined. As long as they can be considered, that ratio will cancel.

Proportional Likelihood Ratio

$\mathbb{P}\left(\boldsymbol{R} \,\middle|\, T^*, \sigma^2\right)$ represents the likelihood of all the responses (adjusted by the other trees) to wind up in the nodes they've wound up in. As we've shown in the previous section that the likelihood for any node is:

$$\mathbb{P}\left(R_{\ell_1},\ldots,R_{\ell_{n_\ell}}\,\middle|\,\sigma^2\right) = h\sqrt{\frac{\sigma^2}{\sigma^2 n_\ell+\sigma_\mu^2}}\,\exp\left(-\frac{1}{2\left(\frac{\sigma^2}{n_\ell}+\sigma_\mu^2\right)}\bar{R}_\ell^2\right)$$

Since the likelihoods are solely determined by the terminal nodes, the proposal tree differs from the original tree by only the ℓth node in the original becoming the ℓ_L and ℓ_R nodes in the proposal. Hence, the proportional likelihood ratio becomes only:

$$\begin{aligned}\frac{\mathbb{P}\left(\boldsymbol{R}\,\middle|\,T^*,\sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R}\,\middle|\,T,\sigma^2\right)} &= h_L\sqrt{\frac{\sigma^2}{\sigma^2 n_{\ell_L}+\sigma_\mu^2}}\,\exp\left(-\frac{1}{2\left(\frac{\sigma^2}{n_{\ell_L}}+\sigma_\mu^2\right)}\bar{R}_{\ell_L}^2\right)\\ &\quad\times h_R\sqrt{\frac{\sigma^2}{\sigma^2 n_{\ell_R}+\sigma_\mu^2}}\,\exp\left(-\frac{1}{2\left(\frac{\sigma^2}{n_{\ell_R}}+\sigma_\mu^2\right)}\bar{R}_{\ell_R}^2\right)\times\\ &\quad\left(h\sqrt{\frac{\sigma^2}{\sigma^2 n_\ell+\sigma_\mu^2}}\,\exp\left(-\frac{1}{2\left(\frac{\sigma^2}{n_\ell}+\sigma_\mu^2\right)}\bar{R}_\ell^2\right)\right)^{-1}\end{aligned}$$

Note that the ratio of the h functions do *not* cancel:

$$\begin{aligned}\frac{h_L h_R}{h} &= \frac{h\left(R_{\ell_{L,1}},\ldots,R_{\ell_{L,n_{\ell,L}}}\mid\sigma^2\right)h\left(R_{\ell_{R,1}},\ldots,R_{R,\ell_{n_{\ell,R}}}\mid\sigma^2\right)}{h\left(R_{\ell_1},\ldots,R_{\ell_{n_\ell}}\mid\sigma^2\right)}\\ &= \frac{\frac{1}{(2\pi\sigma^2)^{n_{\ell_L}/2}}\exp\left(\frac{1}{n_{\ell_L}}\sum_{i=1}^{n_{\ell_L}}\left(R_{\ell_{L,i}}-\bar{R}_{\ell_L}\right)^2\right)\frac{1}{(2\pi\sigma^2)^{n_{\ell_R}/2}}\exp\left(\frac{1}{n_{\ell_R}}\sum_{i=1}^{n_{\ell_R}}\left(R_{\ell_{R,i}}-\bar{R}_{\ell_R}\right)^2\right)}{\frac{1}{(2\pi\sigma^2)^{n_\ell/2}}\exp\left(\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(R_{\ell_i}-\bar{R}_\ell\right)^2\right)}\\ &= \frac{\exp\left(\frac{1}{n_{\ell_L}}\sum_{i=1}^{n_{\ell_L}}\left(R_{\ell_{L,i}}-\bar{R}_{\ell_L}\right)^2+\frac{1}{n_{\ell_R}}\sum_{i=1}^{n_{\ell_R}}\left(R_{\ell_{R,i}}-\bar{R}_{\ell_R}\right)^2\right)}{\exp\left(\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\left(R_{\ell_i}-\bar{R}_\ell\right)^2\right)}\quad(\text{since } n_\ell = n_{\ell_R}+n_{\ell_L})\end{aligned}$$

Hence, we cannot really simplify the above expression for the proportional likelihood ratio.

Alternative Calculation

Let's factorize the $\overset{iid}{\sim}$ normal likelihood like the following:

$$\begin{aligned}\mathbb{P}\left(R_{\ell_1},\ldots,R_{\ell_{n_\ell}}|\mu,\sigma^2\right) &= \frac{1}{(2\pi\sigma^2)^{n_\ell/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n_\ell}\left(R_{\ell_i}-\mu\right)^2\right)\\ &= \frac{1}{(2\pi\sigma^2)^{n_\ell/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n_\ell}R_{\ell_i}^2\right)\underbrace{\exp\left(-\frac{1}{2\sigma^2}\left(-2n_\ell\mu\bar{R}_\ell+n_\ell\mu^2\right)\right)}_{g}\end{aligned}$$

Now, let's try to build the ratio:

$$
\begin{aligned}
&\frac{\mathbb{P}\left(\boldsymbol{R}\,\middle|\,T^*,\sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R}\,\middle|\,T,\sigma^2\right)} \\
=\;& \frac{\mathbb{P}\left(R_{\ell_{L,1}},\ldots,R_{\ell_{L,n_{\ell,L}}}\,\middle|\,\sigma^2\right)\mathbb{P}\left(R_{\ell_{R,1}},\ldots,R_{R,\ell_{n_{\ell,R}}}\,\middle|\,\sigma^2\right)}{\mathbb{P}\left(R_{\ell_1},\ldots,R_{\ell_{n_\ell}}\,\middle|\,\sigma^2\right)} \\
=\;& \frac{\int_{\mathbb{R}}\mathbb{P}\left(R_{\ell_{L,1}},\ldots,R_{\ell_{L,n_{\ell,L}}}\,\middle|\,\mu,\sigma^2\right)\mathbb{P}(\mu)\,\mathrm{d}\mu\quad\int_{\mathbb{R}}\mathbb{P}\left(R_{\ell_{R,1}},\ldots,R_{R,\ell_{n_{\ell,R}}}\,\middle|\,\mu,\sigma^2\right)\mathbb{P}(\mu)\,\mathrm{d}\mu}{\int_{\mathbb{R}}\mathbb{P}\left(R_{\ell_1},\ldots,R_{\ell_{n_\ell}}\,\middle|\,\mu,\sigma^2\right)\mathbb{P}(\mu)\,\mathrm{d}\mu} \\
=\;& \frac{\frac{1}{(2\pi\sigma^2)^{n_{\ell_L}/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n_{\ell_L}}R_{\ell_{L}i}^2\right)\left(\int_{\mathbb{R}}g_{\ell_L}\mathbb{P}(\mu)\,\mathrm{d}\mu\right)\;\frac{1}{(2\pi\sigma^2)^{n_{\ell_R}/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n_{\ell_R}}R_{\ell_{R}i}^2\right)\left(\int_{\mathbb{R}}g_{\ell_R}\mathbb{P}(\mu)\,\mathrm{d}\mu\right)}{\frac{1}{(2\pi\sigma^2)^{n_\ell/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n_\ell}R_{\ell_i}^2\right)\left(\int_{\mathbb{R}}g_{\ell}\mathbb{P}(\mu)\,\mathrm{d}\mu\right)} \\
=\;& \frac{\cancel{\frac{1}{(2\pi\sigma^2)^{n_{\ell_L}/2}}}\cancel{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n_{\ell_L}}R_{\ell_{L}i}^2\right)}\left(\int_{\mathbb{R}}g_{\ell_L}\mathbb{P}(\mu)\,\mathrm{d}\mu\right)\;\cancel{\frac{1}{(2\pi\sigma^2)^{n_{\ell_R}/2}}}\cancel{\exp\left(-\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n_{\ell_R}}R_{\ell_{R}i}^2}\left(\int_{\mathbb{R}}g_{\ell_R}\mathbb{P}(\mu)\,\mathrm{d}\mu\right)}{\cancel{\frac{1}{(2\pi\sigma^2)^{n_\ell/2}}}\cancel{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n_\ell}R_{\ell_i}^2\right)}\left(\int_{\mathbb{R}}g_{\ell}\mathbb{P}(\mu)\,\mathrm{d}\mu\right)} \\
=\;& \frac{\int_{\mathbb{R}}g_{\ell_L}\mathbb{P}(\mu)\,\mathrm{d}\mu\int_{\mathbb{R}}g_{\ell_R}\mathbb{P}(\mu)\,\mathrm{d}\mu}{\int_{\mathbb{R}}g_{\ell}\mathbb{P}(\mu)\,\mathrm{d}\mu} \\
=\;& \frac{\int_{\mathbb{R}}g_{\ell_L}\cancel{\frac{1}{\sqrt{2\pi\sigma_\mu^2}}}\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)\mathrm{d}\mu\int_{\mathbb{R}}g_{\ell_R}\frac{1}{\sqrt{2\pi\sigma_\mu^2}}\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)\mathrm{d}\mu}{\int_{\mathbb{R}}g_{\ell}\cancel{\frac{1}{\sqrt{2\pi\sigma_\mu^2}}}\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)\mathrm{d}\mu} \\
=\;& \frac{1}{\sqrt{2\pi\sigma_\mu^2}}\frac{\int_{\mathbb{R}}g_{\ell_L}\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)\mathrm{d}\mu\int_{\mathbb{R}}g_{\ell_R}\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)\mathrm{d}\mu}{\int_{\mathbb{R}}g_{\ell}\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)\mathrm{d}\mu} \\
=\;& \frac{1}{\sqrt{2\pi\sigma_\mu^2}}\frac{\int_{\mathbb{R}}\exp\left(-\frac{1}{2\sigma^2}\left(-2n_{\ell_L}\mu\bar{R}_{\ell_L}+n_{\ell_L}\mu^2\right)\right)\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)\mathrm{d}\mu\int_{\mathbb{R}}\exp\left(-\frac{1}{2\sigma^2}\left(-2n_{\ell_L}\mu\bar{R}_{\ell_L}+n_{\ell_L}\mu^2\right)\right)\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)\mathrm{d}\mu}{\underbrace{\int_{\mathbb{R}}\exp\left(-\frac{1}{2\sigma^2}\left(-2n_{\ell}\mu\bar{R}_{\ell}+n_{\ell}\mu^2\right)\right)\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)\mathrm{d}\mu}}
\end{aligned}
$$

Let's evaluate the underbraced integral using Mathematica:

$$
\begin{aligned}
\int_{\mathbb{R}}g(\bar{R}_\ell\mid\mu,\sigma^2)\mathbb{P}\left(\mu;\sigma_\mu^2\right)d\mu &= \int_{\mathbb{R}}\exp\left(-\frac{1}{2\sigma^2}\left(-2n_\ell\mu\bar{R}_\ell+n_\ell\mu^2\right)\right)\frac{1}{\sqrt{2\pi\sigma_\mu^2}}\exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right)d\mu \\
&= \sqrt{\frac{2\pi}{\frac{n_\ell}{\sigma^2}+\frac{1}{\sigma_\mu^2}}}\exp\left(\frac{n_\ell^2\bar{R}_\ell^2\sigma_\mu^2}{2\sigma^2\left(\sigma^2+n_\ell\sigma_\mu^2\right)}\right)
\end{aligned}
$$

So now we have:

$$
\begin{aligned}
&\frac{\mathbb{P}\left(\boldsymbol{R}\,\middle|\,T^*,\sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R}\,\middle|\,T,\sigma^2\right)} \\
&= \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \frac{\sqrt{\frac{2\pi}{\frac{n_{\ell_L}}{\sigma^2}+\frac{1}{\sigma_\mu^2}}}\exp\left(\frac{n_{\ell_L}^2\bar{R}_{\ell_L}^2\sigma_\mu^2}{2\sigma^2\left(\sigma^2+n_{\ell_L}\sigma_\mu^2\right)}\right)\sqrt{\frac{2\pi}{\frac{n_{\ell_R}}{\sigma^2}+\frac{1}{\sigma_\mu^2}}}\exp\left(\frac{n_{\ell_R}^2\bar{R}_{\ell_R}^2\sigma_\mu^2}{2\sigma^2\left(\sigma^2+n_{\ell_R}\sigma_\mu^2\right)}\right)}{\sqrt{\frac{2\pi}{\frac{n_\ell}{\sigma^2}+\frac{1}{\sigma_\mu^2}}}\exp\left(\frac{n_\ell^2\bar{R}_\ell^2\sigma_\mu^2}{2\sigma^2\left(\sigma^2+n_\ell\sigma_\mu^2\right)}\right)} \\
&= \sqrt{\frac{\frac{n_\ell}{\sigma^2}+\frac{1}{\sigma_\mu^2}}{\sigma_\mu^2\left(\frac{n_{\ell_L}}{\sigma^2}+\frac{1}{\sigma_\mu^2}\right)\left(\frac{n_{\ell_R}}{\sigma^2}+\frac{1}{\sigma_\mu^2}\right)}}\ \exp\left(\frac{\sigma_\mu^2}{2\sigma^2}\left(\frac{n_{\ell_L}^2\bar{R}_{\ell_L}^2}{\sigma^2+n_{\ell_L}\sigma_\mu^2}+\frac{n_{\ell_R}^2\bar{R}_{\ell_R}^2}{\sigma^2+n_{\ell_R}\sigma_\mu^2}-\frac{n_\ell^2\bar{R}_\ell^2}{\sigma^2+n_\ell\sigma_\mu^2}\right)\right) \\
&= \sqrt{\frac{\sigma^2\left(\sigma^2+n_\ell\sigma_\mu^2\right)}{\left(\sigma^2+n_{\ell_L}\sigma_\mu^2\right)\left(\sigma^2+n_{\ell_R}\sigma_\mu^2\right)}}\ \exp\left(\frac{\sigma_\mu^2}{2\sigma^2}\left(\frac{\left(\sum_{i=1}^{n_{\ell_L}}R_{\ell_L,i}\right)^2}{\sigma^2+n_{\ell_L}\sigma_\mu^2}+\frac{\left(\sum_{i=1}^{n_{\ell_L}}R_{\ell_R,i}\right)^2}{\sigma^2+n_{\ell_R}\sigma_\mu^2}-\frac{\left(\sum_{i=1}^{n_{\ell_L}}R_{\ell,i}\right)^2}{\sigma^2+n_\ell\sigma_\mu^2}\right)\right)
\end{aligned}
$$

Note how the ratio is only a function of the $\sum R_i$'s which makes calculations efficient.

Tree Structure Ratio

We consider each tree separately.

$$
\mathbb{P}\left(T^*\right) = \prod_{\eta\in H}\mathbb{P}\left(\text{splitting }\eta\right)\prod_{\eta\in H}\mathbb{P}\left(\text{assigning a rule to }\eta\right)
$$

Remember from CGM98 that the probability of splitting on a given node η is driven by two hyperparameters, α and β in the following way where d_η is the depth of the η node:

$$
\mathbb{P}\left(\text{splitting }\eta\right) = \frac{\alpha}{\left(1+d_\eta\right)^\beta}
$$

The probability of assigning the specific rule is just picking from all available attributes and then from all available split points so $\frac{1}{p_\eta}\frac{1}{n_\eta}$. Once again, the proposal tree differs from the original tree by only the ℓth node in the original becoming the ℓ_L and ℓ_R nodes in the proposal. This simplifies the tree structure ratio to just:

$$
\frac{\mathbb{P}\left(T^*\right)}{\mathbb{P}\left(T\right)} = \frac{\frac{\alpha}{\left(1+d_{\eta_L}\right)^\beta}\frac{\alpha}{\left(1+d_{\eta_R}\right)^\beta}}{\frac{\alpha}{\left(1+d_\eta\right)^\beta}}\frac{p_\eta n_\eta}{p_{\eta_L}p_{\eta_R}n_{\eta_L}n_{\eta_R}}
$$

We know the depth of the child nodes is just the depth of the parent node incremented by 1. That plus a bit more algebra gives us:

$$\frac{\mathbb{P}(T^*)}{\mathbb{P}(T)} = \alpha \frac{(1+d_\eta)^\beta}{(2+d_\eta)^{2\beta}} \frac{p_\eta n_\eta}{p_{\eta_L} p_{\eta_R} n_{\eta_L} n_{\eta_R}}$$

Now we have a way of calculating r for grow proposals by multiplying all three above results.

Prune Proposal

Transition Ratio

For prune proposals, we move in the opposite direction. We need to hack off a node:

$$\begin{aligned}
\mathbb{P}(T \to T^*) &= \mathbb{P}(\text{PRUNE})\, \mathbb{P}(\text{selecting the } \ell\text{th node to prune from}) \\
&= \mathbb{P}(\text{PRUNE}) \frac{1}{w_2}
\end{aligned}$$

To go the opposite direction, we need to make sure we grow the exact same node so p_{adj}, n_{adj} need to be calculated based on whatever the ℓth node originally was:

$$\begin{aligned}
\mathbb{P}(T^* \to T) &= \mathbb{P}(\text{GROW})\, \mathbb{P}(\text{selecting the } \ell\text{th node to grow from}) \times \\
&\quad \mathbb{P}(\text{selecting the original attribute to split on}) \times \\
&\quad \mathbb{P}(\text{selecting the original value to split on}) \\
&= \mathbb{P}(\text{GROW}) \frac{1}{b-1} \frac{1}{p_{adj}} \frac{1}{n_{adj}}
\end{aligned}$$

We're using $b-1$ here because the proposed tree has one less terminal node due to the pruning than the original tree T had.

Thus, the transition ratio becomes:

$$\frac{\mathbb{P}(T^* \to T)}{\mathbb{P}(T \to T^*)} = \frac{\mathbb{P}(\text{GROW}) \frac{1}{b-1} \frac{1}{p_{adj}} \frac{1}{n_{adj}}}{\mathbb{P}(\text{PRUNE}) \frac{1}{w_2}} = \frac{\mathbb{P}(\text{GROW})}{\mathbb{P}(\text{PRUNE})} \frac{w_2}{(b-1) p_{adj} n_{adj}}$$

Once again, we need to bookkeep and make sure we can actually prune this node. If it's just a root node, then we can't even consider this step at all. Otherwise, they'll cancel.

Proportional Likelihood Ratio

It is pretty obvious this is the inverse of the grow step's proportional likelihood ratio. Now the tree proposal has just one collapsed node where the original has a left and right component:

$$\begin{aligned}
\frac{\mathbb{P}\left(\boldsymbol{R}\,\middle|\,T^*,\sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R}\,\middle|\,T,\sigma^2\right)} &= \left(\sqrt{\frac{\sigma^2\left(n_\ell+\sigma^2\right)}{\sigma_\mu^2\left(n_{\ell_L}+\sigma^2\right)\left(n_{\ell_R}+\sigma^2\right)}}\ \exp\left(\frac{\sigma_\mu^2}{2\sigma^2}\left(\frac{n_{\ell_L}^2\bar{R}_L^2}{\sigma^2+n_{\ell_L}\sigma_\mu^2}+\frac{n_{\ell_R}^2\bar{R}_R^2}{\sigma^2+n_{\ell_R}\sigma_\mu^2}-\frac{n_\ell^2\bar{R}^2}{\sigma^2+n_\ell\sigma_\mu^2}\right)\right)\right)^{-1} \\
&= \sqrt{\frac{\sigma_\mu^2\left(n_{\ell_L}+\sigma^2\right)\left(n_{\ell_R}+\sigma^2\right)}{\sigma^2\left(n_\ell+\sigma^2\right)}}\ \exp\left(\frac{\sigma_\mu^2}{2\sigma^2}\left(\frac{n_\ell^2\bar{R}^2}{\sigma^2+n_\ell\sigma_\mu^2}-\frac{n_{\ell_L}^2\bar{R}_L^2}{\sigma^2+n_{\ell_L}\sigma_\mu^2}-\frac{n_{\ell_R}^2\bar{R}_R^2}{\sigma^2+n_{\ell_R}\sigma_\mu^2}\right)\right)
\end{aligned}$$

Tree Structure Ratio

It is also clear this is just the inverse of the tree structure ratio for the grow step.

$$\begin{aligned}
\frac{\mathbb{P}\left(T^*\right)}{\mathbb{P}\left(T\right)} &= \left(\alpha\frac{\left(1+d_\eta\right)^\beta}{\left(2+d_\eta\right)^{2\beta}}\frac{p_\eta n_\eta}{p_{\eta_L}p_{\eta_R}n_{\eta_L}n_{\eta_R}}\right)^{-1} \\
&= \frac{1}{\alpha}\frac{\left(2+d_\eta\right)^{2\beta}}{\left(1+d_\eta\right)^\beta}\frac{p_{\eta_L}p_{\eta_R}n_{\eta_L}n_{\eta_R}}{p_\eta n_\eta}
\end{aligned}$$

Change and Swap Proposals

Due to the complexity of the bookkeeping, we do not consider these steps.

Implementation Details

We use what we have above to calculate r for grow and prune steps. In practice, we just take the log to avoid numerical problems.

One thing we would like to estimate is the likelihood of the trees over the lifetime of the Gibbs sampler.