Supplemental material to "Supervised neighborhoods for distributed nonparametric regression"

1 Proof of Theorem 1

We restate the assumptions required for Theorem 1 to hold.

Assumption 1. *The training data* $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ *are generated i.i.d. from a joint distribution that satisfies the following properties:*

$$\mathbf{x}_i \sim \text{Uniform}([0,1]^p) \tag{1}$$

$$y_i = g(\mathbf{x}_i) + \omega(\mathbf{x}_i)\epsilon_i \tag{2}$$

where ϵ_i *is independent of* $\mathbf{x}_i$*, the function* $\omega(\mathbf{x})$ *is bounded,* $E(\epsilon_i) = 0$*, and* $E(\epsilon_i^2) < \infty$*. The function* g *must be sufficiently well-behaved such that Assumption 4 can be satisfied. A minimal requirement is that* g *is continuous.*

Assumption 2. *The splits of the constituent regression trees of the random forest are calculated using a dataset that is independent of the training data.*

Assumption 3. *We require that*

$$\min_{\substack{\mathbf{x}\in[0,1]^p \\ j\in(1,\ldots,K)}} k_{\theta_j}(\mathbf{x}) \to \infty \tag{3}$$

i.e., the number of training points contained in each node of each tree of the random forest goes to infinity.

Assumption 4. *For each* $\mathbf{x} \in [0,1]^p$*, the trees are trained in such a way that*

$$\max_{i,j} \left[w(\mathbf{x}_i, \mathbf{x}, \theta_j) |g(\mathbf{x}_i) - g(\mathbf{x})| \right] \xrightarrow{p} 0 \tag{4}$$

i.e., that the cells containing $\mathbf{x}$ *shrink in such away that the maximal variation of the function* g *within a cell shrinks to* 0 *in probability.*

Assumption 5. *The data in each tree are sampled without replacement from the original training set, so that all training points occuring in a particular leaf node have the same weight.*

Theorem 1. *Under Assumptions 1 - 5, for all* $\mathbf{x} \in [0,1]^p$,

$$\hat{g}(\mathbf{x}) - g(\mathbf{x}) \xrightarrow{p} 0 \tag{5}$$

Proof. For convenience, we introduce a shorthand for the random forest weight of training point i, defining $w_i = w_{\mathrm{RF}}(\mathbf{x}_i, \mathbf{x})$. We substitute equation 2 into the definition of $\hat{g}(\mathbf{x})$ in equation 12 from the main text.

$$\hat{g}\left(\mathbf{x}\right) - g\left(\mathbf{x}\right) = \mathbf{U}\left(\mathbf{0}\right)^T \Sigma_{\mathbf{x}}^{-1} \left[\sum_{i=1}^{n} w_i \mathbf{U}\left(\mathbf{x}_i - \mathbf{x}\right)\left(g\left(\mathbf{x}_i\right) - g\left(\mathbf{x}\right) + \omega(\mathbf{x})\epsilon_i\right)\right] \tag{6}$$

Note that we have used the fact that

$$\mathbf{U}\left(\mathbf{0}\right)^T \Sigma_{\mathbf{x}}^{-1} \left[\sum_{i=1}^{n} w_i \mathbf{U}\left(\mathbf{x}_i - \mathbf{x}\right) g\left(\mathbf{x}\right)\right] = g(\mathbf{x}) \tag{7}$$

This is a well-known property of local linear estimators: that they reproduce linear functions (in this case, a constant). See, for example, Proposition 1.12 in [1].

We decompose $\hat{g}(\mathbf{x}) - g(\mathbf{x})$ into a bias-type term and a variance-type term, defining

$$b\left(\mathbf{x}\right) = \mathbf{U}\left(\mathbf{0}\right)^T \Sigma_{\mathbf{x}}^{-1} \left[\sum_{i=1}^{n} w_i \mathbf{U}\left(\mathbf{x}_i - \mathbf{x}\right)\left(g\left(\mathbf{x}_i\right) - g\left(\mathbf{x}\right)\right)\right] \tag{8}$$

$$v\left(\mathbf{x}\right) = \mathbf{U}\left(\mathbf{0}\right)^T \Sigma_{\mathbf{x}}^{-1} \left[\sum_{i=1}^{n} w_i \mathbf{U}\left(\mathbf{x}_i - \mathbf{x}\right)\omega(\mathbf{x}_i)\epsilon_i\right] \tag{9}$$

We will show that each of these terms converges to 0 in probability. □

Lemma 1. $v(\mathbf{x}) \xrightarrow{p} 0$

Proof. For notational convenience, we introduce a shorthand for the indicator that a training point $\mathbf{x}_i$ belongs to the same leaf node as $\mathbf{x}$ in a tree trained with random parameter θ: let $w_i(\theta) = w(\mathbf{x}_i, \mathbf{x}, \theta)$. We drop the dependence on $\mathbf{x}$ in the notation, because we will work with a fixed $\mathbf{x}$ for the duration of the proof. By Assumption 5, a particular training point will only occur once in a leaf node, so $w_i(\theta) \in \{0, 1\}$, and it both indicates the presence of i in the leaf node, and can be used to represent the weight of training point x_i. We define the bandwidth matrix of the random forest to be a diagonal matrix with diagonal elements set to be the largest component-wise distances from $\mathbf{x}$ to a training point that has nonzero weight. Let $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})$, and let $\mathbf{x} = (x_1, \ldots, x_p)$. We define

$$\begin{aligned} h_m &= \max_{i,j} \left[w_i(\theta_j)\left|x_{i,m} - x_m\right|\right] \\ \mathbf{H} &= \mathrm{diag}(1, h_1, \ldots, h_p) \end{aligned} \tag{10}$$

By Assumption 3, the number of training points falling in a leaf node goes to infinity. Using Assumption 2, if we condition on the variables $w_i(\theta_j)$, the subset

of the training data falling in $R(\mathbf{x}, \theta_j)$ is independent and identically distributed uniformly in the rectangle $R(\mathbf{x}, \theta_j)$. By definition, $\left\|\mathbf{H}^{-1}\mathbf{U}\left(\mathbf{x}_i - \mathbf{x}\right)\right\|_\infty < 1$, and we have assumed that $\omega(\mathbf{x})$ is bounded and $E(\epsilon_i) = 0$. We apply the weak law of large numbers (in fact, just a variance calculation) to obtain

$$\frac{1}{k_{\theta_j}(\mathbf{x})} \sum_{i=1}^{n} w_i\left(\theta_j\right) \mathbf{H}^{-1}\mathbf{U}\left(\mathbf{x}_i - \mathbf{x}\right) \omega(\mathbf{x}_i)\epsilon_i \xrightarrow{p} \mathbf{0} \tag{11}$$

Thus, averaging over trees, we have

$$\sum_{i=1}^{n} w_i \mathbf{H}^{-1}\mathbf{U}\left(\mathbf{x}_i - \mathbf{x}\right) \omega(\mathbf{x}_i)\epsilon_i \xrightarrow{p} \mathbf{0} \tag{12}$$

We now examine the covariance matrix $\Sigma_{\mathbf{x}}$.

$$\begin{aligned}
\Sigma_{\mathbf{x}} &= \sum_{i=1}^{n} w_i \mathbf{U}(\mathbf{x}_i - \mathbf{x})\mathbf{U}(\mathbf{x}_i - \mathbf{x})^T \\
&= \frac{1}{K} \sum_{j=1}^{K} \left[\frac{1}{k_{\theta_j}(\mathbf{x})} \sum_{i=1}^{n} w_i(\theta_j)\mathbf{U}(\mathbf{x}_i - \mathbf{x})\mathbf{U}(\mathbf{x}_i - \mathbf{x})^T \right]
\end{aligned}$$

We define the tree-level contributions to $\Sigma_{\mathbf{x}}$ as

$$\Sigma_{\mathbf{x}}(\theta_j) = \frac{1}{k_{\theta_j}(\mathbf{x})} \sum_{i=1}^{n} w_i(\theta_j)\mathbf{U}(\mathbf{x}_i - \mathbf{x})\mathbf{U}(\mathbf{x}_i - \mathbf{x})^T$$

We define $\boldsymbol{\delta}_i = w_i(\theta_j)(\mathbf{x}_i - \mathbf{x})$, and denote the components of $\boldsymbol{\delta}_i$ as $\boldsymbol{\delta}_i = (\delta_{i,1}, \ldots, \delta_{i,p})$. For convenience, we have dropped the dependence on j in the notation for $\boldsymbol{\delta}_i$, but it is to be understood that it is only nonzero for data falling in the leaf node of tree j. Then

$$\Sigma_{\mathbf{x}}\left(\theta_j\right) = \frac{1}{k_{\theta_j}\left(\mathbf{x}\right)} \begin{pmatrix} k_{\theta_j}(\mathbf{x}) & \sum_i \delta_{i,1} & \cdots & \sum_i \delta_{i,p} \\ \sum_i \delta_{i,1} & \sum_i \delta_{i,1}^2 & \cdots & \sum_i \delta_{i,1}\delta_{i,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i \delta_{i,p} & \sum_i \delta_{i,p}\delta_{i,1} & \cdots & \sum_i \delta_{i,p}^2 \end{pmatrix} \tag{13}$$

We define the width of the rectangle $R(\mathbf{x}, \theta_j)$ in each dimension as $\omega_{m,j}$, and define the offsets of $\mathbf{x}$ from the center of the rectangle $R(\mathbf{x}, \theta_j)$ in each dimension as $\Delta_{m,j}$, $m \in \{1, \ldots, p\}, j \in \{1, \ldots, K\}$. Then we have, by the weak law of large numbers, Assumption 2, Assumption 3, and the i.i.d. uniform distribution of $\mathbf{x}_i$,

$$\frac{1}{k_{\theta_j}(\mathbf{x})} \sum_i \delta_{i,m} + \Delta_{m,j} \xrightarrow{p} 0 \tag{14}$$

$$\frac{1}{k_{\theta_j}(\mathbf{x})} \sum_{i=1}^{n} \delta_{i,l}\delta_{i,m} - \Delta_{l,j}\Delta_{m,j} \xrightarrow{p} 0 \text{ for } l \neq m \tag{15}$$

$$\frac{1}{k_{\theta_j}(\mathbf{x})} \sum_i \delta_{i,m}^2 - \frac{1}{3}\left(3\Delta_{m,j}^2 + \frac{\omega_{m,j}^2}{4}\right) \xrightarrow{p} 0 \tag{16}$$

We define the tree-averaged quantities

$$\Delta_m = \frac{1}{K}\sum_{j=1}^{K}\Delta_{m,j} \tag{17}$$

$$\sigma_m^2 = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{3}\left(3\Delta_{m,j}^2 + \frac{\omega_{m,j}^2}{4}\right) \tag{18}$$

We define the vector $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_p)^T$, and define the matrix

$$S = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_p^2 \end{pmatrix} + \Delta\Delta^T \tag{19}$$

The above has shown that

$$\Sigma_{\mathbf{x}} - \begin{pmatrix} 1 & \Delta^T \\ \Delta & S \end{pmatrix} = o_p(1) \tag{20}$$

Then, we will apply the Woodbury formula and the formula for the inverse of a block-partitioned matrix to explicitly calculate the inverse of this matrix. We define

$$\eta = \sum_{m=1}^{p}\frac{\Delta_m^2}{\sigma_m^2} - \frac{1}{1+\sum_{j=1}^{p}\frac{\Delta_i^2}{\sigma_i^2}}\sum_{m=1}^{p}\sum_{l=1}^{p}\frac{\Delta_m^2\Delta_l^2}{\sigma_m^2\sigma_l^2} \tag{21}$$

Then

$$\mathbf{U}(\mathbf{0})^T\begin{pmatrix} 1 & \Delta^T \\ \Delta & S \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{1-\eta} \\ \frac{1}{1-\eta}\left[\frac{\Delta_1}{\sigma_1^2} - \frac{1}{1+\sum_{j=1}^{p}\frac{\Delta_i^2}{\sigma_i^2}}\sum_{m=1}^{p}\frac{\Delta_1\Delta_m^2}{\sigma_1^2\sigma_m^2}\right] \\ \vdots \\ \frac{1}{1-\eta}\left[\frac{\Delta_p}{\sigma_p^2} - \frac{1}{1+\sum_{j=1}^{p}\frac{\Delta_i^2}{\sigma_i^2}}\sum_{m=1}^{p}\frac{\Delta_p\Delta_m^2}{\sigma_p^2\sigma_m^2}\right] \end{pmatrix} \tag{22}$$

Since $\Delta_m^2 = O(\sigma_m^2)$, we see that this vector is

$$\mathbf{U}(\mathbf{0})^T\bar{\Sigma} = \begin{pmatrix} O(1) \\ O\left(\frac{1}{\sigma_1}\right) \\ \vdots \\ O\left(\frac{1}{\sigma_p}\right) \end{pmatrix} \tag{23}$$

Hence, because $h_m = O(\sigma_m)$,

$$\mathbf{U}(\mathbf{0})^T\Sigma_{\mathbf{x}}^{-1}\mathbf{H} = (O_p(1), \ldots, O_p(1)) \tag{24}$$

Thus, we have that $\mathbf{U}(\mathbf{0})^T\Sigma_{\mathbf{x}}^{-1}\sum_{i=1}^{n} w_i\mathbf{U}(\mathbf{x}_i - \mathbf{x}) = o_p(1)$ □

Lemma 2. $b(\mathbf{x}) \xrightarrow{p} 0$

Proof. We have, by definition,

$$b(\mathbf{x}) = \mathbf{U}(\mathbf{0})^T \Sigma_{\mathbf{x}}^{-1} \mathbf{H} \left[\sum_{i=1}^{n} w_i \mathbf{H}^{-1} \mathbf{U} (\mathbf{x}_i - \mathbf{x}) (g(\mathbf{x}_i) - g(\mathbf{x})) \right] \tag{25}$$

where $\mathbf{H}$ was defined in equation 10. By Assumption 4, and the definition of $\mathbf{H}$, we have that

$$\left[\sum_{i=1}^{n} w_i \mathbf{H}^{-1} \mathbf{U} (\mathbf{x}_i - \mathbf{x}) (g(\mathbf{x}_i) - g(\mathbf{x})) \right] = o_p(1) \tag{26}$$

As we have shown in Lemma 1, $\mathbf{U}(\mathbf{0})^T \Sigma_{\mathbf{x}}^{-1} \mathbf{H} = (O_p(1), \ldots, O_p(1))$, hence $b(\mathbf{x}) = o_p(1)$ □

2 Timing Results for Experiments in Spark

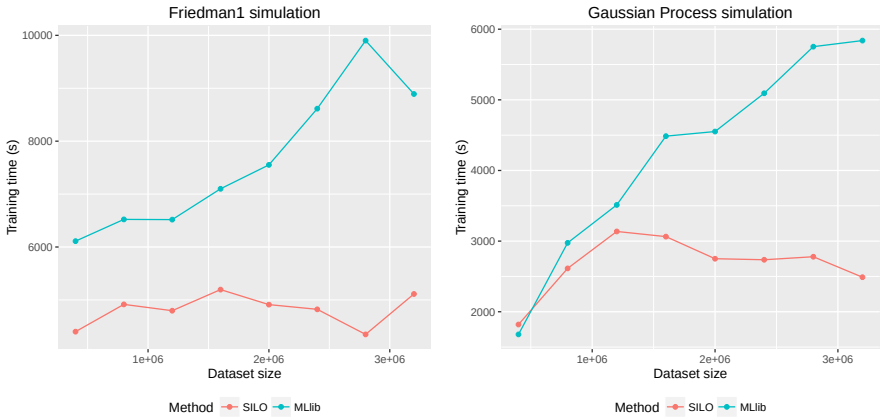

Figure 1: Training timing for Distributed-SILO and MLlib with growing dataset size. The amount of training data per Spark partition was fixed at 100,000 observations. Experiments were run on Amazon EC2 clusters using r3.xlarge instances, which have 4 processors and 30.5 GB of RAM per node. For each experiment, the cluster size was chosen such that the number of partitions was equal to the number of processors. As expected, training time of Distributed SILO is fairly constant as size of the dataset is increased, due to the lack of communication between workers. While MLlib's implementation avoids communication, particularly at deeper nodes in the trees, it does pay some communication penalty as more workers are added.

References

[1] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. Springer-Verlag New York, 2008.