Submitted to the Annals of Applied Statistics

POSTERIOR CONCENTRATION FOR BAYESIAN REGRESSION TREES AND FORESTS

By Veronika Ročková and, Stéphanie van der Pas
Supplemental Material

University of Chicago *
Leiden University †

1. Proof of Theorem 6.1. We aim to establish conditions (2.1), (2.2) and (2.3) for $\varepsilon_n^2 = \sum_{t=1}^{T_0}(\varepsilon_n^t)^2$, where $\varepsilon_n^t = n^{-\alpha^t/(2\alpha^t+q_0^t)}\log^{\beta^t} n$. Our sieve consists of valid forests with either (a) many trees that are small (weak learners), or (b) a few large trees (strong learners). We impose a joint requirement on $\sum_{t=1}^T K^t$ so that the overall number of leaves in the ensemble is small. At the same time, we require that $\sum_{t=1}^T q^t$ (the upper bound on the number of active variables in the ensemble) is small as well. The sieve is constructed as follows:

$$\mathcal{F}_{\mathcal{E}}^n = \bigcup_{T=1}^{\infty} \bigcup_{\boldsymbol{q}:\sum_{t=1}^T q^t \le s_n} \bigcup_{\boldsymbol{\mathcal{S}}:|\mathcal{S}^t|=q^t} \bigcup_{\boldsymbol{K}:\sum_{t=1}^T K^t \le z_n} \mathcal{F}(\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}) \tag{1.1}$$

for some integer values s_n and z_n. Throughout this section we denote $\bar{K} = \frac{1}{T}\sum_{t=1}^T K^t$.

1.1. *Condition 2.1.* We first obtain the following upper bound on the log-covering number

$$\log N\left(\frac{\varepsilon}{36}, \{f_{\mathcal{E},\mathcal{B}} \in \mathcal{F}(\mathcal{E}) : \|f_{\mathcal{E},\mathcal{B}} - f_0\|_n < \varepsilon\}, \|\cdot\|_n\right) \lesssim (T \times \bar{K})\log(108\sqrt{n}) \tag{1.2}$$

where $\mathcal{F}(\mathcal{E}) = \left\{f_{\mathcal{E},\mathcal{B}} : [0,1]^p \to \mathbb{R} : f_{\mathcal{E},\mathcal{B}}(\boldsymbol{x}) = \sum_{t=1}^T f_{\mathcal{T}^t,\boldsymbol{\beta}^t}(\boldsymbol{x}); \boldsymbol{\beta}^t \in \mathbb{R}^{K^t}\right\}$ is the set of all additive step functions supported on a single partition ensemble $\mathcal{E} \in \mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}$. We denote by $\widetilde{\mathcal{T}}(\mathcal{E}) = \{\widetilde{\Omega}_k\}_{k=1}^{K(\mathcal{E})}$ the global partition associated with $\mathcal{E}$, consisting of $K(\mathcal{E})$ global cells. For $\mathcal{B}_1, \mathcal{B}_2 \in \mathbb{R}^{T\times\bar{K}}$, we denote by $f_{\mathcal{E},\mathcal{B}_1}, f_{\mathcal{E},\mathcal{B}_2} \in \mathcal{F}(\mathcal{E})$ two additive regression trees that sit on the same partition ensemble $\mathcal{E}$. Let $\bar{\boldsymbol{\beta}}_1 = \boldsymbol{A}(\mathcal{E})\mathcal{B}_1$ and $\bar{\boldsymbol{\beta}}_2 = \boldsymbol{A}(\mathcal{E})\mathcal{B}_2$ be the aggregated step sizes, as defined in (5.4), where $\boldsymbol{A}(\mathcal{E})$ is the stretching matrix. Then we can write

$$\frac{1}{n}||\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_2||_2^2 \le ||f_{\mathcal{E},\mathcal{B}_1} - f_{\mathcal{E},\mathcal{B}_2}||_n^2 = \sum_{k=1}^{K(\mathcal{E})} \mu(\widetilde{\Omega}_k)(\bar{\beta}_{1k} - \bar{\beta}_{2k})^2 \le ||\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_2||_2^2.$$

Deploying the singular value decomposition $\boldsymbol{A}(\mathcal{E}) = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ we write $\widetilde{\mathcal{B}}_1 = \boldsymbol{V}^T\mathcal{B}_1 \in \mathbb{R}^{\widetilde{K}}$ and $\widetilde{\mathcal{B}}_2 = \boldsymbol{V}^T\mathcal{B}_2 \in \mathbb{R}^{\widetilde{K}}$, where $\widetilde{K} \leq \min\{K(\mathcal{E}), T\bar{K}\}$. Using the fact that $\boldsymbol{U}$ is unitary, we have

$$\frac{1}{n}||\boldsymbol{D}(\tilde{\mathcal{B}}_1 - \tilde{\mathcal{B}}_2)||_2^2 \leq ||f_{\mathcal{E},\mathcal{B}_1} - f_{\mathcal{E},\mathcal{B}_2}||_n^2 \leq ||\boldsymbol{D}(\tilde{\mathcal{B}}_1 - \tilde{\mathcal{B}}_2)||_2^2.$$

We write $f_{\mathcal{E},\mathcal{B}_2}$ to be the projection of f_0 onto $\mathcal{F}(\mathcal{E})$ and note that $\{\mathcal{B} \in \mathbb{R}^{T\bar{K}} : \|f_{\mathcal{E},\mathcal{B}} - f_{\mathcal{E},\mathcal{B}_2}\|_n \leq \varepsilon\} \subset \{\widetilde{\mathcal{B}} : \|\boldsymbol{D}(\widetilde{\mathcal{B}} - \widetilde{\mathcal{B}}_2)\|_2 \leq \sqrt{n}\varepsilon\}$ and $\{\mathcal{B} \in \mathbb{R}^{T\bar{K}} : \|f_{\mathcal{E},\mathcal{B}} - f_{\mathcal{E},\mathcal{B}_2}\|_n \leq \varepsilon/36\} \supset \{\widetilde{\mathcal{B}} : \|\boldsymbol{D}(\widetilde{\mathcal{B}} - \widetilde{\mathcal{B}}_2)\|_2 \leq \varepsilon/36\}$. The covering number of $\{f_{\mathcal{E},\mathcal{B}} \in \mathcal{F}(\mathcal{E}) : \|f_{\mathcal{E},\mathcal{B}} - f_0\|_n \leq \varepsilon\}$ can be thus bounded from above by the minimal number of smaller ellipsoids $\{\widetilde{\mathcal{B}}_1 \in \mathbb{R}^{\widetilde{K}} : \|\boldsymbol{D}(\widetilde{\mathcal{B}}_1 - \widetilde{\mathcal{B}}_2)\|_2 \leq \varepsilon/36\}$ needed to cover a larger ellipsoid $\{\widetilde{\mathcal{B}}_1 \in \mathbb{R}^{\widetilde{K}} : \|\boldsymbol{D}(\widetilde{\mathcal{B}}_1 - \widetilde{\mathcal{B}}_2)\|_2 \leq \sqrt{n}\varepsilon\}$. Since these ellipsoids have *the same scaling factors* $\boldsymbol{D}$, this number is *the same* as the minimal number of little balls $\{\widetilde{\mathcal{B}}_1 \in \mathbb{R}^{\widetilde{K}} : \|\widetilde{\mathcal{B}}_1 - \widetilde{\mathcal{B}}_2\|_2 \leq \varepsilon/36\}$ needed to cover $\{\widetilde{\mathcal{B}}_1 \in \mathbb{R}^{\widetilde{K}} : \|\widetilde{\mathcal{B}}_1 - \widetilde{\mathcal{B}}_2\|_2 \leq \sqrt{n}\varepsilon\}$. The links between coverings of ellipsoids and balls can be found, for instance, in Dumer [1]. This altogether implies that the covering number is bounded by $(108\sqrt{n})^{T\bar{K}}$.

Now we find an upper bound on the number of valid ensembles $\mathcal{E} \in \mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}$ inside the sieve $\mathcal{F}_{\mathcal{E}}^n$. To start, we note that given $(T, \boldsymbol{q}, \boldsymbol{K}, \boldsymbol{\mathcal{S}})$, there are at most $\prod_{t=1}^T (K^t\, q^t\, n)^{K^t}$ valid ensembles $\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}$. This bound is obtained from Lemma 3.1 by combining all possible T-tuples of trees.[1] Given $(T, \boldsymbol{q})$, there are $\prod_{t=1}^T \binom{p}{q^t}$ sets of subsets $\boldsymbol{\mathcal{S}} = \{\mathcal{S}^1, \ldots, \mathcal{S}^T\}$ satisfying the constraint $|\mathcal{S}^t| = q^t$. This leads to an overall upper bound

$$\begin{aligned}
&\sum_{T=1}^{\min\{s_n, z_n\}} \sum_{\boldsymbol{K}:\sum_{t=1}^T K^t \leq z_n} \sum_{\boldsymbol{q}:\sum_{t=1}^T q^t \leq s_n} \prod_{t=1}^T \binom{p}{q^t} \left(K^t\, q^t\, n\right)^{K^t} \\
&\qquad < \sum_{T=1}^{\min\{s_n, z_n\}} \sum_{\boldsymbol{K}:\sum_{t=1}^T K^t \leq z_n} (z_n\, s_n\, n)^{z_n} \sum_{\boldsymbol{q}:\sum_{t=1}^T q^t \leq s_n} \prod_{t=1}^T \left(\frac{p\,\mathrm{e}}{q^t}\right)^{q^t} \\
&\qquad < s_n^{s_n+1} z_n^{z_n+1} \left(z_n\, s_n\, n\right)^{z_n} \left(p\,\mathrm{e}\right)^{s_n}.
\end{aligned}$$

Combining this bound with (1.2), we obtain the following bound

$$\begin{aligned}
&\log N\left(\tfrac{\varepsilon}{36}, \left\{f \in \mathcal{F}_{\mathcal{E}}^n : \|f - f_0\|_n < \varepsilon\right\}, \|.\|_n\right) < (s_n + 1)\log s_n + (z_n + 1)\log z_n \\
&\qquad + z_n \log\left(z_n\, s_n\, n\right) + s_n \log\left(p\,\mathrm{e}\right) + z_n \log\left(108\,\sqrt{n}\right).
\end{aligned} \tag{1.3}$$

[1] The order of trees in $\mathcal{E}$ matters.

Condition 2.1 will be met when (1.3) is smaller than (a constant multiple of) $n\varepsilon_n^2 = \sum_{t=1}^{T_0} n(\varepsilon_n^t)^2$. With the choice $z_n = \lfloor C_z n\varepsilon_n^2 / \log n \rfloor$ and $s_n = \lceil C_s n^{q_0/(2\alpha+q_0)} \log^{2\beta} n / \log(p \vee n) \rceil$, where C_s and C_z are large enough constants to be determined later, this condition is satisfied.

1.2. *Condition* (2.2). To establish Condition (2.2) for tree ensembles, we begin by finding a single additive tree that approximates well. We will heavily leverage our findings from Section 8.2, noting that the problem of approximating an additive function f_0 with a sum of trees can be decomposed into smaller problems of approximating each layer f_0^t separately.

Denote by a_n^t the smallest leaf size of a k-d tree (defined in Remark 3.1) needed to approximate f_0^t with an error smaller than $\varepsilon_n^t/2$, where $\varepsilon_n^t = n^{-\alpha^t/(2\alpha^t+q_0^t)} \log^{\beta^t} n$. Such a tree step function approximation exists according to Lemma 3.2 when $\mathcal{X}$ is $(M, \mathcal{S}_0^t)$-regular. We will denote this approximation with $f_{\widehat{\mathcal{T}}^t, \widehat{\boldsymbol{\beta}}^t}(\boldsymbol{x})$. Moreover, with $\boldsymbol{a}_n = (a_n^1, \ldots, a_n^{T_0})'$ we denote the vector of such minimal tree sizes, where each a_n^t satisfies (8.6) with q_0^t, α^t and ε_n^t. Next, we will denote by $\widehat{\mathcal{E}} = \{\widehat{\mathcal{T}}^1, \ldots, \widehat{\mathcal{T}}^{T_0}\}$ the approximating partition ensemble with step heights $\widehat{\mathcal{B}} = (\widehat{\boldsymbol{\beta}}^{1\prime}, \ldots, \widehat{\boldsymbol{\beta}}^{T_0\prime})'$. The individual tree approximations $f_{\widehat{\mathcal{T}}^t, \widehat{\boldsymbol{\beta}}^t}(\boldsymbol{x})$ are woven into an approximating forest $f_{\widehat{\mathcal{E}}, \widehat{\mathcal{B}}}(\boldsymbol{x}) = \sum_{t=1}^{T_0} f_{\widehat{\mathcal{T}}^t, \widehat{\boldsymbol{\beta}}^t}(\boldsymbol{x}) \in \mathcal{F}(\mathcal{V}\mathcal{E}_{\boldsymbol{\mathcal{S}}_0}^{\boldsymbol{a}_n})$, where $\boldsymbol{\mathcal{S}}_0 = \{\mathcal{S}_0^1, \ldots, \mathcal{S}_0^{T_0}\}$.

Arguing as in Section 8.2, the statement $\|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^t\|_2 < \frac{\varepsilon_n^t}{2}$ for all $1 \le t \le T_0$ implies $\|f_0^t - f_{\widehat{\mathcal{T}}^t, \boldsymbol{\beta}^t}\|_n < \varepsilon_n^t$ for all $1 \le t \le T_0$ and $\boldsymbol{\beta}^t \in \mathbb{R}^{a_n^t}$. This further implies

$$\|f_0 - f_{\widehat{\mathcal{E}}, \mathcal{B}}\|_n \le \sum_{t=1}^{T_0} \|f_0^t - f_{\widehat{\mathcal{T}}^t, \boldsymbol{\beta}^t}\|_n \le \sum_{t=1}^{T_0} \varepsilon_n^t \le \sqrt{T_0}\varepsilon_n,$$

for any $\mathcal{B} = (\boldsymbol{\beta}^{1\prime}, \ldots, \boldsymbol{\beta}^{T_0\prime})' \in \mathbb{R}^{T_0 \times \bar{\boldsymbol{a}}_n}$, where the final inequality is due to Cauchy-Schwarz and where $\bar{\boldsymbol{a}}_n = \frac{1}{T_0}\sum_{t=1}^{T_0} a_n^t$. Denote by $\mathcal{F}(\widehat{\mathcal{E}})$ the set of all additive trees supported on the partition ensemble $\widehat{\mathcal{E}}$. Then we can write

$$\begin{aligned}
&\Pi\left(f \in \mathcal{F}(\widehat{\mathcal{E}}) : \|f_0 - f\|_n \le \varepsilon_n\right) \\
&\quad \ge \Pi\left(\boldsymbol{\beta} \in \mathbb{R}^{\sum_{t=1}^{T_0} a_n^t} : \|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^t\|_2 \le \frac{\varepsilon_n^t}{2\sqrt{T_0}} \quad \text{for each } t = 1, \ldots, T_0\right) \\
&\quad = \prod_{t=1}^{T_0} \Pi\left(\boldsymbol{\beta}^t \in \mathbb{R}^{a_n^t} : \|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^t\|_2 \le \frac{\varepsilon_n^t}{2\sqrt{T_0}}\right).
\end{aligned}$$

Because we assumed $\beta_j^t \sim \mathcal{N}(0, 1/T)$, given T, we can directly use (8.9) to

lower-bound the above with

$$\prod_{t=1}^{T_0} \frac{2^{-a_n^t} \mathrm{e}^{-\|\widehat{\boldsymbol{\beta}}^t\|_2^2 - (\varepsilon_n^t)^2/8}}{\Gamma(\frac{a_n^t}{2}) \frac{a_n^t}{2}} \left(\frac{(\varepsilon_n^t)^2}{4}\right)^{\frac{a_n^t}{2}}. \tag{1.4}$$

Because each tree $\widehat{\mathcal{T}}^t$ is a k-d tree and is by definition balanced, we have $\|\widehat{\boldsymbol{\beta}}^t\|_n^2 \lesssim a_n^t \|f_0^t\|_\infty^2$. Now we can directly apply all our calculations from Section 8.2. In particular, using (1.4) and noting that $\Delta(\mathcal{V}\mathcal{E}_{\boldsymbol{\mathcal{S}}_0}^{\boldsymbol{a}_n}) < \prod_{t=1}^{T_0} \Delta(\mathcal{V}_{\mathcal{S}_0^t}^{a_n^t})$, we obtain

$$\begin{aligned}
\Pi\left(f \in \mathcal{F}_{\mathcal{E}} : \|f_0 - f\|_n \leq \varepsilon_n\right) &\geq \pi(T_0)\pi(\boldsymbol{q}_0 \mid T_0)\pi(\boldsymbol{a}_n \mid T_0)\pi(\boldsymbol{\mathcal{S}}_0 \mid T, \boldsymbol{q}_0) \times \\
&\quad \times \pi(\widehat{\mathcal{E}} \mid \boldsymbol{\mathcal{S}}_0, \boldsymbol{a}_n) \Pi\left(f \in \mathcal{F}(\widehat{\mathcal{E}}) : ||f_0 - f_{\widehat{\mathcal{E}},\mathcal{B}}||_n \leq \varepsilon_n\right) \\
&> \pi(T_0) \prod_{t=1}^{T_0} L(q_0^t, a_n^t, \mathcal{S}_0^t, \widehat{\boldsymbol{\beta}}^t, \varepsilon_n^t),
\end{aligned}$$

where $L(\cdot)$ was defined in (8.8). It follows from Section (8.2) that

$$-\log L(q_0^t, a_n^t, \mathcal{S}_0^t, \widehat{\boldsymbol{\beta}}^t, \varepsilon_n^t) \lesssim n(\varepsilon_n^t)^2$$

for each $1 \leq t \leq T_0$ when $q_0^t \lesssim \log^{\beta^t} n$, $\log p \lesssim \min_{1 \leq t \leq T_0} n^{q^t/(2\alpha^t + q_0^t)}$ and $T_0 \lesssim n$. The last condition $T_0 \lesssim n$ is needed under our prior $K^t \sim \text{Poisson}(\lambda/T)$ so that $-\log \pi(a_n) \lesssim a_n \log a_n + a_n \log T_0 \lesssim n(\varepsilon_n^t)^2 = n^{q_0^t/(2\alpha^t + q_0^t)} \log^{2\beta^t} n$ for $\beta^t \geq 1/2$. Putting all the pieces together, we obtain the following lower bound

$$\Pi\left(f \in \mathcal{F}_{\mathcal{E}} : \|f_0 - f\|_n \leq \varepsilon_n\right) \geq \pi(T_0) \mathrm{e}^{-d\, n \sum_{t=1}^{T_0} (\varepsilon_n^t)^2}$$

for some suitably large $d > 0$. The last requirement needed for Condition (2.2) to be satisfied is that $\pi(T_0) \geq \mathrm{e}^{-d\, n\varepsilon_n^2}$. Our prior $\pi(T) \propto \mathrm{e}^{-C_T T}$ safely satisfies this requirement.

1.3. *Condition* (2.3). The condition entails showing that $\Pi(\mathcal{F}_{\mathcal{E}} \backslash \mathcal{F}_{\mathcal{E}}^n) = o(\mathrm{e}^{-(d+2)\, n\varepsilon_n^2})$ for d deployed in the previous section. It suffices to show that

$$\left[\Pi\left((T, \boldsymbol{K}) : \sum_{t=1}^{T} K^t > z_n\right) + \Pi\left((T, \boldsymbol{q}) : \sum_{t=1}^{T} q^t > s_n\right)\right] \mathrm{e}^{(d+2)\, n\varepsilon_n^2} \to 0.$$

Because we assume $K^t \mid T \overset{iid}{\sim} \text{Poisson}(\lambda/T)$ for some $\lambda \in \mathbb{R}$ (according to our definition in (T4*)), we can apply a similar Chernoff bound as in (8.14).

Namely, we have for any $\gamma > 0$

$$\Pi\left(\sum_{t=1}^{T} K^t > z_n\right) = \sum_{T=1}^{\infty} \pi(T)\Pi\left(\sum_{t=1}^{T} K^t > z_n \mid T\right) \lesssim \mathrm{e}^{-\gamma(z_n+1)} \sum_{T=1}^{\infty} \pi(T)\left(\mathrm{e}^{\mathrm{e}^{\gamma}\lambda/T} - 1\right)^T \lesssim \mathrm{e}^{-\gamma(z_n+1)+\mathrm{e}^{\gamma}\lambda}.$$

With $z_n = \lfloor C_z n\varepsilon_n^2/\log n \rfloor \sim \sum_{t=1}^{T_0} n^{q_0^t/(2\alpha^t+q_0^t)} \log^{2\beta^t-1} n$ and $\gamma = \log z_n$, we have $\Pi\left(\sum_{t=1}^{T} K^t > z_n\right) \mathrm{e}^{d\, n\varepsilon_n^2} \to 0$ for a large enough constant $C_z > 0$. Next, with the independent product prior (T1*), the Chernoff bound gives

$$\Pi\left(\sum_{t=1}^{T} q^t > s_n \mid T\right) \leq \mathrm{e}^{-\gamma(s_n+1)} \prod_{t=1}^{T} \mathbb{E}\left[\mathrm{e}^{\gamma q^t}\right] \lesssim \mathrm{e}^{-\gamma(s_n+1)}\mathrm{e}^{-T\log[1-\mathrm{e}^{\gamma}/(cp^a)]}$$

for any $\gamma > 0$, where we used the fact

$$\mathbb{E}\left[\mathrm{e}^{\gamma q^t}\right] = \sum_{q=0}^{p} \left[\mathrm{e}^{\gamma}/(cp^a)\right]^q < \frac{1}{1-\mathrm{e}^{\gamma}/(cp^a)}.$$

With $\gamma = \log(p \vee n)$ and $a > 2$, we can write

$$\Pi\left(\sum_{t=1}^{T} q^t > s_n \mid T\right) \lesssim \mathrm{e}^{-(s_n+1)\,\log(p\vee n)}\mathrm{e}^{-T\log\left[1-1/(cp^{a-1})\right]} < \mathrm{e}^{-(s_n+1)\log(p\vee n)-T\log[1-1/(cp)]}.$$

Next, we have

$$\Pi\left(\sum_{t=1}^{T} q^t > s_n\right) \lesssim \mathrm{e}^{-(s_n+1)\log(p\vee n)} \sum_{T=1}^{\infty} \pi(T)\mathrm{e}^{T\log[1+1/(cp-1)]}.$$

With $\pi(T) \propto \mathrm{e}^{-C_T T}$, where $C_T > \log 2$, we have

$$\Pi\left(\sum_{t=1}^{T} q^t > s_n\right) \lesssim \mathrm{e}^{-(s_n+1)\log(p\vee n)} \sum_{T=1}^{\infty} \mathrm{e}^{-T(C_T-\log 2)} \lesssim \mathrm{e}^{-(s_n+1)\log(p\vee n)}.$$

With $s_n = \lfloor C_s n\varepsilon_n^2/\log(p \vee n) \rfloor$ we have $\Pi\left(\sum_{t=1}^{T} q^t > s_n\right) \mathrm{e}^{(d+2)\, n\varepsilon_n^2} \to 0$ for a large enough constant C_q.

2. Proof of Theorem 5.1. The sieve will be very similar to (1.1). The only difference is that each tree in the ensemble is now constrained to depend on the same set of active variables $\mathcal{S}$. To mark this difference, we have denoted the partition ensembles with $\mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}}$ instead of $\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}$. Throughout this section, we use the following sieve:

$$\mathcal{F}_{\mathcal{E}}^{n} = \bigcup_{T=1}^{\infty} \bigcup_{q=0}^{q_n} \bigcup_{\mathcal{S}:|\mathcal{S}|=q} \bigcup_{\boldsymbol{K}:\sum_{t=1}^{T} K^t \leq z_n} \mathcal{F}(\mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}}). \tag{2.1}$$

2.1. *Condition 2.1.* Our sieve (2.1) is embedded in (1.1), where the number of ensembles $\mathcal{E}$ inside $\mathcal{F}_{\mathcal{E}}^{n}$ is now upper-bounded by

$$\sum_{T=1}^{z_n} \sum_{q=0}^{q_n} \sum_{\boldsymbol{K}:\sum_{t=1}^{T} K^t \leq z_n} \binom{p}{q} \prod_{t=1}^{T} \left(K^t\, q\, n\right)^{K^t} < z_n^{z_n+1} (q_n+1)\,(z_n\, q_n\, n)^{z_n}\,(p\,\mathrm{e})^{q_n}\,.$$

Using the same arguments as in Section 1.1, we choose $z_n = \lfloor C_z n \varepsilon_n^2 / \log n \rfloor$ and $q_n = \lceil C_q \min\{p, n^{q_0/(2\alpha+q_0)} \log^{2\beta} n / \log(p \vee n)\} \rceil$, for which the condition holds.

2.2. *Condition (2.2).* The key ingredient for establishing Condition (2.2) is the following lemma on the existence of a tree ensemble that approximates f_0 well.

LEMMA 2.1. *Assume $f \in \mathcal{H}_p^{\alpha} \cap \mathcal{C}(\mathcal{S})$, where $|\mathcal{S}| = q$, and that $\mathcal{X}$ is $(M, \mathcal{S})$-regular. Then for any $s \in \mathbb{N}\backslash\{0\}$, there exists an additive tree function $f_{\widehat{\mathcal{E}},\widehat{\boldsymbol{\mathcal{B}}}} \in \mathcal{F}(\mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}})$ consisting of $T = 2^{s\,q-1}$ trees, each with $K^t = s\,q+1$ leaves, such that*

$$\|f - f_{\widehat{\mathcal{E}},\widehat{\boldsymbol{\mathcal{B}}}}\|_n \leq ||f||_{\mathcal{H}^{\alpha}} C\, M^{\alpha}\, q / K(\widehat{\mathcal{E}})^{\alpha/q}$$

for some $C > 0$, where $K(\widehat{\mathcal{E}}) = 2^{s\,q}$.

PROOF. From Lemma 3.2, we know that there exists a single tree function $f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}} \in \mathcal{F}(\mathcal{V}_{\mathcal{S}}^{K})$ with $\widehat{K} = 2^{s\,q}$ leaves which approximates well. We regard the full symmetric tree $\widehat{\mathcal{T}}$ as the global partition of the approximating partition ensemble $\widehat{\mathcal{E}}$, i.e. $\widetilde{\mathcal{T}}(\widehat{\mathcal{E}}) = \widehat{\mathcal{T}}$. Moreover, $\widehat{\boldsymbol{\beta}}$ is regarded as the vector of aggregated steps, i.e. $\bar{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$ (the definition of the aggregated steps is in (5.4)). The actual ensemble $\widehat{\mathcal{E}}$ is obtained from $\widehat{\mathcal{T}}$ by redistributing the cuts among T trees, each with $K^t = K$ leaves, in the following way. We take completely imbalanced trees that keep refining one cell until the resolution reaches the tree depth $\log_2 \widehat{K}$. These trees have $K^t = \log_2 \widehat{K} + 1$

leaves and we need $T = \widehat{K}/2$ of those to sum up towards $\widehat{\mathcal{T}}$. This decomposition is illustrated in Figure 3, where a full symmetric tree $\widehat{\mathcal{T}}$ (Figure 4) with $\widehat{K} = 8$ leaves has been trimmed into $T = \widehat{K}/2 = 4$ smaller trees with $K^t = \log_2 \widehat{K} + 1 = 4$ leaves. The decomposition yields a tree ensemble $\widehat{\mathcal{E}} = \{\mathcal{T}^1, \ldots, \mathcal{T}^T\}$ with a stretching matrix $\boldsymbol{A}(\widehat{\mathcal{E}}) = [\mathrm{I}_{\widehat{K}}, \boldsymbol{A}_1]$ (after a suitable permutation of columns), where $\boldsymbol{A}_1$ is some binary matrix. It follows from Lemma 1(g) of Govaerts and Pryce [2] that $\lambda^2_{min}(\widehat{\mathcal{E}}) = 1 + \sigma^2_{min}(\boldsymbol{A}_1) \geq 1$, where $\sigma_{min}(\boldsymbol{A}_1)$ denotes the smallest singular value of $\boldsymbol{A}_1$. Moreover, we have $K(\widehat{\mathcal{E}}) = \widehat{K}$. Finally, we use (5.4) to obtain the individual tree steps via $\widehat{\mathcal{B}} = (\boldsymbol{A}(\widehat{\mathcal{E}})'\boldsymbol{A}(\widehat{\mathcal{E}}))^\dagger \boldsymbol{A}(\widehat{\mathcal{E}})'\bar{\boldsymbol{\beta}}$, where $\boldsymbol{A}^\dagger$ denotes the Moore pseudoinverse of $\boldsymbol{A}$. The rest follows from Lemma 3.2. □

Now we proceed with Condition 2.2. Denote by $\widehat{\mathcal{E}}$ the approximating ensemble from Lemma 2.1. Recall that the global partition $\widetilde{\mathcal{T}}(\widehat{\mathcal{E}}) = \{\widetilde{\Omega}_k\}_{k=1}^{K(\widehat{\mathcal{E}})}$ is a k-d tree, which is balanced in the sense that $C^2_{min}/K(\widehat{\mathcal{E}}) \leq \mu(\widetilde{\Omega}_k) \leq C^2_{max}/K(\widehat{\mathcal{E}})$ for some constants $C_{min} < 1 < C_{max}$ and $k = 1, \ldots, K(\widehat{\mathcal{E}})$. Next, we find the smallest $K(\widehat{\mathcal{E}})$ such that $||f||_{\mathcal{H}^\alpha} C\, M^\alpha\, q_0/K(\widehat{\mathcal{E}})^{\alpha/q_0} < \varepsilon_n/2$. This value will be denoted by a_n and it satisfies (8.6). Next, we denote by $\widehat{T} = a_n/2$ the number of approximating trees and by $\widehat{\boldsymbol{K}} = (\widehat{K}^1, \ldots, \widehat{K}^T)'$ the vector of leaves, where $\widehat{K}^t = \log_2 a_n + 1$ (again we are using the construction from Lemma 2.1). Then, using similar arguments as in Section 1.2 we can lower-bound $\Pi(f \in \mathcal{F}_{\mathcal{E}} : \|f - f_0\|_n \leq \varepsilon_n)$ with

$$\frac{\pi(\widehat{T})\pi(\widehat{\boldsymbol{K}} \mid \widehat{T})\pi(q_0)}{\left(\frac{e\,p}{q_0}\right)^{q_0} \prod_{t=1}^{\widehat{T}} (\widehat{K}^t\, q_0\, n)^{\widehat{K}^t}} \Pi\left(f \in \mathcal{F}(\widehat{\mathcal{E}}) : ||f - f_0||_n \leq \varepsilon_n\right), \tag{2.2}$$

where $\mathcal{F}(\widehat{\mathcal{E}})$ consists of all additive tree functions supported on $\widehat{\mathcal{E}}$. We denote by $\widetilde{a}_n = \sum_{t=1}^{\widehat{T}} \widehat{K}^t = a_n/2(\log_2 a_n + 1)$ and by $\widehat{\mathcal{B}} \in \mathbb{R}^{\widetilde{a}_n}$ the steps of the approximating additive trees from Lemma 2.1. Because $\mu(\widetilde{\Omega}_k) \leq C^2_{max}/K(\widehat{\mathcal{E}})$ we obtain for any arbitrary vector $\mathcal{B} \in \mathbb{R}^{\widetilde{a}_n}$ (similarly as in Section 1.1)

$$\|f_{\widehat{\mathcal{E}},\mathcal{B}} - f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}}\|_n \leq C_{max}\lambda_{max}(\widehat{\mathcal{E}})/\sqrt{K(\widehat{\mathcal{E}})}\|\mathcal{B} - \widehat{\mathcal{B}}\|_2$$

and thereby

$$\|\mathcal{B} - \widehat{\mathcal{B}}\|_2 \geq \sqrt{K(\widehat{\mathcal{E}})}/(C_{max}\lambda_{max}(\widehat{\mathcal{E}}))\big|\|f_0 - f_{\widehat{\mathcal{E}},\mathcal{B}}\|_n - \|f_0 - f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}}\|_n\big|.$$

Combined with the fact $\lambda^2_{\max}(\mathcal{E}) \leq K(\mathcal{E})\widetilde{a}_n$ (as shown in the proof of Lemma 4.1), the statement $\|\mathcal{B} - \widehat{\mathcal{B}}\|_2 < \frac{\varepsilon_n}{2}\frac{1}{C_{max}\sqrt{\widetilde{a}_n}}$ implies $\|f_0 - f_{\widehat{\mathcal{E}},\mathcal{B}}\|_n < \varepsilon_n$. Therefore we have

$$\Pi\left(f \in \mathcal{F}(\widehat{\mathcal{E}}) : \|f - f_0\|_n \leq \varepsilon_n\right) > \Pi\left(\mathcal{B} \in \mathbb{R}^{\widetilde{a}_n} : \|\mathcal{B} - \widehat{\mathcal{B}}\|_2 < \frac{\varepsilon_n}{2}\frac{1}{C_{max}\sqrt{\widetilde{a}_n}}\right).$$

Moreover, because $\mu(\widetilde{\Omega}_k) \geq C_{min}^2/K(\widehat{\mathcal{E}})$ for some $C_{min} < 1$, we have

$$\|\widehat{\mathcal{B}}\|_2 \leq \frac{\sqrt{a_n}}{C_{min}\lambda_{min}(\widehat{\mathcal{E}})}\|f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}}\|_n \leq \frac{\sqrt{a_n}}{C_{min}}\left(\frac{\varepsilon_n}{2} + \|f_0\|_\infty\right),$$

where we used the fact $\lambda_{min}^2(\widehat{\mathcal{E}}) \geq 1$ (proof of Lemma 2.1). Therefore we have $\|\widehat{\mathcal{B}}\|_2^2 \leq C_2 a_n \|f_0\|_\infty^2$ for some $C_2 > 0$. Following the calculations from Section 8.2 (namely (8.10)), we continue to lower-bound (2.2) with

(2.3)
$$\frac{\pi(\widehat{T})\pi(\widehat{\boldsymbol{K}} \mid \widehat{T})\pi(q_0)\mathrm{e}^{-\frac{\varepsilon_n^2}{8C_{max}^2\widetilde{a}_n} - a_n(C_2\|f_0\|_\infty^2 + \log 2)}}{\left(\frac{\mathrm{e}\,p}{q_0}\right)^{q_0}\left[(\log_2 a_n + 1)\, q_0\, n\right]^{\widetilde{a}_n}}\left(\frac{\varepsilon_n^2}{4C_{max}^2\widetilde{a}_n}\right)^{\frac{\widetilde{a}_n}{2}}\left(\frac{2}{\widetilde{a}_n}\right)^{\widetilde{a}_n/2+1}.$$

This quantity should be at least $\mathrm{e}^{-d\,n\varepsilon_n^2}$ for some suitably large $d > 0$. Now, with our prior (T4*) we can write

$$\pi(\widehat{\boldsymbol{K}} \mid \widehat{T}) \gtrsim \prod_{t=1}^{\widehat{T}} \mathrm{e}^{\widehat{K}^t \log(\lambda/\widehat{T}) - \widehat{K}^t \log \widehat{K}^t} \gtrsim \mathrm{e}^{-\widetilde{a}_n \log(\log_2 a_n + 1) + \widetilde{a}_n \log(2\lambda/a_n)}.$$

This quantity can be lower-bounded by $\mathrm{e}^{-D\,\widetilde{a}_n \log a_n}$ for some $D > 0$. Then we can write

$$\frac{\pi(\widehat{T})\pi(\widehat{\boldsymbol{K}} \mid \widehat{T})\pi(q_0)}{\left(\frac{\mathrm{e}\,p}{q_0}\right)^{q_0}\left[(\log_2 a_n + 1)\, q_0\, n\right]^{\widetilde{a}_n}} > \mathrm{e}^{-C_T a_n/2}\mathrm{e}^{-D\,\widetilde{a}_n \log a_n - q_0 \log(c\,\mathrm{e}\,p^{a+1}/q_0)}\mathrm{e}^{-\widetilde{a}_n \log(q_0 n(\log_2 a_n + 1))}.$$

By our assumptions $q_0 \lesssim \log^\beta n$ and $p \lesssim n^{q_0/(2\alpha+q_0)}$, the term $\mathrm{e}^{-q_0 \log(c\,\mathrm{e}\,p^{a+1}/q_0)}$ will safely be larger than $\mathrm{e}^{-d_1 n\varepsilon_n^2}$ for some $d_1 > 0$. We take all the remaining important terms in (2.3), aiming to show that (i) $\widetilde{a}_n \log(\widetilde{a}_n/\varepsilon_n^2)$, (ii) $a_n\|f_0\|_\infty^2$, (iii) $\widetilde{a}_n \log(q_0 n \log_2 a_n)$ and (iv) $\widetilde{a}_n \log \widetilde{a}_n$ are bounded by a constant multiple of $n\varepsilon_n^2$. From (8.6), we obtain $a_n \lesssim n^{q_0/(2\alpha+q_0)}$ under our assumption $2C_0 q_0 \lesssim \log^\beta n$. Then we can write

$$\widetilde{a}_n = a_n/2(\log_2 a_n + 1) \lesssim n^{q_0/(2\alpha+q_0)} \log n. \tag{2.4}$$

Using this bound, we verify that (i)-(iv) are bounded by a constant multiple of $n\varepsilon_n^2 = n^{\frac{q_0}{2\alpha+q_0}} \log^{2\beta} n$. First, note that

$$\widetilde{a}_n \log(\widetilde{a}_n/\varepsilon_n^2) \lesssim n^{\frac{q_0}{2\alpha+q_0}} \log n \log\left(n^{\frac{q_0}{2\alpha+q_0}} n^{\frac{2\alpha}{2\alpha+q_0}} \log^{1-2\beta} n\right).$$

This quantity is bounded by a multiple of $n^{\frac{q_0}{2\alpha+q_0}} \log^{2\beta} n$ when $\beta \geq 1$. Next, we can write $a_n \|f_0\|_\infty^2 \lesssim n^{\frac{q_0}{2\alpha+q_0}} \log^{2\beta} n$. Lastly, it follows from (2.4) that $\widetilde{a}_n \log \widetilde{a}_n \lesssim n^{q_0/(2\alpha+q_0)} \log^2 n$ and $\widetilde{a}_n \log(q_0 n \log_2 a_n) \lesssim n^{q_0/(2\alpha+q_0)} \log^2 n$. To sum up, there exists $d > 0$ such that $\Pi(f \in \mathcal{F}_\mathcal{E} : \|f - f_0\|_n \leq \varepsilon_n) > \mathrm{e}^{-d\, n\varepsilon_n^2}$ for $\beta \geq 1$.

2.3. *Condition* (2.3). The condition $\Pi(\mathcal{F}_\mathcal{E} \backslash \mathcal{F}_\mathcal{E}^n)\mathrm{e}^{(d+2)\, n\varepsilon_n^2} \to 0$ is verified similarly as in Section 1.3 and Section 8.3. For $\Pi(q > q_n)$, we use the bound from Section 8.3 with $q_n = \lceil C_q \min\{p, n^{q_0/(2\alpha+q_0)} \log^{2\beta} n / \log(p \vee n)\} \rceil$ and a large enough constant C_q. For $\Pi\left((T, \boldsymbol{K}) : \sum_{t=1}^T K^t > z_n\right)$ we use the bound from Section 1.3 with $z_n = \lfloor C_z n\varepsilon_n^2 / \log n \rfloor$ and a large enough constant C_z.

3. Proof of Lemma 3.2. We start with an auxiliary statement showing that when f is α-Hölder continuous, we can grow a step function on any given (tree) partition so that the approximation error will be governed by cell diameters.

Namely, for $f \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S})$ and a valid tree partition $\mathcal{T} = \{\Omega_k\}_{k=1}^K \in \mathcal{V}_\mathcal{S}^K$, there exists a step function $f_{\mathcal{T},\widehat{\boldsymbol{\beta}}}(\boldsymbol{x}) = \sum_{k=1}^K \widehat{\beta}_k \mathbb{I}_{\Omega_k}(\boldsymbol{x})$ such that $\|f - f_{\mathcal{T},\widehat{\boldsymbol{\beta}}}\|_n \leq ||f||_{\mathcal{H}^\alpha} \sqrt{\sum_{k=1}^K \mu(\Omega_k) \mathrm{diam}^{2\alpha}(\Omega_k; \mathcal{S})}$. Indeed, given $\mathcal{T}$ and design points $\mathcal{X}$, we take $\widehat{\beta}_k = \frac{1}{n_k} \sum_{i=1}^n f(\boldsymbol{x}_i) \mathbb{I}_{\Omega_k}(\boldsymbol{x}_i)$, where $n_k = \mu(\Omega_k) n$. Then for $\boldsymbol{x}_j \in \Omega_k \cap \mathcal{X}$ we have, from Hölder continuity,

$$
\begin{aligned}
|f(\boldsymbol{x}_j) - f_{\mathcal{T},\widehat{\boldsymbol{\beta}}}(\boldsymbol{x}_j)| &< \frac{1}{n_k} \sum_{\boldsymbol{x}_i \in \Omega_k} |f(\boldsymbol{x}_j) - f(\boldsymbol{x}_i)| \\
&\leq ||f||_{\mathcal{H}^\alpha} \frac{1}{n_k} \sum_{\boldsymbol{x}_i \in \Omega_k} ||\boldsymbol{x}_{j\mathcal{S}} - \boldsymbol{x}_{i\mathcal{S}}||_2^\alpha \leq ||f||_{\mathcal{H}^\alpha} \mathrm{diam}^\alpha(\Omega_k; \mathcal{S}).
\end{aligned}
$$

Then the approximation error satisfies

$$
||f - f_{\mathcal{T},\widehat{\boldsymbol{\beta}}}||_n \leq ||f||_{\mathcal{H}^\alpha} \sqrt{\sum_{k=1}^K \mu(\Omega_k) \mathrm{diam}^{2\alpha}(\Omega_k; \mathcal{S})}. \tag{3.1}
$$

To continue with the proof of (3.5), we grow a k-d tree partition $\widehat{\mathcal{T}} = \{\widehat{\Omega}_k\}_{k=1}^K$ (as explained in Remark 3.1) and construct an approximating step function $f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}$, as outlined above.

Using (3.1), the statement (3.5) then follows from

$$
||f - f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}||_n < C_{\max} ||f||_{\mathcal{H}^\alpha} \max_{1 \leq k \leq K} \mathrm{diam}^\alpha(\widehat{\Omega}_k; \mathcal{S}),
$$

where we used the fact that $\mu(\widehat{\Omega}_k) \leq C_{max}^2/K$ in k-d trees. A minor modification of the proof of Proposition 6 in [3] yields $\sum_{k=1}^{K} \mu(\widehat{\Omega}_k)\text{diam}(\widehat{\Omega}_k; \mathcal{S}) \leq \frac{q}{K^{1/q}}$. The rest follows from Definition 3.3. □

4. Auxiliary Result.

LEMMA 4.1. *Assume a valid ensemble $\mathcal{E}$ consisting of T trees, each with K^t leaves. Let $\lambda_{max}^2(\mathcal{E})$ be the largest eigenvalue of $\widetilde{\boldsymbol{A}}(\mathcal{E}) = \boldsymbol{A}(\mathcal{E})'\boldsymbol{A}(\mathcal{E})$. Then*

$$\lambda_{max}^2(\mathcal{E}) \leq K(\mathcal{E})(T \times \bar{K}), \tag{4.1}$$

where $\bar{K} = \frac{1}{T}\sum_{t=1}^{T} K^t$ and where $K(\mathcal{E})$ denotes the number of rows of $\boldsymbol{A}(\mathcal{E})$.

PROOF. By the Gershgorin circle theorem, all eigenvalues of $\widetilde{\boldsymbol{A}}(\mathcal{E}) = (\widetilde{a}_{ij})$ lie inside the union of intervals $[\widetilde{a}_{ii} - \sum_{j\neq i}\widetilde{a}_{ij}, \widetilde{a}_{ii} + \sum_{j\neq i}\widetilde{a}_{ij}]$ for $i = 1, \ldots, T\times \bar{K}$. As explained in Section 5.1, the diagonal and off-diagonal entries of $\widetilde{\boldsymbol{A}}(\mathcal{E})$ quantify the persistence and the overlap in terms of the number of intersecting global partitioning cells. The magnitude $|\widetilde{a}_{ij}|$ is no larger than $K(\mathcal{E})$ for each $1 \leq i, j \leq T\times\bar{K}$. The upper bound on the maximal eigenvalue $\lambda_{max}^2(\mathcal{E})$ is thus $K(\mathcal{E})(T \times \bar{K})$. □

References.

[1] I. Dumer. Covering an ellipsoid with equal balls. *Journal of Combinatorial Theory, Series A*, 113:1667–1676, 2006.

[2] W. Govaerts and J. Pryce. A singular value inequality for block matrices. *Linear algebra and its applications*, 125:141–145, 1989.

[3] N. Verma, S. Kpotufe, and S. Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 565–574. AUAI Press, 2009.