Submitted to the Annals of Statistics

SUPPLEMENT TO "CONVERGENCE RATES OF OBLIQUE REGRESSION TREES FOR FLEXIBLE FUNCTION LIBRARIES"

By Matias D. Cattaneo[a*], Rajita Chandak[b], and Jason M. Klusowski[c†]

Department of Operations Research and Financial Engineering, Princeton University
[a]*cattaneo@princeton.edu*
[b]*rchandak@princeton.edu*
[c]*jason.klusowski@princeton.edu*

This supplement provides detailed proofs of theoretical results presented in the main paper.

APPENDIX A

Proof of Theorems 2.3 and 3.2. We begin by splitting the MSE (averaging only with respect to the joint distribution of $\{\mathcal{A}_t : t \in [T_k]\}$) into two terms, $\mathbb{E}_{T_k}[\|\mu - \widehat{\mu}(T_k)\|^2] = E_1 + E_2$, where

$$E_1 = \mathbb{E}_{T_k}[\|\mu - \widehat{\mu}(T_k)\|^2] - 2(\mathbb{E}_{T_K}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2] - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) - \alpha(n,k) - \beta(n)$$
$$E_2 = 2(\mathbb{E}_{T_k}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2] - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) + \alpha(n,k) + \beta(n),$$

and where $\alpha(n,k)$ and $\beta(n)$ are positive sequences that will be specified later.

To bound $\mathbb{E}[E_1]$, we split our analysis into two cases based on the observed data y_i. Accordingly, we have

$$\mathbb{E}[E_1] = \mathbb{E}[E_1 \mathbb{1}(\forall i : |y_i| \le B)] + \mathbb{E}[E_1 \mathbb{1}(\exists i : |y_i| > B)], \quad B \ge 0. \tag{A.1}$$

Bounded term. We start by looking at the first term on the right hand side of (A.1).

Proceeding, we introduce a few useful concepts and definitions for studying data-dependent partitions, due to Nobel [3]. Let

$$\Lambda_{n,k} = \{\mathcal{P}(\{(\tilde{y}_1, \tilde{\mathbf{x}}_1^{\mathrm{T}}), \ldots, (\tilde{y}_n, \tilde{\mathbf{x}}_n^{\mathrm{T}})\}) : (\tilde{y}_i, \tilde{\mathbf{x}}_i^{\mathrm{T}}) \in \mathbb{R}^{1+p}\}$$

be the family of all achievable partitions $\mathcal{P}$ by growing a depth k oblique decision tree on n data points with split boundaries of the form $\mathbf{x}^{\mathrm{T}}\mathbf{a} = b$, where $\|\mathbf{a}\|_{\ell_0} \le d$. In particular, note that $\Lambda_{n,k}$ contains all data-dependent partitions. We also define

$$M(\Lambda_{n,k}) = \max\{|\mathcal{P}| : \mathcal{P} \in \Lambda_{n,k}\}$$

to be the maximum number of terminal nodes among all partitions in $\Lambda_{n,k}$. Note that $M(\Lambda_{n,k}) \le 2^k$ (this statement does not rely on the specific algorithm used to grow a depth k oblique tree, as long as the tree generates a partition of $\mathcal{X}$ at each level). Given a set $\mathbf{z}^n = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\} \subset \mathbb{R}^p$, define $\Gamma(\mathbf{z}^n, \Lambda_{n,k})$ to be the number of distinct partitions of $\mathbf{z}^n$ induced by elements of $\Lambda_{n,k}$, that

*Financial support from the National Science Foundation (SES-2019432 and SES-2241575) is gratefully acknowledged.

†Financial support from the National Science Foundation (CAREER DMS-2239448, DMS-2054808, and HDR TRIPODS CCF-1934924) is gratefully acknowledged.

MSC2020 subject classifications: Primary 62G08; secondary 62L12.

Keywords and phrases: decision trees; neural networks; projection pursuit regression; CART; random forest.

is, the number of different partitions $\{\mathbf{z}^n \cap A : A \in \mathcal{P}\}$, for $\mathcal{P} \in \Lambda_{n,k}$. The partitioning number $\Gamma_{n,k}(\Lambda_{n,k})$ is defined by

$$\Gamma_{n,k}(\Lambda_{n,k}) = \max\{\Gamma(\mathbf{z}^n, \Lambda_{n,k}) : \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n \in \mathbb{R}^p\},$$

i.e., the maximum number of different partitions of any n point set that can be induced by members of $\Lambda_{n,k}$. Finally, let $\mathcal{F}_{n,k}(R)$ denote the collection of all functions (bounded by R) that output an element of span($\mathcal{H}$) on each region from a partition $\mathcal{P} \in \Lambda_{n,k}$.

We can deduce that the partitioning number is bounded by

$$\Gamma_{n,k}(\Lambda_{n,k}) \leq \left(\binom{p}{d} n^d\right)^{2^k} \leq \left(\left(\frac{ep}{d}\right)^d n^d\right)^{2^k} = \left(\frac{enp}{d}\right)^{d2^k}.$$

The bound on $\Gamma_{n,k}$ follows from the maximum number of ways in which n data points can be split by a hyperplane in d dimensions. The $\binom{p}{d}$ factor accounts for the number of ways in which a d-dimensional hyperplane can be constructed in a p-dimensional space. Note that this bound is not derived from the specific algorithm used to select the splitting hyperplanes; it is purely combinatorial.

Then, by slightly modifying the calculations in Györfi et al. [1, p. 240] and combining them with Györfi et al. [1, Lemma 13.1, Theorem 9.4], we have the following bound for the covering number $\mathcal{N}(r, \mathcal{F}_{n,k}(R), \mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n}))$ of $\mathcal{F}_{n,k}(R)$ by balls of radius $r > 0$ in $\mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n})$ with respect to the empirical discrete measure $\mathbb{P}_{\mathbf{x}^n}$ on $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^p$:

$$\begin{aligned}
\mathcal{N}\left(\frac{\beta(n)}{40R}, \mathcal{F}_{n,k}(R), \mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n})\right) &\leq \Gamma_{n,k}(\Lambda_{n,k}) \left(3\left(\frac{6eR}{\frac{\beta(n)}{40R}}\right)^{2\mathrm{VC}(\mathcal{H})}\right)^{2^k} \\
&\leq \left(\left(\frac{enp}{d}\right)^d\right)^{2^k} \left(3\left(\frac{240eR^2}{\beta(n)}\right)^{2\mathrm{VC}(\mathcal{H})}\right)^{2^k} \\
&= \left(3\left(\frac{enp}{d}\right)^d\right)^{2^k} \left(\frac{240eR^2}{\beta(n)}\right)^{\mathrm{VC}(\mathcal{H})2^{k+1}},
\end{aligned} \tag{A.2}$$

where we use $\mathrm{VC}(\mathcal{H})$ to denote the VC dimension of span($\mathcal{H}$). According to (2.6), we know that the regression function is uniformly bounded, $\|\mu\|_\infty \leq M'$. Let $R = QB$. We assume, without loss of generality, that $R \geq M'$ so that $\|\mu\|_\infty \leq R$ and $\|\widehat{\mu}(T_k)\|_\infty \leq R$ almost surely, if $\max_{1 \leq i \leq n} |y_i| \leq B$. By Györfi et al. [1, Theorem 11.4], with $\varepsilon = 1/2$ (in their notation),

$$\begin{aligned}
&\mathbb{P}\Big(\exists f \in \mathcal{F}_{n,k}(R) : \|\mu - f\|^2 \geq 2(\|\mathbf{y} - \mathbf{f}\|_n^2 - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) + \alpha(n,k) + \beta(n),\ \forall i : |y_i| \leq B\Big) \\
&\leq 14 \sup_{\mathbf{x}^n} \mathcal{N}\left(\frac{\beta(n)}{40R}, \mathcal{F}_{n,k}(R), \mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n})\right) \exp\left(-\frac{\alpha(n,k)n}{2568R^4}\right).
\end{aligned}$$

Then, we have the following probability concentration

$$\begin{aligned}
&\mathbb{P}(\mathbb{E}_{T_K}[\|\mu - \widehat{\mu}(T_k)\|] \geq 2(\mathbb{E}_{T_K}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2] - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) + \alpha(n,k) + \beta(n),\ \forall i : |y_i| \leq B) \\
&\leq 14 \sup_{\mathbf{x}^n} \mathcal{N}\left(\frac{\beta(n)}{40R}, \mathcal{F}_{n,k}(R), \mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n})\right) \exp\left(-\frac{\alpha(n,k)n}{2568R^4}\right).
\end{aligned} \tag{A.3}$$

This inequality follows from the fact that, on the event $\{\forall i : |y_i| \leq B\}$, if

$$\mathbb{E}_{T_K}[\|\mu - \widehat{\mu}(T_k)\|^2 - 2\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2] \geq -2\|\mathbf{y} - \boldsymbol{\mu}\|_n^2 + \alpha(n,k) + \beta(n)$$

holds, then there exists a realization $\widehat{\mu}(T_k') \in \mathcal{F}_{n,k}(R)$ such that

$$\|\mu - \widehat{\mu}(T_k')\|^2 - 2\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k')\|_n^2 \geq -2\|\mathbf{y} - \boldsymbol{\mu}\|_n^2 + \alpha(n,k) + \beta(n),$$

and hence

$$\mathbb{P}\Big(\mathbb{E}_{T_K}[\|\mu - \widehat{\mu}(T_k)\|] \geq 2(\mathbb{E}_{T_K}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2] - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) + \alpha(n,k) + \beta(n),\ \forall i : |y_i| \leq B\Big)$$
$$\leq \mathbb{P}\Big(\exists f \in \mathcal{F}_{n,k}(R) : \|\mu - f\|^2 \geq 2(\|\mathbf{y} - \mathbf{f}\|_n^2 - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) + \alpha(n,k) + \beta(n),\ \forall i : |y_i| \leq B\Big).$$

We can now plug in the result of (A.2) into (A.3) to obtain

$$\mathbb{P}(E_1 \geq 0,\ \forall i : |y_i| \leq B) \leq 14\left(3\left(\frac{enp}{d}\right)^d\right)^{2^k}\left(\frac{240eR^2}{\beta(n)}\right)^{\mathrm{VC}(\mathcal{H})2^{k+1}} \exp\left(-\frac{\alpha(n,k)n}{2568R^4}\right). \tag{A.4}$$

We choose

$$\alpha(n,k) = \frac{2568R^4\left(2^k d\log(enp/d) + 2^k\log(3) + \mathrm{VC}(\mathcal{H})2^{k+1}\log(\frac{240eR^2}{\beta(n)}) + \log(14n^2)\right)}{n}$$

$$\beta(n) = \frac{240eR^2}{n^2}$$

so that $\mathbb{P}(E_1 \geq 0,\ \forall i : |y_i| \leq B) \leq 1/n^2$. Thus,

$$E_1\mathbb{1}(\forall i : |y_i| \leq B) \leq (\mathbb{E}_{T_K}[\|\mu - \widehat{\mu}(T_k)\|^2] + 2\|\mathbf{y} - \boldsymbol{\mu}\|_n^2)\mathbb{1}(\forall i : |y_i| \leq B) \leq 12R^2,$$

and so we have

$$\mathbb{E}[E_1\mathbb{1}(\forall i : |y_i| \leq B)] \leq 12R^2\mathbb{P}(E_1 \geq 0,\ \forall i : |y_i| \leq B) \leq \frac{12R^2}{n^2} = \frac{12Q^2B^2}{n^2}. \tag{A.5}$$

Unbounded term. We now look at the second term on the right hand side of (A.1). Because we have $|\widehat{\mu}(T_k)\|_\infty \leq Q \cdot \sqrt{\max_{1\leq i\leq n}\frac{1}{i}\sum_{\ell=1}^i y_\ell^2}$ almost surely, we can bound

$$\mathbb{E}[\|\mu - \widehat{\mu}(T_k)\|^2\mathbb{1}(\exists i : |y_i| > B)] \leq (Q+1)^2\mathbb{E}\Big[\max_{1\leq i\leq n}\max\{y^2, y_i^2\}\mathbb{1}(\exists i : |y_i| > B)\Big].$$

Using the fact that the sum of non-negative variables upper bounds their maximum, and the exponential concentration of the conditional distribution of y given $\mathbf{x}$ (Assumption 2) together with a union bound, we can then apply Cauchy-Schwarz to obtain

$$\mathbb{E}[\|\mu - \widehat{\mu}(T_k)\|^2\mathbb{1}(\exists i : |y_i| > B)] \leq (Q+1)^2\sqrt{(n+1)\mathbb{E}[y^4]}\sqrt{nc_1\exp(-c_2(B-M)^\gamma)}.$$

Setting $B = B_n = M + ((6/c_2)\log(n+1))^{1/\gamma} \geq M'$, we have that

$$\mathbb{E}[\|\mu - \widehat{\mu}(T_k)\|^2\mathbb{1}(\exists i : |y_i| > B)] \leq \frac{(Q+1)^2\sqrt{c_1\mathbb{E}[y^4]}}{n^2}. \tag{A.6}$$

Thus combining (A.5) and (A.6), we have

$$\begin{aligned}\mathbb{E}[E_1] &= \mathbb{E}[E_1\mathbb{1}(\forall i : |y_i| \leq B)] + \mathbb{E}[E_1\mathbb{1}(\exists i : |y_i| > B)] \\ &\leq \frac{12Q^2B^2}{n^2} + \frac{(Q+1)^2\sqrt{c_1\mathbb{E}[y^4]}}{n^2} = O\left(\frac{\log^{2/\gamma}(n)}{n^2}\right).\end{aligned} \tag{A.7}$$

Next, we turn our attention to $\mathbb{E}[E_2]$. Since

$$\mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2 - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2] = \|\mu - g\|^2 + \mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2 - \|\mathbf{y} - \mathbf{g}\|_n^2],$$

it follows that

$$\mathbb{E}[E_2] = 2\|\mu - g\|^2 + 2\mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2 - \|\mathbf{y} - \mathbf{g}\|_n^2] + \alpha(n,k) + \beta(n). \tag{A.8}$$

Finally, combining the bounds (A.7) and (A.8) and simplifying $\alpha(n,k)$ and $\beta(n)$,

$$\begin{aligned}&\mathbb{E}[\|\mu - \widehat{\mu}(T_K)\|^2] \\ &\leq 2\|\mu - g\|^2 + 2\mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\|_n^2 - \|\mathbf{y} - \mathbf{g}\|_n^2] + C\frac{2^K(d + \mathrm{VC}(\mathcal{H}))\log(np/d)\log^{4/\gamma}(n)}{n},\end{aligned} \tag{A.9}$$

for some positive constant $C = C(c_1, c_2, \gamma, M, Q)$.

Pruned tree. We now consider the pruned tree, T_{opt}. Let $\mathbb{E}_{T_{\mathrm{opt}}}[\|\mu - \widehat{\mu}(T_{\mathrm{opt}})\|^2] = E_1' + E_2'$, where

$$\begin{aligned}E_1' &= \mathbb{E}_{T_{\mathrm{opt}}}[\|\mu - \widehat{\mu}(T_{\mathrm{opt}})\|^2] - 2(\mathbb{E}_{T_{\mathrm{opt}}}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_{\mathrm{opt}})\|_n^2] - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) - 2\lambda|T_{\mathrm{opt}}| \\ E_2' &= 2(\mathbb{E}_{T_{\mathrm{opt}}}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_{\mathrm{opt}})\|_n^2] - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) + 2\lambda|T_{\mathrm{opt}}|.\end{aligned}$$

Note that, for each $k = 1, 2, \ldots, n-1$,

$$\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_{\mathrm{opt}})\|_n^2 + \lambda|T_{\mathrm{opt}}| \leq \|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2 + \lambda 2^k,$$

and hence, for each $k \geq 1$,

$$\mathbb{E}[E_2'] \leq 2\|\mu - g\|^2 + 2\mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_k)\|_n^2 - \|\mathbf{y} - \mathbf{g}\|_n^2] + \lambda 2^{k+1}. \tag{A.10}$$

Choose $\lambda = \lambda_n$ such that $\alpha(n,k) + \beta(n) \leq \lambda_n 2^{k+1}$. This implies that $\lambda_n \gtrsim \frac{(d+\mathrm{VC}(\mathcal{H}))\log(np/d)\log^{4/\gamma}(n)}{n}$. For each realization of T_{opt}, there exists k such that $|T_{\mathrm{opt}}| \geq 2^k$. By a union bound and the result established in (A.4), we have

$$\begin{aligned}P(E_1' \geq 0) &\leq \mathbb{P}(\mathbb{E}_{T_{\mathrm{opt}}}[\|\mu - \widehat{\mu}(T_{\mathrm{opt}})\|^2] \geq 2(\mathbb{E}_{T_{\mathrm{opt}}}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_{\mathrm{opt}})\|_n^2] - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) + 2\lambda_n|T_{\mathrm{opt}}|) \\ &\leq \sum_{1\leq k\leq n-1} \mathbb{P}(\exists f \in \mathcal{F}_{n,k}(R) : \|\mu - f\|^2 \geq 2(\|\mathbf{y} - \mathbf{f}\|_n^2 - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) + \lambda_n 2^{k+1}) \\ &\leq \sum_{1\leq k\leq n-1} \mathbb{P}(\exists f \in \mathcal{F}_{n,k}(R) : \|\mu - f\|^2 \geq 2(\|\mathbf{y} - \mathbf{f}\|_n^2 - \|\mathbf{y} - \boldsymbol{\mu}\|_n^2) + \alpha(n,k) + \beta(n)) \\ &\leq \sum_{1\leq k\leq n-1} n^{-2} \leq 1/n.\end{aligned}$$

Once again, we split the expectation, $\mathbb{E}[E_1']$ into two cases, as in (A.1), and bound each case separately. The argument is identical to that for the un-pruned tree so we omit details here. Combining this bound on $\mathbb{E}[E_1']$ with (A.10) gives as an analogous result to (A.9), namely, for all $K \geq 1$,

$$\begin{aligned}&\mathbb{E}[\|\mu - \widehat{\mu}(T_{\mathrm{opt}})\|^2] \\ &\leq 2\|\mu - g\|^2 + 2\mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\|_n^2 - \|\mathbf{y} - \mathbf{g}\|_n^2] + C\frac{2^K(p + \mathrm{VC}(\mathcal{H}))\log^{1+4/\gamma}(n)}{n},\end{aligned} \tag{A.11}$$

for some positive constant $C = C(c_1, c_2, \gamma, M, Q)$.

The next part of the proof entails bounding $\mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\|_n^2 - \|\mathbf{y} - \mathbf{g}\|_n^2]$, depending on the assumptions we make. Note that for the constant output $\hat{y}_t(\mathbf{x}) \equiv \bar{y}_t$, we have $Q = 1$ and $\mathrm{VC}(\mathcal{H}) = 1$.

For Theorem 2.3: We bound $\mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\|_n^2 - \|\mathbf{y} - \mathbf{g}\|_n^2]$ using Lemma 2.2. The inequality (9) follows directly from (A.9) and the inequality (10) follows directly from (A.11).

For Theorem 3.2: Taking $g = \mu \in \mathcal{G}$ and $d = p$, we bound $\mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\|_n^2 - \|\mathbf{y} - \mathbf{g}\|_n^2]$ using Lemma 3.1. The inequality (3.2) follows directly from (A.9). To show (13), we use (3.2) and

$$\inf_{K \geq 1} \left\{ \frac{2AV^2}{4^{(K-1)/q}} + C \frac{2^{K+1} p \log^{4/\gamma+1}(n)}{n} \right\}$$
$$= 2(2+q) \left(\frac{AV^2}{q} \right)^{q/(2+q)} \left(\frac{Cp \log^{4/\gamma+1}(n)}{n} \right)^{2/(2+q)}.$$

This completes the proof of both Theorem 2.3 and Theorem 3.2. ∎

Proof of Corollary 2.4. Because our risk bounds allow for model misspecification, one can easily establish consistency of $\widehat{\mu}(T_K)$, even when $\mu \in \mathcal{F} \setminus \mathcal{G}$. Recall that $\mathcal{F} = \mathrm{cl}(\mathcal{G})$, that is,

$$\mathcal{F} = \left\{ f(\mathbf{x}) = \sum_{k=1}^{M} f_k(\mathbf{a}_k^{\mathrm{T}} \mathbf{x}),\ \mathbf{a}_k \in \mathbb{R}^p,\ f_k : \mathbb{R} \mapsto \mathbb{R} \right\}.$$

Importantly, $\mathcal{F}$ includes functions whose $\mathcal{L}_1$ norm may be infinite. Consider such a function μ that belongs to $\mathcal{F}$ but not to $\mathcal{G}$. Furthermore, grant Assumptions 1 and 2, which entail $\mu \in \mathscr{L}_\infty(\mathbb{R}^p)$. Let $\mathcal{G}' \subset \mathcal{G}$ denote the set of all single-hidden layer feed-forward neural networks with activation function that is non-constant and of bounded variation (and hence bounded). Then by Hornik [2, Theorem 1], $\mathcal{G}'$ is dense in $\mathscr{L}_\infty(\mathbb{R}^p) \subset \mathscr{L}_2(\mathbb{P}_\mathbf{x})$. Therefore, we can choose a sequence $\{g_n\} \subset \mathcal{G}'$, where each component function g_{nk} is bounded, non-constant, and of bounded variation, such that $\lim_{n\to\infty} \|\mu - g_n\| = 0$ and $\|g_n\|_{\mathcal{L}_1} < \infty$ for each n. Define a subsequence $\{g_{a_n}\}$ by $a_n = \max\left\{ m \leq n : \|g_m\|_{\mathcal{L}_1} \leq D\sqrt{K_n / \log(n+1)} \right\}$, where D is a positive constant large enough so that $\|g_1\|_{\mathcal{L}_1} \leq D\sqrt{K_n/\log(n+1)}$ for all n. Then, by construction, we have $\|\mu - g_{a_n}\| \to 0$ and $\|g_{a_n}\|_{\mathcal{L}_1} = o(\sqrt{K_n})$ as $n \to \infty$. Finally, according to (9) (and similarly (10)), since $\{g_{a_n}\} \subset \mathcal{F}$, we have $\lim_{n\to\infty} \mathbb{E}[\|\mu - \widehat{\mu}(T_K)\|^2] = 0$.

An analogous argument holds for the pruned tree T_{opt}. ∎

Proof of Corollary 2.5. The proof follows directly from the assumptions and Theorem 2.3. ∎

Proof of Theorem 4.1. Since we assume the subsample selection is independent of the splitting direction subset selection at each node, we have the following decomposition of the law of the process that governs each tree in the forest:

$$\Pi_\Theta = \Pi_K \times \Pi_\mathcal{I},$$

where $\mathcal{I} \subset \{1, \ldots, n\}$ is the set of indices of the subsampled data set of size N.

Part 1: Training error bound. By Jensen's inequality,

$$\mathbb{E}_{\Pi_\Theta}[\|\mu - \widehat{\mu}(\boldsymbol{\Theta})\|^2] \leq \mathbb{E}_{\Pi_\Theta}[\|\mu - \widehat{\mu}(T_K(\Theta))\|^2].$$

Additionally, by the law of total expectation,

$$\mathbb{E}_{\Pi_\Theta}[\|\mu - \widehat{\mu}(T_K(\Theta))\|^2] = \mathbb{E}_\mathcal{I}[\mathbb{E}_{\Pi_K}[\|\mu - \widehat{\mu}(T_K(\Theta))\|^2 \mid \mathcal{I}]].$$

We can prove a training error bound analogous to that of Lemma 3.1 by considering the modified definitions of excess training error. Define excess training error at each node conditional on the subsampled data as

$$R_K^{\mathcal{I}}(\mathrm{t}) = \|\mathbf{y} - \bar{y}_{\mathrm{t}}\|_{\mathrm{t},\mathcal{I}}^2 - \|\mathbf{y} - \mathbf{g}\|_{\mathrm{t}}^2,$$

and the excess training error of the tree as

$$R_K^{\mathcal{I}} = \|\mathbf{y} - \bar{y}\|_{\mathcal{I}}^2 - \|\mathbf{y} - \mathbf{g}\|_{\mathcal{I}}^2 .$$

Since we do the subset selection independently at each node, any terminal node t of T_{K-1} is independent of Π_K, conditional on Π_{K-1}. We can then apply the law of iterated expectation to the conditional training error, just as in the proof of Lemma 2.2 and the bound follows directly.

Part 2: Oracle inequality. The second part of this proof is analogous to the proof Theorem 2.3 where the averaging over the data set is replaced by averaging over the subsampled data.

This completes the proof. ■

A.1. Sedrakyan's Inequality. For completeness, we reproduce Sedrakyan's inequality [4] in its generalized form below.

LEMMA A.1 (Sedrakyan's inequality [4]). *Let U and V be two non-negative random variables with $V > 0$ almost surely. Then*

$$\mathbb{E}\left[\frac{U}{V}\right] \geq \frac{(\mathbb{E}[\sqrt{U}])^2}{\mathbb{E}[V]}.$$

PROOF OF LEMMA A.1. By the Cauchy-Schwarz inequality,

$$\mathbb{E}[\sqrt{U}] = \mathbb{E}\left[\sqrt{\frac{U}{V}}\sqrt{V}\right] \leq \sqrt{\mathbb{E}\left[\frac{U}{V}\right]}\sqrt{\mathbb{E}[V]}.$$

Rearranging the above inequality gives the desired result. ■

REFERENCES

[1] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression* **1**. Springer.

[2] HORNIK, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4** 251-257.

[3] NOBEL, A. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics* **24** 1084 – 1105.

[4] SEDRAKYAN, N. (1997). About the applications of one useful inequality. *Kvant Journal* **97** 42–44.