Supplementary File for Reinforcement Learning Trees

Ruoqing Zhu [*], Donglin Zeng [†] and Michael R. Kosorok [‡]

February 11, 2015

[*]R. Zhu (E-mail: rzhu@live.unc.edu) is a student; [†] D. Zeng (E-mail: dzeng@email.unc.edu) is Professor; and [‡] M. R. Kosorok (E-mail: Kosorok@unc.edu) is W. R. Kenan, Jr. Distinguished Professor and Chair; Department of Biostatistics, CB#7420, University of North Carolina, Chapel Hill, NC 27599-7420. The authors thank the editors and reviewers for their careful review and thoughtful suggestions which led to a significantly improved paper. The authors were funded in part by grant CA142538 from the National Cancer Institute.

1. PROOF OF THEOREM 3.6

Proof. The proof consists of the following four steps.

Step 1: We first establish the asymptotic results for the variable importance measure. Without further specification, the proof of Step 1 is conditional on an internal node A with sample size n_A and number of non-muted variables equal to p_A. We denote the internal node dataset by $\mathcal{D}_A = \{(X_i, Y_i), i \in A\}$. Let $\mathbb{P}$ be the probability measure of $(\mathbf{X}, Y)$ and let $\mathbb{P}$ be the corresponding empirical measure.

First, we observe that $VI_A(j)$ is bounded. By Assumption 3.1, f is Lipschitz continuous with Lipschitz constant c_f,

$$\begin{aligned} & VI_A(j) \\ = \quad & \frac{E[E[(f(X_i^{(1)}, ..., \tilde{X}_i^{(j)}, ..., X_i^{(p)}) - f(X_i^{(1)}, ..., X_i^{(j)}, ..., X_i^{(p)}))^2 | X_i^{(j)}] | A]}{\sigma^2} \\ \leq \quad & \frac{E[E[(c_f \cdot (b_j - a_j))^2 | X_i^{(j)}] | A]}{\sigma^2} = \frac{c_f^2 \cdot (b_j - a_j)^2}{\sigma^2}. \end{aligned} \tag{1}$$

Hence $VI_A(j)$ is also bounded above by the interval length of $X^{(j)}$, i.e., $(b_j - a_j)$, in A. It can be further bounded above by $\frac{c_f^2}{\sigma^2}$ since $(b_j - a_j) < 1$ for any internal node A.

Now we show that $\widehat{VI}_A(j)$ converges to $VI_A(j)$ at an exponential rate. For simplicity, assume that the embedded model $\widehat{f}_A^*$ randomly selects half of $\mathcal{D}_A$ to fit the model, denoted by $\mathcal{D}_{A_1}$, and the variable importance is calculated using the other half of the data, denoted by $\mathcal{D}_{A_2}$. Note that this is exactly (except for the proportion of each subset) what we do for each tree in a standard random forests model. However, with the potential use of other models, this simplifies the formulation. Further, since the j-th variable importance measure is calculated by randomly permuting the values of $X_i^{(j)}$ in $\mathcal{D}_{A_2}$, which we denote by $\tilde{X}_i^{(j)}$, we assume that this permutation is done infinitely many times. Then, for the i-th observation in $\mathcal{D}_{A_2}$, the squared error after permutation is $E_{\tilde{X}_i^{(j)}} \big(\widehat{f}_A^*(X_i^{(1)}, ..., \tilde{X}_i^{(j)}, ..., X_i^{(p)}) - Y_i \big)^2$. Hence the j-th variable importance can be formulated as:

$$\begin{aligned}
&\widehat{VI}_A(j)\\
=& \frac{\frac{1}{n_A/2}\sum_{\mathbf{X}_i\in\mathcal{D}_{A_2}} E_{\tilde{X}^{(j)}}\left(\widehat{f}^*_A(X_i^{(1)},...,\tilde{X}^{(j)},...,X_i^{(p)})-Y_i\right)^2}{\frac{1}{n_A/2}\sum_{\mathbf{X}_i\in\mathcal{D}_{A_2}} E_{\tilde{X}^{(j)}}\left(\widehat{f}^*_A(X_i^{(1)},...,X_i^{(j)},...,X_i^{(p)})-Y_i\right)^2}-1\\
=& \frac{\frac{1}{n}\sum_{\mathbf{X}_i\in\mathcal{D}} E_{\tilde{X}^{(j)}}\left(\widehat{f}^*_A(X_i^{(1)},...,\tilde{X}^{(j)},...,X_i^{(p)})-Y_i\right)^2 I_{[\mathbf{X}_i\in A_2]}}{\frac{1}{n}\sum_{\mathbf{X}_i\in\mathcal{D}} E_{\tilde{X}^{(j)}}\left(\widehat{f}^*_A(X_i^{(1)},...,X_i^{(j)},...,X_i^{(p)})-Y_i\right)^2 I_{[\mathbf{X}_i\in A_2]}}-1,
\end{aligned} \tag{2}$$

where $I[\mathbf{X}_i \in A_2]$ is the indicator function denoting whether $\mathbf{X}_i$ falls into the internal node A, and is randomized with probability $\frac{1}{2}$ to $\mathcal{D}_{A_2}$ for calculating variable importance. Let the set $(X_i^{(1)},...,X_i^{(j-1)},X_i^{(j+1)}...,X_i^{(p)})$ be $X_i^{(-j)}$. Then the numerator of the first term of (2) can be broken down into:

$$\begin{aligned}
&\frac{1}{n}\sum_{\mathbf{X}_i\in\mathcal{D}} E_{\tilde{X}^{(j)}}\left(\widehat{f}^*_A(X_i^{(1)},...,\tilde{X}^{(j)},...,X_i^{(p)})-Y_i\right)^2 I_{[\mathbf{X}_i\in A_2]}\\
=\;& \mathbb{P}_n\Big(E_{\tilde{X}^{(j)}}\big(\widehat{f}^*_A(X^{(-j)},\tilde{X}^{(j)})-Y\big)^2 I_{[\mathbf{X}\in A_2]}\Big)\\
=\;& (\mathbb{P}_n-\mathbb{P})\Big(E_{\tilde{X}^{(j)}}(\widehat{f}^*_A(X^{(-j)},\tilde{X}^{(j)})-Y)^2 I_{[\mathbf{X}\in A_2]}\Big)\\
&+\mathbb{P}\Big(E_{\tilde{X}^{(j)}}\big(\widehat{f}^*_A(X^{(-j)},\tilde{X}^{(j)})-f_A(X^{(-j)},\tilde{X}^{(j)})\big)^2 I_{[\mathbf{X}\in A_2]}\Big)\\
&+\mathbb{P}\Big(E_{\tilde{X}^{(j)}}\big(f_A(X^{(-j)},\tilde{X}^{(j)})-f_A(X^{(-j)},X^{(j)})\big)^2 I_{[\mathbf{X}\in A_2]}\Big)\\
&+\mathbb{P}\Big(E_{\tilde{X}^{(j)}}\big(f_A(X^{(-j)},X^{(j)})-Y\big)^2 I_{[\mathbf{X}\in A_2]}\Big)\\
=:\;& \tilde{T}_1+\tilde{T}_2+\tilde{T}_3+\tilde{T}_4.
\end{aligned} \tag{3}$$

Now we bound each of the four terms in Equation (3). We will first show the bound for $\tilde{T}_1$ and then for $\tilde{T}_2$, following the same idea. We use Theorem 8 in van de Geer and Lederer (2011) to establish the bound for $\tilde{T}_1$. The Theorem states that for any function $g(\mathbf{X})$ that lives in a collection of functions $\mathcal{G}$, if the Bernstein condition

$$\sup_{g\in\mathcal{G}} E|g|^m \le \frac{m!}{2}K^{m-2},\ m=2,3,... \tag{4}$$

is satisfied for some constant $K \ge 1$, then $\sqrt{n}(\mathbb{P}_n-\mathbb{P})g$ has exponential tail.

By Assumption 3.4, $\widehat{f}^*$ has exponential tail. On the other hand, $Y = f(\mathbf{X}) + \epsilon$, and $f(\mathbf{X})$ are bounded, and hence Y also satisfies the moment condition by Assumption 3.5. Hence, we can find some constant K such that the following Bernstein condition is satisfied:

$$\sup_{\widehat{f}^*} E\Big|f_A^*(X^{(-j)}, \tilde{X}^{(j)}) - Y\Big|^m \leq \frac{m!}{2} K^{m-2},\ m = 2, 3, ... \tag{5}$$

Furthermore, since $\widehat{f}^*$ has finite entropy integral by Assumption 3.4, we can use Theorem 8 in van de Geer and Lederer (2011) and reorganize the terms to find a constant $K_1^* > 0$ such that:

$$P\left(\sup\Big|\sqrt{n}\tilde{T}_1\Big| \geq t\right) \leq e^{-t/K_1^*}. \tag{6}$$

For $\tilde{T}_2$, we first write it into a conditional probability $\mathbb{P}_{A_2}$ such that

$$\begin{aligned}\tilde{T}_2 &= \mathbb{P}_{A_2}\Big(E_{\tilde{X}^{(j)}}\big(\widehat{f}_A^*(X^{(-j)}, \tilde{X}^{(j)}) - f_A(X^{(-j)}, \tilde{X}^{(j)})\big)^2\Big)P(A_2)\\ &= \tilde{T}_2^* P(A_2).\end{aligned} \tag{7}$$

For $\tilde{T}_2^*$ in the above equation, noting Assumption 3.4 for the error bound of f_A^*, and following similar arguments as applied to $\tilde{T}_1$, we have for some constant $K_2^* > 0$:

$$P\left(\sup\Big|\sqrt{n_A^{\eta(p_A)}}\tilde{T}_2^*\Big| \geq t\right) \leq e^{-t/K_2^*}. \tag{8}$$

For the other two terms, it is easy to see by Definition 2.1 that $\tilde{T}_3 = VI_A(j)\sigma^2 P(A_2)$, and $\tilde{T}_4 = \sigma^2 P(A_2)$ by Assumption 3.5.

Note that the denominator of the first term in (2) can be decomposed into four terms: T_1, T_2, T_3^* and T_4, similar to (3) but with $X_i^{(j)}$ in lieu of $\tilde{X}_i^{(j)}$. The first two terms can be bounded in the same way as the above. The third term equals 0 since $\tilde{X}_i^{(j)}$ is replaced by $X_i^{(j)}$. And the fourth term $T_4 = \sigma^2 P(A_2)$.

Hence, together with (6), (8) for the numerator, and the above arguments for the denominator, we can derive that

$$\begin{aligned}
& P\Big(\Big|\widehat{VI}_A(j) - VI_A(j)\Big| > C\Big) \\
= \ & P\Big(\Big|\frac{\tilde{T}_1 + \tilde{T}_2^* P(A_2) + \sigma^2 P(A_2) VI_A(j) + \sigma^2 P(A_2)}{T_1 + T_2^* P(A_2) + 0 + \sigma^2 P(A_2)} - 1 - VI_A(j)\Big| > C\Big) \\
\leq \ & P\Big(\Big|\frac{\tilde{T}_1}{T_1 + T_2^* P(A_2) + \sigma^2 P(A_2)}\Big| > C/3\Big) \\
& + P\Big(\Big|\frac{\tilde{T}_2^* P(A_2)}{T_1 + T_2^* P(A_2) + \sigma^2 P(A_2)}\Big| > C/3\Big) \\
& + P\Big(\Big|\frac{\sigma^2 P(A_2)(VI_A(j)+1)}{T_1 + T_2^* P(A_2) + \sigma^2 P(A_2)} - 1 - VI_A(j)\Big| > C/3\Big) \\
= \ & P\Big(\Big|\frac{\tilde{T}_1}{T_1 + T_2^* P(A_2) + \sigma^2 P(A_2)}\Big| > C/3\Big) \\
& + P\Big(\Big|\frac{\tilde{T}_2^* P(A_2)}{T_1 + T_2^* P(A_2) + \sigma^2 P(A_2)}\Big| > C/3\Big) \\
& + P\Big(\Big|\frac{(T_1 + T_2^* P(A_2))(1 + VI_A(j))}{T_1 + T_2^* P(A_2) + \sigma^2 P(A_2)}\Big| > C/3\Big).
\end{aligned} \tag{9}$$

Noting that all the T terms are positive, and $VI_A(j)$ is also positive and bounded above, we have:

$$\begin{aligned}
& P\Big(\Big|\widehat{VI}_A(j) - VI_A(j)\Big| > C\Big) \\
\leq \ & P\Big(\Big|\frac{\tilde{T}_1}{\sigma^2 P(A_2)}\Big| > C/3\Big) + P\Big(\Big|\frac{\tilde{T}_2^* P(A_2)}{\sigma^2 P(A_2)}\Big| > C/3\Big) + \\
& P\Big(\Big|\frac{T_1(1 + VI_A(j))}{\sigma^2 P(A_2)}\Big| > C/6\Big) + P\Big(\Big|\frac{T_2^* P(A_2)(1 + VI_A(j))}{\sigma^2 P(A_2)}\Big| > C/6\Big) \\
\leq \ & e^{-C \cdot P(A_2) \cdot n/3K_1} + e^{-C \cdot n_A^{\eta(p_A)}/3K_2} + e^{-C \cdot P(A_2) \cdot n/3K_3} + e^{-C \cdot n_A^{\eta(p_A)}/3K_4} \\
\leq \ & e^{-C \cdot n_A^{\eta(p_A)}/K_5}.
\end{aligned} \tag{10}$$

Noting that this is the tail probability for $\widehat{VI}_A(j)$ when p_A variables are considered in the embedded model, we can easily generalize it to the situation at an internal node where only p_0 variables are considered. In this case, we replace $\eta(p)$ by $\eta(p_0)$, yielding a faster convergence rate. In the derivation, the constant K_5 can possibly depend on p_A, however, since $p_A < p$, which is finite, we can always choose a larger K_5 such that the equation holds for all values of p_A. Consequently, K_5 does not depend on the choice of internal node A.

Now, two situations can arise for $VI_A(j)$:

Situation 1: $X^{(j)}$ is a noise variable. Since changing the value of $X^{(j)}$ will not change $f(X)$, $f(X^{(1)},...,\tilde{X}^{(j)},...,X^{(p)}) \equiv f(X^{(1)},...,X^{(j)},...,X^{(p)})$, and thus $VI_A(j) \equiv 0$.

Situation 2: $X^{(j)}$ is a strong variable. According to Assumption 3.2, $VI_A(j)$ is bounded below by $\psi_1(\delta) \cdot \psi_2(b_j - a_j)$, where $\delta = \min_{i \in \{\mathcal{S} \backslash j\}} (b_i - a_i)$. We further note that since the internal node size is n_A, the interval length of any variable is at least $\frac{n_A}{n}$ even if all splits are made on that variable. Hence both δ and $b_j - a_j$ are larger than $\frac{n_A}{n}$. Hence $VI_A(j) \geq \psi_1(\frac{n_A}{n}) \cdot \psi_2(\frac{n_A}{n})$ for any strong variable.

Hence, to sum up situations (1) and (2), we have

$$VI_A(j) \begin{cases} \geq \psi_1(\frac{n_A}{n}) \cdot \psi_2(\frac{n_A}{n}), & \text{if } j \in \mathcal{S}. \\ = 0, & \text{if } j \in \mathcal{S}^c. \end{cases} \tag{11}$$

Step 2: Now we prove a) of this Theorem. Let $\widehat{j}_A$ be the selected splitting variable at internal node A, i.e., $\widehat{j}_A = \arg\max_j VI_A(j)$. Without loss of generality, we assume that at this internal node A, the true variable importance measures are in the order $VI_A(1) \geq VI_A(2) \geq \cdots \geq VI_A(p_1) > VI_A(p_1 + 1) = \cdots = VI_A(p) = 0$. Then the probability that the selected splitting variable $\widehat{j}_A^*$ belongs to the set of strong variables satisfies the following inequality:

$$\begin{aligned} & P(\widehat{j}_A \in \mathcal{S}) \\ = \;& 1 - P(\widehat{j}_A \in \mathcal{S}^c) \\ = \;& 1 - \sum_{i \in \mathcal{S}^c} P(\widehat{j}_A = i) \\ \geq \;& 1 - \sum_{i \in \mathcal{S}^c} P\big(\widehat{VI}_A(i) > \widehat{VI}_A(j), \text{ for all } j \in \mathcal{S}\big) \\ \geq \;& 1 - p_1 \sum_{i \in \mathcal{S}^c} P(\widehat{VI}_A(i) > \widehat{VI}_A(p_1)). \end{aligned} \tag{12}$$

Let $\widehat{\Delta}_j = \widehat{VI}_A(j) - VI_A(j)$. Using equation (10) and noting that $VI_A(i) = 0$ for all $i \in \mathcal{S}^c$, the above probability can be bounded below by

$$\begin{aligned}
& P(\widehat{j}_A \in \mathcal{S}) \\
\geq\ & 1 - p_1 \sum_{i \in \mathcal{S}^c} P\big(\widehat{\Delta}_j + 0 > \widehat{\Delta}_{p_1} + VI_A(p_1)\big) \\
\geq\ & 1 - p_1 \sum_{i \in \mathcal{S}^c} \Big[P\big(|\widehat{\Delta}_{p_1}| > \frac{VI_A(p_1)}{2}\big) + P\big(\widehat{\Delta}_j > \frac{VI_A(p_1)}{2}\big)\Big] \\
=\ & 1 - p_1 \sum_{i \in \mathcal{S}^c} 4 \cdot e^{-\frac{VI_A(p_1)}{2} \cdot n_A^{\eta} / K_5} \\
=\ & 1 - 4 p_1 p_2 \cdot e^{-\frac{VI_A(p_1)}{2} \cdot n_A^{\eta} / K_5}. \qquad (13)
\end{aligned}$$

Using Equation (11), we have, for any internal node A with sample size n_A, and with p_A nonmuted variables,

$$P(\widehat{j}_A \in \mathcal{S}) \geq 1 - 4 p_1 p_2 \cdot e^{-\psi_1(\frac{n_A}{n}) \cdot \psi_2(\frac{n_A}{n}) \cdot n_A^{\eta(p_A)} / (K_5 \cdot 2)}. \qquad (14)$$

Since p_1, p_2 and K_5 are all constant, the proof for a) is concluded.

Step 3: We show b) using a similar structure as the proof of a). Note that at any internal node A, the probability that the maximum true variable importance is larger than twice that of the selected splitting variable is

$$P\Big(\max_j VI_A(j) > 2 VI_A\big(\widehat{j}_A\big)\Big).$$

By defining the variable with the true maximum variable importance at node A as $j_A^m = \arg\max_j VI_A(j)$, the above equation can be bounded by

$$\begin{aligned}
& P\big(VI_A(j_A^m) > 2 VI_A(\widehat{j}_A)\big) \\
\leq\ & P\Big(VI_A(j_A^m) > VI_A(\widehat{j}_A) + \psi_1(\frac{n_A}{n}) \cdot \psi_2(\frac{n_A}{n})\Big) \\
=\ & P\Big(VI_A(j_A^m) - \widehat{VI}_A(j_A^m) > VI_A(\widehat{j}_A) - \widehat{VI}_A(j_A^m) + \psi_1(\frac{n_A}{n}) \cdot \psi_2(\frac{n_A}{n})\Big) \\
=\ & P\Big(VI_A(j_A^m) - \widehat{VI}_A(j_A^m) > VI_A(\widehat{j}_A) - \widehat{VI}_A(\widehat{j}_A) \\
& \qquad + \widehat{VI}_A(\widehat{j}_A) - \widehat{VI}_A(j_A^m) + \psi_1(\frac{n_A}{n}) \cdot \psi_2(\frac{n_A}{n})\Big).
\end{aligned}$$

Note that $\widehat{VI}_A(\widehat{j}_A) - \widehat{VI}_A(j_A^m) \geq 0$ since $\widehat{j}_A$ is the selected variable. Adapting the notation of

$\widehat{\Delta}$ used in Step 2, we now have

$$\begin{aligned} & P\Big(VI_A(j_A^m) > 2VI_A\big(\widehat{j}_A\big)\Big) \\ \leq\ & P\Big(\widehat{\Delta}_{j_A^m} > \widehat{\Delta}_{\widehat{j}_A} + 0 + \psi_1(\frac{n_A}{n}) \cdot \psi_2(\frac{n_A}{n})\Big) \\ \leq\ & P\Big(|\widehat{\Delta}_{j_A^m}| > \frac{\psi_1(\frac{n_A}{n}) \cdot \psi_2(\frac{n_A}{n})}{2}\Big) \\ & +P\Big(|\widehat{\Delta}_{\widehat{j}_A}| > \frac{\psi_1(\frac{n_A}{n}) \cdot \psi_2(\frac{n_A}{n})}{2}\Big) \\ \leq\ & 4e^{-\psi_1(\frac{n_A}{n})\cdot\psi_2(\frac{n_A}{n})\cdot n_A^{\eta(p_A)}/(K_5\cdot 2)}. \end{aligned} \tag{15}$$

Thus the proof for b) is concluded.

Step 4: We now show c), that the protected set $\mathcal{P}_A^0$ for the entire tree contains all strong variables with probability close to 1, provided the number of protected variables p_0 is greater than p_1. It is sufficient to show this property at the root node, where $A = [0,1]^p$, since the protected set will only increase after a split. Note that when $p_0 > p_1$, if a strong variable is not in the protected set, there must be at least one noise variable with larger $\widehat{VI}$. Hence we have:

$$\begin{aligned} & P(\mathcal{S} \in \mathcal{P}_A^0) \\ \geq\ & 1 - P(\exists j \in \mathcal{S} \text{ and } i \in \mathcal{S}^c, s.t.\, \widehat{VI}_A(j) < \widehat{VI}_A(i)) \\ \geq\ & 1 - \sum_{j \in \mathcal{S}, i \in \mathcal{S}^c} P(\widehat{VI}_A(j) < \widehat{VI}_A(i)) \\ \geq\ & 1 - p_1 p_2 P(\widehat{VI}_A(p_1) < \widehat{VI}_A(p_1+1)). \end{aligned}$$

By similar arguments to those used in Steps 2), and noting that $n_A = n$ at the root node, we can bound the above probability by:

$$\begin{aligned} & P(\mathcal{S} \in \mathcal{P}_A^0) \\ \geq\ & 1 - p_1 p_2 e^{VI_A(p_1)\cdot n^{\eta(p)}/(K_5\cdot 2)}. \end{aligned} \tag{16}$$

Since at the root node, all the variable importance measures, including $VI_A(p_1)$, are fixed constants, the proof for c) is concluded. □

2. LEMMA 2.1

Lemma 2.1. *Let $\mathcal{A}_{nT}$ denote the set of terminal hypercubes. Under the same assumptions of Theorem 3.7, it holds that*

$$\max_{A\in\mathcal{A}_{nT}, j\in\mathcal{S}} VI_A(j) = O_p(n^{-2r\gamma log_q(1-q)/(rp_1)^{p_1+1}}),$$

where r is a constant satisfying $r > 1$ and $2(1-q)^{2r}/q^2 \leq 1$, p_1 is the number of strong variables, and q is the lower quantile for the random splitting point generation.

Proof. For any terminal hypercube $A \in \mathcal{A}_{nT}$, let $A_1 \to A_2 \to \ldots \to A_N = A$ be the constructed chain of the nodes leading to A, where A_{k+1} is the daughter node of A_k. Since at each node, the splitting point is chosen uniformly between the $100q$ and $100(1-q)$ quantiles of the current range of the splitting variable for some $q \in (0, \frac{1}{2})$, and since the terminal node is the last node having $\geq n^\gamma$ observations, it is easy to see that $-\gamma \log_q(n) \leq N \leq -\gamma \log_{(1-q)}(n)$. Let $j_k = \text{argmax}_{j\in\mathcal{S}} \widehat{VI}_{A_k}(j)$ be the index of the variable selected for splitting at node A_k and, moreover, define $m_j = \sum_{k=1}^N I(j_k = j)$, the number of times the jth variable is used for splitting. Let $N_j = \max\{k :, k = 1, ..., N, j_k = j\}$, the index of the last node split with the jth variable.

Before presenting the main proof, we state two simple properties:

Property 1. For $j \in \mathcal{S}$, $VI_{A_{N_j}}(j) \leq c_1(1-q)^{m_j}$. This is because after node A_{N_j}, the interval of the jth variable has been split m_j times so its length is at most $(1-q)^{m_j-1}$. Therefore, according to the proof of Theorem 3.6, $VI_{A_{N_j}}(j) \leq c_1(1-q)^{2m_j}$.

Property 2. For $k = 1, ..., N-1$ and any $j \in \mathcal{S}$, $V_{A_{k+1}}(j) \leq 2VI_{A_k}(j_k)/q^2$. That is, the importance of any variable in the daughter node is no larger than the importance of the selected variable at the current node by a factor of $2/q^2$. This follows from Theorem 3.6 (b): $2VI_{A_k}(j_k) \geq \max_j VI_{A_k}(j)$. On other hand, for any $j \in \mathcal{S}$, since $A_{k+1} \subset A_k$ and $|A_{k+1}| \geq |A_k|/q$, we have

$$\begin{aligned} VI_{A_k}(j) &= \frac{E\Big[\big(f(X^{(-j)}, X^{(j)}) - f(X^{(-j)}, \tilde{X}^{(j)})\big)^2 I(X \in A_k, \tilde{X} \in A_k)\Big]}{\sigma^2 P(X \in A_k)} \\ &\geq \frac{E\Big[\big(f(X^{(-j)}, X^{(j)}) - f(X^{(-j)}, \tilde{X}^{(j)})\big)^2 I(X \in A_{k+1}, \tilde{X} \in A_{k+1})\Big]/q}{\sigma^2 P(X \in A_{k+1})q} \\ &= VI_{A_{k+1}}(j)/q^2. \end{aligned}$$

Thus, $V_{A_{k+1}}(j) \le VI_{A_k}(j)/q^2 \le 2VI_{A_k}(j_k)/q^2$. With these two properties, we now proceed to prove the lemma. First, we define the following sequence:

$$N > \frac{N}{(rp_1)^1} > \cdots > \frac{N}{(rp_1)^{p_1}} > 0, \tag{17}$$

where r is a constant satisfying $r > 1$ and $2(1-q)^{2r}/q^2 = c \le 1$. Since $0 < q < 1/2$, r can always be properly chosen. Correspondingly, we obtain intervals $W_k = [N/(rp_1)^k, N/(rp_1)^{k-1})$ for $k = 1, ..., p_1$ and $W_{p_1+1} = [0, N/(rp_1)^{p_1})$. Recall the definition of m_j, the number of times the jth variable is selected for splitting. Since $\sum_{k=1}^{p_1} m_j = N$, there must be at least one j such that $m_j \ge N/(rp_1)$ and $m_j \in W_1$. Furthermore, since there are (p_1+1) intervals, there exists an integer $p_1 + 1 \ge k_0 \ge 2$ such that $m_j \notin W_{k_0}$ for any $j = 1, ..., p_1$. Hence, we can define two sets:

$$\mathcal{S}_1 = \{j : m_j \ge N/(rp_1)^{k_0-1}\}$$

and

$$\mathcal{S}_2 = \{j : m_j < N/(rp_1)^{k_0}\},$$

so that $\mathcal{S}_1 \ne \emptyset$ and $\mathcal{S}_1 \cup \mathcal{S}_2 = \{1, ..., p_1\}$.

Let j^* be the variable in $\mathcal{S}_1$ which is split last among all the variables in $\mathcal{S}_1$, and let N^* be the node index where this variable is split last. In other words, the variables selected in the nodes A_k for $k > N^*$ are all from $\mathcal{S}_2$. Then using Property 1, we have $VI_{A_{N^*}}(j^*) \le c_1(1-q)^{2m_{j^*}}$. Using the fact that $j^* \in \mathcal{S}_1$, we obtain

$$VI_{A_{N^*}}(j^*) \le c_2(1-q)^{2N/(rp_1)^{k_0-1}}.$$

Since all splitting variables after node A_{N^*} are from $\mathcal{S}_2$, and the number of distinct variables is at most $(p_1 - 1)$, and the number of possible splits after $A_{N^*} = N - N^*$ is no larger than $(p_1 - 1)N/(rp_1)^{k_0}$, we can build the relationship between $VI_{A_{N^*}}(j^*)$ and all other variable importance measures at the terminal node A. Hence we conclude: (a) if $N^* = N$, then

$$\begin{aligned} VI_A(j) &= VI_{A_N}(j) \le 2VI_{A_N}(j_N) = 2VI_{A_{N^*}}(j^*) \\ &\le 2c_1(1-q)^{2N/(rp_1)^{k_0-1}} 2c_1 \le (1-q)^{2N/(rp_1)^{p_1}}. \end{aligned}$$

(b) if $N^* < N$, then according to Property 2,

$$VI_A(j) = VI_{A_N}(j) \le (\frac{2}{q^2})^{N-N^*} VI_{A_N^*}(j^*) \le (\frac{2}{q^2})^{(p_1-1)N/(rp_1)^{k_0}} VI_{A_N^*}(j^*).$$

Thus,

$$\begin{aligned}
VI_A(j) &\leq \frac{2c_3}{(1-q)^2q^2}\left(\frac{2(1-q)^{2r}}{q^2}\right)^{(p_1-1)N/(rp_1)^{k_0}} (1-q)^{2rN/(rp_1)^{k_0}} \\
&\leq c_4(1-q)^{2rN/(rp_1)^{k_0}} \\
&\leq c_4(1-q)^{2rN/(rp_1)^{p_1+1}}, \qquad (18)
\end{aligned}$$

where c_4 is a constant depending only on p_1 and q, and where we used the fact that $2(1-q)^{2r}/q^2 < 1$. Finally, since $-\gamma \log_q(n) \leq N \leq -\gamma \log_{(1-q)}(n)$, we obtain

$$\begin{aligned}
\max_{j\in\mathcal{S}} VI_A(j) &\leq c_5(1-q)^{-2r\gamma \log_q(n)/(rp_1)^{p_1+1}} \\
&= c_5 n^{-2r\gamma log_q(1-q)/(rp_1)^{p_1+1}},
\end{aligned}$$

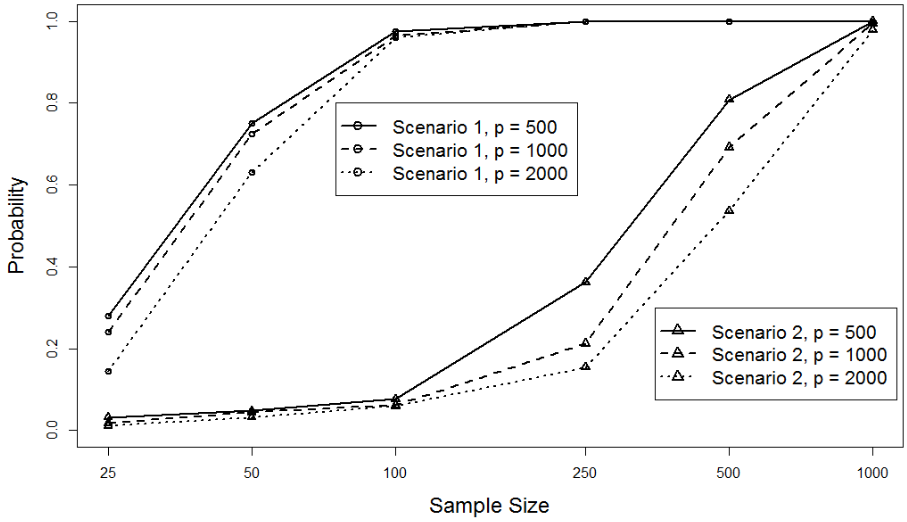
where c_5 is a constant depending only on p_1, q and r. Thus the lemma holds. □

3. TOY EXAMPLE FOR THEOREM 3.6

This toy example serves as a numerical demonstration of Theorem 3.6. We show that as the sample size increases, the probability of using a strong (or the strongest) variable as the splitting rule approaches 1. In other words, the variable importance from an embedded model should behave well as the sample size increase. However, this probability should also depend on the size of p, and the complexity of the true model structure. Following the settings of our simulation study in section 4.2, we consider two scenarios, Scenario 1: $E(Y) = 0.5X^{(50)} + 0.5X^{(100)} + X^{(150)}$, and Scenario 2: $E(Y) = 2X^{(50)}X^{(100)}$. We consider $n = 25, 50, 100, 250, 500, 1000$, and $p = 500, 1000, 2000$. For each setting, we fit an embedded model to the generated data and record whether $X^{(150)}$ has the highest $\widehat{VI}$ in Scenario 1, and whether $X^{(50)}$ or $X^{(100)}$ has the highest $\widehat{VI}$ in Scenario 2. This is repeated 500 times and the probabilities are summarized in the following plot.

Since the monotone effects of Scenario 1 are easier to detect by trees, the higher selection probability of Scenario 1 is expected. When the sample size is 250 or larger, the strongest variable $X^{(150)}$ is almost always selected. For a checkerboard model, however, we need to increase the sample size to 1000 to almost guarantee a "correct" selection. For both scenarios, there is a large randomness when $n = 25$. This indicates that when approaching terminal nodes, the splitting variable selection process can behave like a random pick.

Figure 1: Probability of selecting the most important variable as the splitting rule



4. ADDITIONAL SIMULATION RESULTS

Table 1: Classification/prediction error (SD) of RLT under $\alpha = 0.01$, $p = 200$

Muting	Linear combination	Scenario 1	Scenario 2	Scenario 3	Scenario 4
No	1	14.8% (4.0%)	4.09 (1.00)	5.43 (0.75)	4.36 (0.52)
	2	16.9% (4.2%)	5.35 (1.27)	5.74 (0.61)	2.63 (0.39)
	5	22.4% (3.2%)	7.82 (1.22)	5.94 (0.62)	2.81 (0.44)
Moderate	1	11.8% (3.4%)	3.20 (0.84)	4.74 (0.74)	3.27 (0.39)
	2	12.5% (3.6%)	3.72 (1.02)	4.85 (0.72)	1.99 (0.29)
	5	18.1% (3.4%)	6.00 (1.24)	4.94 (0.72)	2.08 (0.32)
Aggressive	1	10.3% (3.2%)	2.79 (0.71)	4.87 (0.81)	3.23 (0.39)
	2	10.0% (3.4%)	2.76 (0.77)	4.90 (0.79)	1.75 (0.23)
	5	14.3% (3.6%)	4.59 (1.10)	4.83 (0.78)	1.71 (0.24)

Note that α does not affect the performance of 1 linear combination under any circumstances.

5. DATA ANALYSIS RESULTS

Table 2: Classification/prediction error (SD) of RLT under $\alpha = 0.01$, $p = 500$

Muting	Linear combination	Scenario 1	Scenario 2	Scenario 3	Scenario 4
	1	17.8% (4.0%)	4.93 (1.20)	6.96 (0.98)	4.89 (0.62)
No	2	20.7% (4.1%)	6.48 (1.35)	7.10 (0.86)	2.97 (0.42)
	5	24.8% (3.2%)	8.80 (1.10)	7.27 (0.82)	3.20 (0.47)
	1	14.9% (3.9%)	3.88 (1.11)	6.43 (1.08)	3.69 (0.47)
Moderate	2	16.9% (4.2%)	4.83 (1.33)	6.49 (0.98)	2.30 (0.32)
	5	21.7% (3.4%)	7.30 (1.23)	6.59 (0.95)	2.45 (0.35)
	1	12.8% (3.8%)	3.39 (1.04)	6.13 (1.09)	3.35 (0.44)
Aggressive	2	13.6% (4.2%)	3.62 (1.23)	6.11 (1.05)	1.90 (0.24)
	5	17.8% (3.8%)	6.03 (1.26)	6.18 (1.05)	1.93 (0.25)

Note that α does not affect the performance of 1 linear combination under any circumstances.

Table 3: Classification/prediction error (SD) of RLT under $\alpha = 0.01$, $p = 1000$

Muting	Linear combination	Scenario 1	Scenario 2	Scenario 3	Scenario 4
	1	18.8% (4.4%)	5.64 (1.51)	7.81 (1.07)	5.08 (0.60)
No	2	21.4% (3.8%)	7.37 (1.53)	7.83 (0.97)	3.07 (0.45)
	5	25.4% (2.5%)	9.29 (1.09)	8.03 (0.83)	3.26 (0.48)
	1	16.0% (5.0%)	4.50 (1.47)	7.48 (1.26)	3.81 (0.45)
Moderate	2	18.1% (4.7%)	5.81 (1.63)	7.51 (1.06)	2.40 (0.38)
	5	22.9% (3.1%)	8.15 (1.30)	7.62 (0.99)	2.54 (0.40)
	1	13.7% (4.9%)	4.01 (1.38)	7.20 (1.22)	3.36 (0.42)
Aggressive	2	14.7% (5.1%)	4.49 (1.52)	7.08 (1.14)	1.92 (0.28)
	5	19.2% (3.9%)	6.98 (1.45)	7.18 (1.15)	1.99 (0.33)

Note that α does not affect the performance of 1 linear combination under any circumstances.

Table 4: Data analysis results for 10 machine learning datasets

Dataset	RF-all	ET	BART	LASSO	Boosting	RLT-naive	RLT
Boston housing	16.67 (3.02)	18.15 (2.65)	23.21 (2.79)	31.55 (3.12)	21.81 (2.99)	18.85 (2.40)	16.79 (2.04)
parkinson	38.89% (2.04%)	38.20% (2.00%)	39.21% (2.03%)	41.96% (1.65%)	38.42% (1.72%)	38.43% (1.99%)	38.05% (2.04%)
sonar	0.25 (0.06)	0.22 (0.05)	0.25 (0.06)	0.30 (0.06)	0.20 (0.05)	0.24 (0.06)	0.23 (0.06)
white wine	0.64 (0.02)	0.63 (0.02)	0.65 (0.02)	0.69 (0.07)	0.65 (0.02)	0.63 (0.02)	0.62 (0.02)
red wine	0.49 (0.02)	0.48 (0.02)	0.49 (0.02)	0.50 (0.03)	0.48 (0.02)	0.47 (0.02)	0.46 (0.02)
parkinson-Oxford	11.40% (4.94%)	13.92% (4.83%)	16.24% (5.76%)	17.71% (5.51%)	13.31% (4.50%)	13.98% (4.92%)	9.08% (4.17%)
ozone	21.54 (1.91)	20.83 (1.87)	22.22 (1.94)	25.13 (2.56)	22.16 (1.79)	21.05 (1.68)	19.21 (1.80)
concrete	95.07 (18.20)	98.40 (11.88)	87.63 (8.01)	162.11 (17.22)	98.34 (10.65)	98.51 (10.57)	67.79 (8.80)
breastcancer	3.36% (0.63%)	3.04% (0.56%)	4.13% (0.89%)	4.42% (0.85%)	3.57% (0.69%)	3.71% (0.79%)	3.26% (0.67%)
auto MPG	9.15 (1.07)	10.96 (1.49)	10.41 (1.25)	13.40 (1.61)	9.89 (1.11)	10.89 (1.43)	8.48 (1.00)

For each simulation run of each dataset, a random training sample of size 150 is use and the remaining data are used as testing sample. Extra covariates are included to increase p to 500. The simulation repeats 200 times. RF-all represents the best performance among RF, RF-$\sqrt{p}$, and RF-$\log p$.

6. TUNING PARAMETERS

In the following table, we summarize the tuning parameters that are implemented in the current version of the "RLT" package (as of date Feb 9, 2015). The code is available at `https://sites.google.com/site/teazrq/software`. Based on our experiments in both simulation and real data analysis, we found that n_{min}, p_d and k significantly affect the performance. The default tunings are $n_{min} = n^{1/3}$, $p_d = 80\%$, and $k = 2$, and it is strongly encouraged to tune p_d and k. The package also incorporates tunings for the embedded tree model. The default tunings in the embedded model are $ntrees\_embed = 100$ (number of trees), $resample\_prob\_embed = 85\%$ (bootstrap sample rate), $mtry\_embed = 2/3 \cdot |\mathcal{P} \setminus \mathcal{P}_A^d|$ (number of splitting variables tested), and $nspliteach\_embed = 1$ (number of random splitting points). However, we did not find significant impact for tuning the embedded model, since they tend to provide stable ranking of the variables in our analysis.

Table 5: Default/suggested tuning parameters in "RLT" package

Parameter	notation in paper	value(s)	Description
$ntrees$	M	100	number of trees
$resample\_prob$	—	100%	bootstrap ratio for fitting each tree
$nmin$	n_{min}	$n^{1/3}$	minimum number of observations in a terminal node
$nspliteach$	—	1	number of random splitting point
$muting\_percent$	p_d	0, 50%, 80%	muting rate
$combsplit$	k	1 to 5	nonzero components in linear combination split
$combsplit\_th$	α	0.25	threshold in a linear combination
$protectVar$	p_0	$\log(p)$	number of protected variables

REFERENCES

van de Geer, S., and Lederer, J. (2011), "The Bernstein–Orlicz norm and deviation inequalities," *Probability Theory and Related Fields*, pp. 1–26.