Supplementary material for

Stochastic tree ensembles for regularized nonlinear regression

June 3, 2021

A Connection between XBART and CART theoretical split criterion, Condition 1

In this section, we first establish the connection between theoretical split criterion of XBART and that of CART. Suppose the current parent node is $\mathcal{A}$, the XBART theoretical split criterion of cutpoint candidate c_{jk} is

$$\begin{aligned}L^*_{\text{XBART}}(c_{jk}) = \frac{1}{\sigma^2}\mathbb{P}(\mathrm{x}^{(i)} \leq c_{jk} \mid \mathrm{x} \in \mathcal{A})\Big[\mathbb{E}(Y \mid \mathrm{x}^{(i)} \leq c_{jk}, \mathrm{x} \in \mathcal{A})\Big]^2 \\ + \frac{1}{\sigma^2}\mathbb{P}(\mathrm{x}^{(i)} > c_{jk} \mid \mathrm{x} \in \mathcal{A})\Big[\mathbb{E}(Y \mid \mathrm{x}^{(i)} > c_{jk}, \mathrm{x} \in \mathcal{A})\Big]^2.\end{aligned}$$

The CART theoretical split criterion is

$$\begin{aligned}L^*_{\text{CART}}(c_{jk}) = \mathbb{V}(Y \mid \mathrm{x} \in \mathcal{A}) - \mathbb{P}(\mathrm{x}^{(j)} \leq c_{jk} \mid \mathrm{x} \in \mathcal{A})\mathbb{V}(Y \mid \mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x} \in \mathcal{A}) \\ - \mathbb{P}(\mathrm{x}^{(j)} > c_{jk} \mid \mathrm{x} \in \mathcal{A})\mathbb{V}(Y \mid \mathrm{x}^{(j)} > c_{jk}, \mathrm{x} \in \mathcal{A}).\end{aligned}$$

Recall that the cuts are always parallel to axis,

$$\mathbb{V}(Y \mid \mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x} \in \mathcal{A}) = \mathbb{E}(Y^2 \mid \mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x} \in \mathcal{A}) - \Big[\mathbb{E}(Y \mid \mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x} \in \mathcal{A})\Big]^2.$$

We have

$$\mathbb{E}(Y^2 \mid \mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x} \in \mathcal{A}) = \frac{1}{\Omega\big(\{\mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x} \in \mathcal{A}\}\big)} \int_{\mathrm{x} \in \{\mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x} \in \mathcal{A}\}} m^2(\mathrm{x}) d\mathrm{x},$$

where $\Omega(A)$ represents volume of the cube $\mathcal{A}$. Observe that

$$\mathbb{P}(\mathrm{x}^{(j)} \leq c_{jk} \mid \mathrm{x} \in \mathcal{A}) = \frac{\Omega\big(\{\mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x} \in \mathcal{A}\}\big)}{\Omega(\mathcal{A})},$$

it yields

$$\begin{aligned}&\mathbb{E}(Y^2 \mid \mathrm{x} \in \mathcal{A}) - \mathbb{P}(\mathrm{x}^{(j)} \leq c_{jk} \mid \mathrm{x} \in \mathcal{A})\mathbb{E}(Y^2 \mid \mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x} \in \mathcal{A}) \\ &\quad - \mathbb{P}(\mathrm{x}^{(j)} > c_{jk} \mid \mathrm{x} \in \mathcal{A})\mathbb{E}(Y^2 \mid \mathrm{x}^{(j)} > c_{jk}, \mathrm{x} \in \mathcal{A}) \\ &= \frac{1}{\Omega(\mathcal{A})}\int_{\mathrm{x}\in\mathcal{A}} m^2(\mathrm{x})d\mathrm{x} - \frac{1}{\Omega(A)}\int_{\mathrm{x}\in\{\mathrm{x}^{(j)} \leq c_{jk}, \mathrm{x}\in\mathcal{A}\}} m^2(\mathrm{x})d\mathrm{x} - \frac{1}{\Omega(A)}\int_{\mathrm{x}\in\{\mathrm{x}^{(j)} > c_{jk}, \mathrm{x}\in\mathcal{A}\}} m^2(\mathrm{x})d\mathrm{x} \\ &= 0.\end{aligned}$$

As a result, the CART theoretical split criterion is equivalent to

$$\begin{aligned} L^*_{\text{CART}}(c_{jk}) &= [\mathbb{E}(Y \mid \mathrm{x} \in \mathcal{A})]^2 - \mathbb{P}(\mathrm{x}^{(i)} \leq c_{jk} \mid \mathrm{x} \in \mathcal{A})\Big[\mathbb{E}(Y \mid \mathrm{x}^{(i)} \leq c_{jk}, \mathrm{x} \in \mathcal{A})\Big]^2 \\ &\quad - \mathbb{P}(\mathrm{x}^{(i)} > c_{jk} \mid \mathrm{x} \in \mathcal{A})\Big[\mathbb{E}(Y \mid \mathrm{x}^{(i)} > c_{jk}, \mathrm{x} \in \mathcal{A})\Big]^2 \\ &= [\mathbb{E}(Y \mid \mathrm{x} \in \mathcal{A})]^2 - \sigma^2 L^*_{\text{XBART}}(c_{jk}). \end{aligned}$$

Since $[\mathbb{E}(Y \mid \mathrm{x} \in \mathcal{A})]^2$ and σ^2 are constants, and we maximize $L^*_{\text{XBART}}(c_{jk})$ but minimize $L^*_{\text{CART}}(c_{jk})$ in practice, we claim that the two theoretical split criterion are equivalent.

Scornet et al. (2015) show that Condition 1 (their Lemma 1) is valid under the assumption of f below

Assumption 1 (A1)**.**

$$y = \sum_{k=1}^{p} f^{(k)}(\mathrm{x}^{(k)}) + \epsilon,$$

where $\mathrm{x} = (\mathrm{x}^{(1)}, \cdots, \mathrm{x}^{(p)})$ *is uniformly distributed on* $[0,1]^p$*.* $\epsilon \sim N(0, s^2)$*. Each* $f^{(k)}(x)$ *is continuous on* $[0,1]$*.*

Since the theoretical split criterion of CART and XBART are equivalent, their proof applies directly without modification to XBART under the same Assumption (A1). We refer readers to Scornet et al. (2015) for details.

However, a weaker replacement of Assumption (A1) is to assume Condition 1 is valid directly, with extra assumption $||f||_\infty < \infty$, continuous on $[0,1]^p$. Although this is perhaps less interpretable than an assumption of an additive model, it is also presumably a weaker assumption in that it may be satisfied by non-additive models. Therefore in our paper, we assume Condition 1 is valid rather than Assumption A1.

B Proof of Lemma 1

Essentially the proof shows that adding small perturbation $\mathbf{c}'_k = (c'_1, \cdots, c'_k)$ to the **sequence** of cutpoints $\mathbf{c}_k = (c_1, \cdots, c_k)$, the change of split criterion at the bottom node $|L_{n,k}(\mathrm{x}, \mathbf{c}'_k) - L_{n,k}(\mathrm{x}, \mathbf{c}_k)|$ is small. In the proof, we show two different bounding strategies as follows,

S1 Perturb cutpoint of the **parent** of the bottom node.

S2 Perturb cutpoint of **nodes above the parent** (two levels above, or higher) of the bottom node.

We show that **S1** and **S2** are valid first, in the setting of $k = 1$ and $k = 2$, and demonstrate proof for general k with the two strategies.

B.1 Bounding strategy S1, Proof of Lemma 1 for the case $k = 1$

In this section, we start from **S1**, bounding the variation when perturb cutpoint of the parent node. Without loss of generality, we assume $k = 1$ and consider the first cut at the root node. Note that the case $L_{n,1}(\mathrm{x}, \cdot)$ does not depend on x, we write $L_{n,1}(\cdot)$ instead of $L_{n,1}(\mathrm{x}, \cdot)$.

Preliminary results Let $Z_i = \max_{1 \le i \le n} |\epsilon_i|$, we have

$$\mathbb{P}(Z_i \ge t) = 1 - \exp[n \ln(1 - 2\mathbb{P}(\epsilon_i \ge t))].$$

The tail of Gaussian distribution has a standard bound:

$$\mathbb{P}(\epsilon_i \ge t) \le \frac{\sigma}{t\sqrt{2\pi}} \left(-\frac{t^2}{2\sigma^2} \right).$$

As a result, there exist a positive constant C_ρ and $N_1 \in \mathbb{N}^*$ such that with probability $1 - \rho$, for all $n > N_1$,

$$\max_{1 \le i \le n} |\epsilon_i| \le C_\rho \sqrt{\log(n)}.$$

In addition, we have

$$\mathbb{P}\left[\left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \right| \ge \alpha \right] \le \frac{\sigma}{\alpha\sqrt{n}} \exp\left(-\frac{\alpha^2 n}{2\sigma^2} \right).$$

Let $\mathrm{N}_n(A)$ denotes number of data observations in a set A. Next we derive from the inequality above and union bound inequality that there exists $N_2 \in \mathbb{N}^*$ such that with probability $1 - \rho$, for all $n > N_2$ and all $0 \le a_n \le b_n \le 1$ satisfying $\mathrm{N}_n([a_n, b_n] \times [0,1]^{p-1}) > \sqrt{n}$,

$$\left| \frac{1}{\mathrm{N}_n([a_n, b_n] \times [0,1]^{p-1})} \sum_{i: \mathrm{x}_i \in [a_n, b_n] \times [0,1]^{p-1}} \epsilon_i \right| \le \alpha.$$

and

$$\frac{1}{\mathrm{N}_n([a_n, b_n] \times [0,1]^{p-1})} \sum_{i: \mathrm{x}_i \in [a_n, b_n] \times [0,1]^{p-1}} \epsilon_i^2 \le \tilde{\sigma}^2.$$

Furthermore, it's easy to verify

$$\left| \frac{1}{\mathrm{N}_n([a_n, b_n] \times [0,1]^{p-1})} \sum_{i: \mathrm{x}_i \in [a_n, b_n] \times [0,1]^{p-1}} Y_i \right| \le ||f||_\infty + \alpha, \tag{1}$$

where f is the true function, and

$$\left| \frac{1}{\mathrm{N}_n([a_n, b_n] \times [0,1]^{p-1})} \sum_{i: \mathrm{x}_i \in [a_n, b_n] \times [0,1]^{p-1}} Y_i^2 \right| \le ||f||_\infty^2 + \tilde{\sigma}^2 + 2\alpha ||f||_\infty. \tag{2}$$

By the Glivenko-Cantelli theorem, there exist $N_3 \in \mathbb{N}^*$ such that with probability $1-\rho$, for all $0 \leq a \leq b \leq 1$ and all $n > N_3$,

$$(b - a - \delta^2)n \leq \mathrm{N}_n([a_n, b_n] \times [0,1]^{p-1}) \leq (b - a + \delta^2)n. \tag{3}$$

In the following proof, we assume to be on the event that all claims above holds with probability $1 - 3\rho$ for all $n > N = \max\{N_1, N_2, N_3\}$. Take $c_1, c_2 \in [0,1]$ such that $|c_1 - c_2| < \delta$ and assume that $c_1 < c_2$. We partition the space $[0,1]^p$ into several pieces as follows, see Figure 1 for an illustration of notations for $k = 1$, where we project the p dimensional cells onto the first variable.

$$\begin{cases} A_{L,\sqrt{\delta}} = [0, \sqrt{\delta}] \times [0,1]^{p-1} \\ A_{R,\sqrt{\delta}} = [1-\sqrt{\delta}, 1] \times [0,1]^{p-1} \\ A_{C,\sqrt{\delta}} = [\sqrt{\delta}, 1-\sqrt{\delta}] \times [0,1]^{p-1} \end{cases}.$$

Similarly,

$$\begin{cases} A_{L,1} = [0, c_1] \times [0,1]^{p-1} \\ A_{R,1} = [c_1, 1] \times [0,1]^{p-1} \\ A_{L,2} = [0, c_2] \times [0,1]^{p-1} \\ A_{R,2} = [c_2, 1] \times [0,1]^{p-1} \\ A_C = [c_1, c_2] \times [0,1]^{p-1} \end{cases}.$$

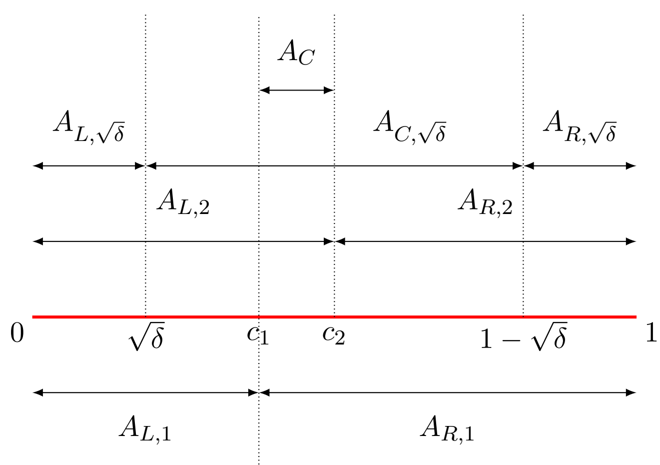

Figure 1: Illustration of notations for $k = 2$. Projection of the cells onto the first variable.

For simplicity, we write the split criterion of the first cut $\mathbf{c} = (1, c)$ as $L_{n,1}(1, c)$ denoting split at the first variable and value c. Recall that our split criterion is defined as

$$
\begin{aligned}
L_{n,1}(1,c) = & \frac{\tau \mathrm{N}_n(A_L)}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_L))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}\le c} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}\le c}(y_i-\bar{y}_l)^2\right) \\
& + \frac{\tau \mathrm{N}_n(A_R)}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_R))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}> c} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}> c}(y_i-\bar{y}_r)^2\right) \\
& + \frac{\gamma_x}{n}.
\end{aligned}
$$

The difference of split criterion on two cutvalues c_1 and c_2, on the first variable is

$$
\begin{aligned}
& L_{n,1}(1,c_1) - L_{n,1}(1,c_2) \\
= & \frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}\le c_1} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}\le c_1}(y_i-\bar{y}_{A_{L,1}})^2\right) \\
& + \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{R,1}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}> c_1} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}> c_1}(y_i-\bar{y}_{A_{R,1}})^2\right) \\
- & \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}\le c_2} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}\le c_2}(y_i-\bar{y}_{A_{L,2}})^2\right) \\
& - \frac{\tau \mathrm{N}_n(A_{R,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{R,2}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}> c_2} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}> c_2}(y_i-\bar{y}_{A_{R,2}})^2\right) \\
+ & \frac{\gamma_{c_1}}{n} - \frac{\gamma_{c_2}}{n}.
\end{aligned}
\tag{4}
$$

We need to prove Lemma 1 for all possible cases depending on location of c_1 and c_2. For notation simplicity, note that after collecting terms, the difference of split criterion can be represented as summation of points for the range of index $\{i : \mathrm{x}_i^{(1)} < c_1\}$, $\{i : \mathrm{x}_i^{(1)} \in [c_1, c_2]\}$ and $\{i : \mathrm{x}_i^{(1)} > c_2\}$. We will use the same decomposition throughout the proof.

First case

Assume that $c_1, c_2 \in A_{C,\sqrt{\delta}}$, two cutpoint candidates are not close to the edge. Consider the split criterion

$$
\begin{aligned}
&L_{n,1}(1,c_1) - L_{n,1}(1,c_2) \\
=&\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}\le c_1} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}\le c_1}(y_i-\bar{y}_{A_{L,1}})^2\right) \\
&\quad + \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{R,1}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}> c_1} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}> c_1}(y_i-\bar{y}_{A_{R,1}})^2\right) \\
-&\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}\le c_2} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}\le c_2}(y_i-\bar{y}_{A_{L,2}})^2\right) \\
&\quad - \frac{\tau \mathrm{N}_n(A_{R,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{R,2}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}> c_2} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}> c_2}(y_i-\bar{y}_{A_{R,2}})^2\right) \\
+&\frac{\gamma_{c_1}}{n} - \frac{\gamma_{c_2}}{n} \\
&= J_1 + J_2 + J_3 + \frac{\gamma_{c_1}}{n} - \frac{\gamma_{c_2}}{n}.
\end{aligned}
$$

First, take n large enough, we have

$$
\left|\frac{\gamma_{c_1}}{n} - \frac{\gamma_{c_2}}{n}\right| \le \alpha.
$$

Let J_2 correspond to $\{i \mid \mathrm{x}_i^{(1)} \in [c_1,c_2]\}$

$$
\begin{aligned}
J_2 &= \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{R,1}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}\in[c_1,c_2]} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}\in[c_1,c_2]}(y_i-\bar{y}_{A_{R,1}})^2\right) \\
&\quad - \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}\in[c_1,c_2]} y_i^2 - \sum_{i:\mathrm{x}_i^{(1)}\in[c_1,c_2]}(y_i-\bar{y}_{A_{L,2}})^2\right) \\
&= \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{R,1}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\in[c_1,c_2]} y_i^2\right) - \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\in[c_1,c_2]} y_i^2\right) \\
&\quad + \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\in[c_1,c_2]}(y_i-\bar{y}_{A_{L,2}})^2\right) \\
&\qquad - \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{R,1}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\in[c_1,c_2]}(y_i-\bar{y}_{A_{R,1}})^2\right) \\
&= J_{21} + J_{22}.
\end{aligned}
$$

Note that $|ax - by| \le |a||x-y| + |a-b||y|$, we have

$$
\begin{aligned}
|J_{22}| = & \left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} \left(\frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} (y_i - \bar{y}_{A_{L,2}})^2 \right) \right. \\
& \left. - \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{R,1}))} \left(\frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} (y_i - \bar{y}_{A_{R,1}})^2 \right) \right| \\
\leq & \left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} \right| \times \left| \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} (y_i - \bar{y}_{A_{L,2}})^2 - \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} (y_i - \bar{y}_{A_{R,1}})^2 \right| \\
& + \left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} - \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{R,1}))} \right| \times \left| \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} (y_i - \bar{y}_{A_{R,1}})^2 \right|.
\end{aligned}
$$

Since we assume that $c_1, c_2 \in A_{C,\sqrt{\delta}}$, by equation (3)

$$
\begin{aligned}
\left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} \right| & \leq \left| \frac{\tau(\delta^2 - \sqrt{\delta})n}{\sigma^2(\sigma^2 + \tau(1 - \delta^2 - \sqrt{\delta})n)} \right| \\
& \leq \left| \frac{\tau(\delta^2 - \sqrt{\delta})}{\sigma^2(\tau(1 - \delta^2 - \sqrt{\delta}))} \right| \\
& = C(\delta) \to 0 \text{ as } \delta \to 0.
\end{aligned} \tag{5}
$$

Note that this bound is valid for $\mathrm{N}_n(A_{L,1}), \mathrm{N}_n(A_{L,2}), \mathrm{N}_n(A_{R,1})$ and $\mathrm{N}_n(A_{R,2})$. By inequality (1) and (2), it is obvious that

$$
\left| \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} (y_i - \bar{y}_{A_{R,1}})^2 \right| \leq \left| \frac{1}{N(A_C)} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} (y_i - \bar{y}_{A_{R,1}})^2 \right| \leq M
$$

by a constant M. Furthermore

$$
\left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} - \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{R,1}))} \right| \leq 2C(\delta). \tag{6}
$$

The bound of second term follows equation (8) in supplementary materials of Scornet et al. (2015) directly,

$$
|J_{22}| \leq C(\delta) \times 4(||m||_\infty + \alpha)\big((\delta + \delta^2)(2||m||_\infty + \alpha) + \alpha\big) + 2C(\delta)M.
$$

The other term J_{21} is

$$
\begin{aligned}
|J_{21}| = & \left| \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{R,1}))} \left(\frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} y_i^2 \right) - \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} \left(\frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} y_i^2 \right) \right| \\
\leq & \left| \frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{R,1}))} - \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} \right| \times \left| \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \in [c_1,c_2]} y_i^2 \right|.
\end{aligned}
$$

The bound of coefficient here is slightly different from equation (5) and (6)

$$
\begin{aligned}
&\left|\frac{\tau \mathrm{N}_n(A_{R,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{R,1}))}-\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\right| \\
=&\left|\frac{\tau (\mathrm{N}_n(A_{R,1})-\mathrm{N}_n(A_{L,2}))}{(\sigma^2+\tau \mathrm{N}_n(A_{R,1}))(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\right| \\
\leq&\left|\frac{\tau}{(\sigma^2+\tau \mathrm{N}_n(A_{R,1}))(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\right|(|\mathrm{N}_n(A_{R,1})|+|\mathrm{N}_n(A_{L,2})|) \\
\leq&\frac{2\tau(1-\sqrt{\delta}+\delta^2)n}{\left(\sigma^2+\tau(1-\sqrt{\delta}-\delta^2)n\right)^2}=g(\delta,n)\to 0 \text{ when } n \text{ is large.}
\end{aligned}
\tag{7}
$$

Note that the upper bounds in equation (5) and (6) can be arbitrarily small if $\delta \to 0$, but the upper bound in equation (7) relies on making n large. Use the tail bound of non-central χ^2 distribution, result of supplementary materials of Scornet et al. (2015), and similar to J_{22}

$$
|J_{21}| \leq g(\delta,n)M, \tag{8}
$$

which can be arbitrarily small when n is large.

Next we switch to J_1, corresponding to $i \mid \mathrm{x}_i^{(1)} \in [0, c_1]$, we proceed with similar decomposition,

$$
\begin{aligned}
J_1 =&\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}\leq c_1} y_i^2-\sum_{i:\mathrm{x}_i^{(1)}\leq c_1}(y_i-\bar{y}_{A_{L,1}})^2\right) \\
&-\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\frac{1}{n}\left(\sum_{i:\mathrm{x}_i^{(1)}\leq c_1} y_i^2-\sum_{i:\mathrm{x}_i^{(1)}\leq c_1}(y_i-\bar{y}_{A_{L,2}})^2\right) \\
=&\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\leq c_1} y_i^2\right)-\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\leq c_1} y_i^2\right) \\
&+\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\leq c_1}(y_i-\bar{y}_{A_{L,2}})^2\right) \\
&-\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\leq c_1}(y_i-\bar{y}_{A_{L,1}})^2\right) \\
=&J_{11}+J_{12}.
\end{aligned}
$$

$$
\begin{aligned}
|J_{12}| = &\left|\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}(y_i-\bar{y}_{A_{L,2}})^2\right)\right. \\
&\left.\quad -\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}(y_i-\bar{y}_{A_{L,1}})^2\right)\right| \\
&\le \left|\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\right| \times \left|\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}(y_i-\bar{y}_{A_{L,2}})^2-\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}(y_i-\bar{y}_{A_{L,1}})^2\right| \\
&\quad + \left|\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}-\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\right| \times \left|\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}(y_i-\bar{y}_{A_{L,1}})^2\right|.
\end{aligned}
$$

Same as J_{22},

$$
\begin{aligned}
|J_{12}| &\le C(\delta)\times\left|\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\in[c_1,c_2]}(y_i-\bar{y}_{A_{R,1}})^2\right| + 2C(\delta)M \\
&\le C(\delta)\times 5(||f||_\infty\sqrt{\delta}+\alpha)+2C(\delta)M.
\end{aligned}
$$

The second equation above use result of equation (9) of supplementary material of Scornet et al. (2015). Similar to J_{21}, we have

$$
\begin{aligned}
|J_{11}| &= \left|\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}y_i^2\right)-\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}y_i^2\right)\right| \\
&\le \left|\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}-\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\right|\times\left|\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}y_i^2\right| \\
&\le g(\delta,n)M.
\end{aligned}
$$

J_3 have the same bound as J_1. Collecting all terms yields

$$
\begin{aligned}
&|J_1| \le g(\delta,n)M + C(\delta)\times 25(||f||_\infty\sqrt{\delta}+\alpha)+2C(\delta)M \\
&|J_2| \le g(\delta,n)M + C(\delta)\times 4(||f||_\infty+\alpha)\big((\delta+\delta^2)(2||f||_\infty+\alpha)+\alpha\big)+2C(\delta)M \\
&|J_3| \le g(\delta,n)M + C(\delta)\times 25(||f||_\infty\sqrt{\delta}+\alpha)+2C(\delta)M \\
&|L_{n,1}(1,c_1)-L_{n,1}(1,c_2)| \le |J_1|+|J_2|+|J_3|
\end{aligned}.
$$

Consequently, for all n large enough and δ small enough, we have

$$
|L_{n,1}(1,c_1)-L_{n,1}(1,c_2)| \le 3\alpha.
$$

Second case

Assume that $c_1, c_2 \in A_{L,\sqrt{\delta}}$, take same arguments as above, we have

$$\mathrm{N}_n(A_{L,1}), \mathrm{N}_n(A_{L,2}) \leq (\sqrt{\delta} + \delta^2)n.$$

Different from the first case, now both c_1 and c_2 are close to the left edge, which is corresponding to term J_1. Note that $|J_2|$ and $|J_3|$ are the same as the first case since the control over region A_C and $A_{R,1} \times A_{R,2}$ and not changed.

$$\begin{aligned}
|J_{12}| &= \left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} \left(\frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \leq c_1} (y_i - \bar{y}_{A_{L,2}})^2 \right) \right. \\
&\quad \left. - \frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,1}))} \left(\frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \leq c_1} (y_i - \bar{y}_{A_{L,1}})^2 \right) \right| \\
&\leq \left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} \right| \times \left| \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \leq c_1} (y_i - \bar{y}_{A_{L,2}})^2 - \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \leq c_1} (y_i - \bar{y}_{A_{L,1}})^2 \right| \\
&\quad + \left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} - \frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,1}))} \right| \times \left| \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \leq c_1} (y_i - \bar{y}_{A_{L,1}})^2 \right|.
\end{aligned}$$

We have

$$\left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{L,2}))} \right| \leq \left| \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2 \tau \mathrm{N}_n(A_{L,2})} \right| = \frac{1}{\sigma^2}$$

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \leq c_1} (y_i - \bar{y}_{A_{L,2}})^2 - \frac{1}{n} \sum_{i:\mathrm{x}_i^{(1)} \leq c_1} (y_i - \bar{y}_{A_{L,1}})^2 \right| \\
&= 2|\bar{y}_{A_{L,1}} - \bar{y}_{A_{L,2}}| \times \frac{1}{n} \left| \sum_{i:\mathrm{x}_i^{(1)} < c_1} \left(y_i - \frac{\bar{y}_{A_{L,1}} + \bar{y}_{A_{L,2}}}{2} \right) \right| \\
&\leq 4(||f||_\infty + \alpha) \left(\frac{(||f||_\infty + \alpha)\mathrm{N}_n(A_{L,1})}{n} + \frac{1}{n} \left| \sum_{i:\mathrm{x}_i^{(1)} < c_1} f(\mathrm{x}_i) + \epsilon_i \right| \right) \\
&\leq 4(||f||_\infty + \alpha) \Big((||f||_\infty + \alpha)(\sqrt{\delta} + \delta^2) + \frac{\mathrm{N}_n(A_{L,1})}{n} (||f||_\infty + \alpha) \Big) \\
&\leq 4(||f||_\infty + \alpha) \Big((||f||_\infty + \alpha + 1)(\sqrt{\delta} + \delta^2) \Big)
\end{aligned}$$

$$
\begin{aligned}
&\left|\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))} - \frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\right| \times \left|\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}(y_i-\bar{y}_{A_{L,1}})^2\right| \\
=&\left|\frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))} - \frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\right| \times \frac{\mathrm{N}_n(A_{L,1})}{n}\left|\frac{1}{\mathrm{N}_n(A_{L,1})}\sum_{i:\mathrm{x}_i^{(1)}\le c_1}(y_i-\bar{y}_{A_{L,1}})^2\right| \\
\le&\frac{2}{\sigma^2}(\sqrt{\delta}+\delta^2)M.
\end{aligned}
$$

As a result

$$
|J_{12}| \le \frac{1}{\sigma^2}4(||f||_\infty+\alpha)\Big((||f||_\infty+\alpha+1)(\sqrt{\delta}+\delta^2)\Big) + \frac{2}{\sigma^2}(\sqrt{\delta}+\delta^2)M \to 0.
$$

$$
\begin{aligned}
|J_{11}| &= \left|\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1} y_i^2\right) - \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\left(\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1} y_i^2\right)\right| \\
&\le \left|\frac{\tau \mathrm{N}_n(A_{L,1})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,1}))} - \frac{\tau \mathrm{N}_n(A_{L,2})}{\sigma^2(\sigma^2+\tau \mathrm{N}_n(A_{L,2}))}\right| \times \left|\frac{1}{n}\sum_{i:\mathrm{x}_i^{(1)}\le c_1} y_i^2\right| \\
&\le \frac{2}{\sigma^2}(\sqrt{\delta}+\delta^2)M \to 0.
\end{aligned}
$$

Consequently we conclude that for all $n > N$ and all δ small enough,

$$
|L_{n,1}(1,c_1) - L_{n,1}(1,c_2)| \le 3\alpha.
$$

The other cases $\{c_1, c_2 \in A_{R,\sqrt{\delta}}\}$, $\{c_1 \in A_{L,\sqrt{\delta}}, c_2 \in A_{C,\sqrt{\delta}}\}$ and $\{c_1 \in A_{C,\sqrt{\delta}}, c_2 \in A_{R,\sqrt{\delta}}\}$ can be proved using similar strategy. Details are omitted.

B.2 Bounding strategy S2, Proof of Lemma 1 for the case $k = 2$

Next we prove **S2**, when adding variation to nodes above the parent of bottom nodes, the variation of split criterion is bounded. First, we assume $k = 2$.

Preliminary results

Similarly, Laurent and Massart (2000) gives tail bound of χ^2 distribution,

$$
\mathbb{P}\big[\chi_n^2 \ge 5n\big] \le \exp(-n).
$$

By the tail bound above, it's straightforward to show that

Suppose X follows χ^2 distribution with degrees of freedom k and non-central parameter λ

$$
P(X \ge x) \le \frac{\sqrt{\pi}}{2e}\Phi(\sqrt{x})I_{\frac{k}{2}}(1)M_{k-1},
$$

where I_v is a modified Bessel function of the first kind, $M_{k-1} = E(y^{k-1})$ and y is a Gaussian $(\mu, 1)$ random variable truncated on $(\sqrt{x}, \infty)$. So we can claim that with probability $1-\rho$, the term $\frac{1}{n}\sum_{i=1}^{n} y_i^2$ is bounded.

Follow the notation of Scornet et al. (2015), let $d_1' = (1, c_1')$ and $d_2' = (2, x_2'$ be such that $|c_1 - c_1'| \le \delta$ and $|c_2 - x_2'| \le \delta$.

There exist a constant $C_\rho > 0$ and N_1 such that, with probability $1 - \rho$, for all $n > N_1$,

$$\max_{1 \le i \le n} |\epsilon_i| \le C_\rho \sqrt{\log(n)} \tag{9}$$

and

$$\max_{1 \le i \le n} |\epsilon_i^2| \le C_\rho^2 \log(n). \tag{10}$$

Fix $\rho > 0$, there exist N_2 such that, with probability $1 - \rho$, for all $n > N_2$ and all $A_n = [a_n^{(1)}, b_n^{(1)}] \times [a_n^{(2)}, b_n^{(2)}] \subset [0,1]^2$ satisfying $\mathrm{N}_n(A_n) > \sqrt{n}$,

$$\left| \frac{1}{\mathrm{N}_n(A_n)} \sum_{i:\mathrm{x}_i \in A_n} \epsilon_i \right| \le \alpha$$

and

$$\frac{1}{\mathrm{N}_n(A_n)} \sum_{i:\mathrm{x}_i \in A_n} \epsilon_i^2 \le \tilde{\sigma}^2.$$

Furthermore, it's easy to verify

$$\left| \frac{1}{\mathrm{N}_n(A_n)} \sum_{i:\mathrm{x}_i \in A_n} y_i \right| \le ||f||_\infty + \alpha \tag{11}$$

and

$$\left| \frac{1}{\mathrm{N}_n(A_n)} \sum_{i:\mathrm{x}_i \in A_n} y_i^2 \right| \le ||f||_\infty^2 + \tilde{\sigma}^2 + 2\alpha ||f||_\infty. \tag{12}$$

Similar to the $k = 1$ case, we denote partition of space as

$$\begin{cases} A_{R,1} = [c_1, 1] \times [0,1]^{p-1} \\ A_{B,2} = [c_1, 1] \times [0, c_2] \times [0,1]^{p-2} \\ A_{H,2} = [c_1, 1] \times [c_2, 1] \times [0,1]^{p-2} \\ A'_{B,2} = [c_1', 1] \times [0, c_2'] \times [0,1]^{p-2} \\ A'_{H,2} = [c_1', 1] \times [c_2', 1] \times [0,1]^{p-2}. \end{cases}$$

Figure 2 shows projection of the cells onto the first two variables.

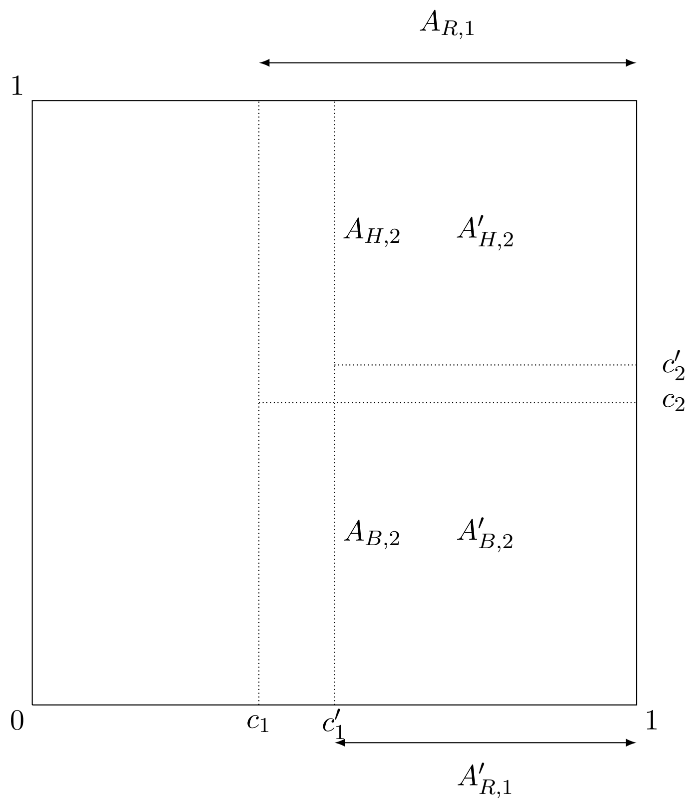

Figure 2: Illustration of notations for $k = 2$. Projection of cells onto the first two variables (assuming they are different variables).

Let $d_1 = (1, c_1)$ denotes cutpoint that splits at the first variable, value c_1, similarly $d_2 = (2, c_2)$, $d'_1 = (1, c'_1)$ and $d'_2 = (2, c'_2)$ be four cutpoints and $|c_1 - c'_1| < \delta$, $|c_2 - c'_2| < \delta$, then

$$\begin{aligned} L_n(d_1, d_2) - L_n(d'_1, d'_2) = L_n(d_1, d_2) - L_n(d'_1, d_2) \\ + L_n(d'_1, d_2) - L_n(d'_1, d'_2). \end{aligned} \tag{13}$$

It is noteworthy that equation (13) decomposes the variation to two terms, where the second term applies bounding strategies **S1** directly, and the first term is variation when the cutpoint of grandparent (two levels above bottom node) is perturbed.

$$
\begin{aligned}
& L_n(d_1, d_2) - L_n(d_1', d_2) \\
=& \frac{\tau \mathrm{N}_n(A_{B,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{B,2}))} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} \leq c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1} - \sum_{i:\mathrm{x}_i^{(2)} \leq c_2} (y_i - \bar{y}_{A_{B,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1} \right) \\
& \quad + \frac{\tau \mathrm{N}_n(A_{H,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{H,2}))} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1} - \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1} \right) \\
-& \frac{\tau \mathrm{N}_n(A'_{B,2})}{\sigma^2\left(\sigma^2 + \tau \mathrm{N}_n(A'_{B,2})\right)} \frac{1}{\mathrm{N}_n(A'_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} \leq c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} \leq c_2} (y_i - \bar{y}_{A'_{B,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right) \\
& \quad - \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2\left(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\right)} \frac{1}{\mathrm{N}_n(A'_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2'} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right) \\
+& \frac{\gamma_{c_1,c_2}}{\mathrm{N}_n(A_{R,1})} - \frac{\gamma_{c_1',c_2}}{\mathrm{N}_n(A'_{R,1})} \\
&= A_1 + B_1
\end{aligned}
$$

$$
\begin{aligned}
A_1 =& \frac{\tau \mathrm{N}_n(A_{H,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{H,2}))} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1} - \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1} \right) \\
& \quad - \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2\left(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\right)} \frac{1}{\mathrm{N}_n(A'_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2'} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right) \\
=& A_{1,1} + A_{1,2} + A_{1,3}
\end{aligned}
$$

$$
\begin{aligned}
A_{1,1} =& \frac{\tau \mathrm{N}_n(A_{H,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{H,2}))} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right) \\
& \quad - \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2\left(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\right)} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right)
\end{aligned}
$$

$$
\begin{aligned}
A_{1,2} =& \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2\left(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\right)} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right) \\
& \quad - \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2\left(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\right)} \frac{1}{\mathrm{N}_n(A'_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2'} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right)
\end{aligned}
$$

$$A_{1,3} = \frac{\tau \mathrm{N}_n(A_{H,2})}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{H,2}))} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} \in [c_1, c_1']} - \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} \in [c_1, c_1']} \right)$$

$$\begin{aligned}
A_{1,1} =& \frac{\tau \mathrm{N}_n(A_{H,2}}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{H,2}))} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right) \\
& - \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2 \Big(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\Big)} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right) \\
=& \left(\frac{\tau}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{H,2}))} \frac{\mathrm{N}_n(A_{H,2})}{\mathrm{N}_n(A_{R,1})} - \frac{\tau}{\sigma^2 \Big(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\Big)} \frac{\mathrm{N}_n(A'_{H,2})}{\mathrm{N}_n(A_{R,1})} \right) \sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \\
& + \left(\frac{\tau}{\sigma^2 \Big(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\Big)} \frac{\mathrm{N}_n(A'_{H,2})}{\mathrm{N}_n(A_{R,1})} \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right. \\
& \left. - \frac{\tau}{\sigma^2(\sigma^2 + \tau \mathrm{N}_n(A_{H,2}))} \frac{\mathrm{N}_n(A_{H,2})}{\mathrm{N}_n(A_{R,1})} \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right)
\end{aligned}$$

which goes to zero using the same argument as $k = 1$ case.

$$\begin{aligned}
A_{1,2} =& \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2 \Big(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\Big)} \frac{1}{\mathrm{N}_n(A_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right) \\
& - \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2 \Big(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\Big)} \frac{1}{\mathrm{N}_n(A'_{R,1})} \left(\sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \sum_{i:\mathrm{x}_i^{(2)} > c_2'} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right) \\
=& \left(\frac{\tau}{\sigma^2 \Big(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\Big)} \frac{1}{\mathrm{N}_n(A_{R,1})} - \frac{\tau}{\sigma^2 \Big(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\Big)} \frac{1}{\mathrm{N}_n(A'_{R,1})} \right) \\
& \times \left(\mathrm{N}_n(A'_{H,2}) \sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} - \mathrm{N}_n(A'_{H,2}) \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right)
\end{aligned}$$

$$\begin{aligned}
|A_{1,2}| \leq & \left| \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2 \Big(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\Big)} \frac{\mathrm{N}_n(A'_{H,2})}{\mathrm{N}_n(A_{R,1})} - \frac{\tau \mathrm{N}_n(A'_{H,2})}{\sigma^2 \Big(\sigma^2 + \tau \mathrm{N}_n(A'_{H,2})\Big)} \frac{\mathrm{N}_n(A'_{H,2})}{\mathrm{N}_n(A'_{R,1})} \right| \\
& \times \left(\left| \frac{1}{\mathrm{N}_n(A'_{H,2})} \sum_{i:\mathrm{x}_i^{(2)} > c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right| + \left| \frac{1}{\mathrm{N}_n(A'_{H,2})} \sum_{i:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A'_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} > c_1'} \right| \right)
\end{aligned}$$

Same as before, the second term is bounded and

$$|A_{1,2}| \leq M\left|\frac{\mathrm{N}_n(A'_{H,2})}{\mathrm{N}_n(A_{R,1})} - \frac{\mathrm{N}_n(A'_{H,2})}{\mathrm{N}_n(A'_{R,1})}\right| \to 0$$

$$\begin{aligned}
|A_{1,3}| \leq & \left|\frac{\tau}{\sigma^2(\sigma^2+\tau\mathrm{N}_n(A_{H,2}))}\frac{\mathrm{N}_n(A_{H,2})}{\mathrm{N}_n(A_{R,1})}\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big)\right| \\
& \times \left|\frac{1}{\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big)} \sum_{i:\mathrm{x}_i^{(2)}>c_2} y_i^2 \mathbb{1}_{\mathrm{x}_i^{(1)} \in [c_1,c_1']}\right| \\
& + \left|\frac{\tau}{\sigma^2(\sigma^2+\tau\mathrm{N}_n(A_{H,2}))}\frac{\mathrm{N}_n(A_{H,2})}{\mathrm{N}_n(A_{R,1})}\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big)\right| \\
& \times \left|\frac{1}{\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big)} \sum_{i:\mathrm{x}_i^{(2)}>c_2} (y_i - \bar{y}_{A_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} \in [c_1,c_1']}\right| \\
& = A_{1,3,1} + A_{1,3,2}.
\end{aligned}$$

Note that $\frac{\tau\mathrm{N}_n(A_{H,2})}{\sigma^2\big(\sigma^2+\tau\mathrm{N}_n(A_{H,2})\big)}$ is bounded by a constant M as n is large,

$$\left|\frac{\tau}{\sigma^2(\sigma^2+\tau\mathrm{N}_n(A_{H,2}))}\frac{\mathrm{N}_n(A_{H,2})}{\mathrm{N}_n(A_{R,1})}\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big)\right| \leq M\frac{\delta^2+\delta}{\delta^2-\sqrt{\delta}} \to 0.$$

So we have $A_{1,3,1} \to 0$ if n is large and δ is small.

1. If $\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big) < \sqrt{n}$,

$$\left|\frac{1}{\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big)} \sum_{l:\mathrm{x}_i^{(2)}>c_2} (y_i - \bar{y}_{A_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} \in [c_1,c_1']}\right| \leq \frac{C_\rho^2\log(n)}{\sqrt{n}}.$$

2. If $\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big) > \sqrt{n}$, note that

$$|1-c_1| \geq \xi,$$

$$\mathrm{N}_n(A_{R,1}) > \mathrm{N}_n(\xi) > (\xi-\delta^2)n,$$

and

$$\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big) \leq \mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\}\Big) \leq (\delta+\delta^2)n.$$

As a result

$$\left|\frac{\mathrm{N}_n\Big(\Big\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\Big\} \times \Big\{\mathrm{x}_i^{(2)} > c_2\Big\}\Big)}{\mathrm{N}_n(A_{R,1})}\right| \leq \frac{\delta-\delta^2}{\xi+\delta^2} \leq \frac{\delta}{\xi}$$

$$\left|\frac{\mathrm{N}_n\left(\left\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\right\} \times \left\{\mathrm{x}_i^{(2)} > c_2\right\}\right)}{\mathrm{N}_n(A_{R,1})}\right|$$

$$\times \left|\frac{1}{\mathrm{N}_n\left(\left\{\mathrm{x}_i^{(1)} \in [c_1, c_1']\right\} \times \left\{\mathrm{x}_i^{(2)} > c_2\right\}\right)} \sum_{l:\mathrm{x}_i^{(2)} > c_2} (y_i - \bar{y}_{A_{H,2}})^2 \mathbb{1}_{\mathrm{x}_i^{(1)} \in [c_1, c_1']}\right|$$

$$\leq \frac{\delta}{\xi}\big(3(||f||_\infty + \alpha)^2 + ||f||_\infty^2 + \tilde{\sigma}^2 + 2||f||_\infty^2 \alpha\big).$$

Therefore $A_{1,3,2} \to 0$. Collecting all bounds, we conclude that $A_1 \to 0$. Bounding strategy **S1** proves that $B_1 \to 0$, thus we have $L_n(d_1, d_2) - L_n(d_1', d_2') \to 0$.

Remark Bounding strategy **S2** applies to perturbation at higher nodes as well. For example if we consider a sequence of three cutpoints $\mathbf{c}_3 = (d_1, d_2, d_3)$ and perturbation $\mathbf{c}_3' = (d_1', d_2', d_3')$. We can show that $L_n(d_1, d_2, d_3) - L_n(d_1', d_2, d_3)$ is bounded using the same argument as **S2** since the second cutpoint d_2 is the same, and the cells can be projected on space spanned by variable 1 and 3 similarly as Figure 2.

B.3 Proof of Lemma 1 for the case $k > 2$

The proof for general $k > 2$ is based on the two bounding strategies **S1** and **S2** above. We scratch the essential ideas behind the proof in this section.

First, same as equation (4) and (13), we express the variation of split criterion as sum of terms on one cutpoint at a time

$$\begin{aligned}
&L_{n,k}(d_1, d_2, \cdots, d_{k-1}, d_k) - L_{n,k}(d_1', d_2', \cdots, d_{k-1}', d_k') \\
=&L_{n,k}(d_1, d_2, \cdots, d_{k-1}, d_k) - L_{n,k}(d_1', d_2, \cdots, d_{k-1}, d_k) \quad \mathbf{S2} \\
+&L_{n,k}(d_1', d_2, \cdots, d_{k-1}, d_k) - L_{n,k}(d_1', d_2', \cdots, d_{k-1}, d_k) \quad \mathbf{S2} \\
+&\cdots \\
+&L_{n,k}(d_1', d_2', \cdots, d_{k-1}, d_k) - L_{n,k}(d_1', d_2', \cdots, d_{k-1}', d_k) \quad \mathbf{S2} \\
+&L_{n,k}(d_1', d_2', \cdots, d_{k-1}', d_k) - L_{n,k}(d_1', d_2', \cdots, d_{k-1}', d_k') \quad \mathbf{S1}
\end{aligned} \tag{14}$$

Following remark in section B.2, the first $k-1$ terms are perturbation at cutpoints above the parent of bottom nodes (two levels above, or higher) and all cutpoints in between are fixed, therefore we can project cells to the variable being perturbed and the last cutpoint d_k similarly as in Figure 2 to estimate the bound. The last term is perturbation of the parent node of the bottom, therefore **S1** applies.

References

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 1302–1338.

Scornet, E., G. Biau, J.-P. Vert, et al. (2015). Consistency of random forests. *The Annals of Statistics 43*(4), 1716–1741.