UNIFORM CONVERGENCE OF RANDOM FORESTS VIA ADAPTIVE CONCENTRATION

By Stefan Wager and Guenther Walther

Stanford University

We study the convergence of the predictive surface of regression trees and forests. To support our analysis we introduce a notion of *adaptive concentration*. This approach breaks tree training into a model selection phase in which we pick the tree splits, followed by a model fitting phase where we find the best regression model consistent with these splits; a similar formalism holds for forests. We show that the fitted tree or forest predictor concentrates around the optimal predictor with the same splits: as d and n_{obs} get large, the discrepancy is with high probability bounded on the order of $\sqrt{\log(d)\log(n_{obs})/k}$ uniformly over the whole regression surface, where d is the dimension of the feature space, n_{obs} is the number of training examples, and k is the minimum leaf size for each tree. We also provide rate-matching lower bounds for this adaptive concentration statement. From a practical perspective, our result implies that random forests should have stable predictive surfaces whenever the minimum leaf size k is reasonable. Thus, forests can be used for principled estimation and data visualization, and need not only be considered as black box predictors.

1. Introduction. Trees [10] and random forests [8] are among the most widely used machine learning predictors today, with applications in a broad variety of fields such as ecology [14, 33], genetics [17, 39], and remote sensing [20, 32]. While allowing for flexible predictive surfaces and complicated interactions, trees and especially random forests have proven to be surprisingly resilient to over-fitting. Unlike competing non-parametric techniques such as kernel methods or neural networks, random forests require very little tuning; experience has shown that we can often obtain good predictive models out-of-the-box with standard software like `randomForest` for `R` [27].

The empirical stability of random forests suggests that, beyond just using the forest for prediction at randomly drawn test examples, we should also be able to interpret the shape of the random forest predictive surface. However, from the perspective of available theoretical results, we have no reason to believe so. The best existing convergence results for random forests either only provide asymptotic consistency guarantees without rates of convergence [36], or assume a substantially simplified training procedure where tree splits are chosen without looking at the training data [4]. The goal of this paper is to introduce a new theoretical framework, *adaptive concentration*, for describing the statistical properties of forests. While our framework is weaker than the classical notion of pointwise consistency, it lets us provide tight bounds on the behavior of the forest predictive surface.

The idea of adaptive concentration is to view training trees (and similarly forests) as occurring in two stages: a model selection stage where we decide on which splits to make, and a model fitting stage where we find the best regression tree conditional on having made these splits. We then treat the splits made by the tree as fixed, and show that the fitted regression tree is not much worse than the optimal regression tree with the same splits; in other words, the fitted trees concentrate relative to the adaptively selected splitting scheme. Figure 1 illustrates this goal for a one-dimensional tree. The guarantee that the fitted tree is close to the optimal tree with the same splits means that any big jumps in the fitted tree correspond to some big changes in the underlying data-generating process, although the location and magnitude of the measured jump may have been affected by the position of the tree splits.

The motivation for our two-stage approach is closely related to the valid post-selection inference framework of Berk et al. [3], who provide convergence guarantees for estimated linear regression parameters that hold even if the regression model is selected after looking at the data. The main difficulty is in bounding the amount by which a tree can overfit to the training data during the splitting (or model selection) phase. The guarantees provided by our method have a similar practical flavor to those given by Berk et al. [3], in that both approaches protect against statistical instability without promising to recover the actual optimal conditional mean response function.

We study an asymptotic regime where the dimension d of the feature space, the number n_{obs} of training examples and the minimum leaf size k go to infinity together. Under mild regularity conditions, we show that regression trees and forest satisfy an adaptive convergence bound that scales as $\sqrt{\log(n_{obs}) \log(d)/k}$; more specifically, there is a universal constant C such that the distance between the fitted forest and the optimal forest with the same splits is bounded by $C\sqrt{\log(n_{obs}) \log(d)/k}$ with high probability over the whole sample space. We also show that this rate of convergence is tight to within a constant factor. This bound does not require any special modifications to the random forest training routine, and in particular holds for the CART algorithm [10] and the original proposal of Breiman [8]. The reason there is a dependence on n_{obs} in the numerator of our bound is that, as the sample size grows, the trees comprising the random forest can become

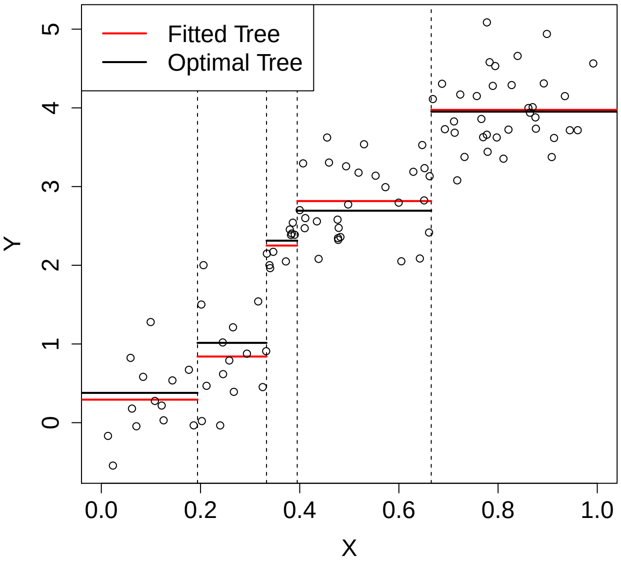

Fig 1: Adaptive concentration compares the prediction surface of the fitted decision tree with that of the optimal decision tree with the same splits. Here, the splits produced by recursive partitioning are denoted by dashed vertical lines. The regression tree was fit using the `R`-package `tree` [42].

deeper and so the model family becomes larger.

Our result lets us theoretically vindicate empirically-observed properties of random forests. First, it can help explain the observation that random forests adapt well to high-dimensional data: most generalization bounds for linear classifiers worsen linearly in d, whereas our result implies that the instability of random forests scales only logarithmically in d. Second, in a seeming disconnect between theory and practice, random forests have been long known to perform well with a small leaf-size k (e.g., `randomForest` uses a default value $k = 5$ regardless of d and n_{obs}), but standard theoretical analyses of random forests [28, 30] require k to grow with n_{obs} to achieve consistency. Our result lets us mediate between these two points of view by showing that, although k must in fact grow with n_{obs} to achieve good generalization, a logarithmically slow growth rate is enough.

1.1. *Related Work.* Ever since their introduction by Breiman [8], many authors have studied the theoretical properties of random forests in order to explain their performance [1, 4, 5, 9, 16, 28, 30, 35, 36, 37, 43]. In particular, Scornet et al. [36] prove that Breiman's original forests are L_2 consistent assuming that the conditional mean surface $\mathbb{E}\left[Y \,\middle|\, X = x\right]$ is an additive function, while Biau [4] and Denil et al. [16] discuss the properties of some random forests where trees are grown using only a development set, without looking at the training data used for predictions. To our knowledge, however, our adaptive concentration result is the first performance guarantee for random forests that provides strong local convergence rates for random forests as they are used in practice and holds in an asymptotic regime where n and d go to infinity jointly.

As discussed above, our analysis of random forests is motivated by work on post-selection inference following the worst-case approach of Berk et al. [3], who seek to provide a simultaneous convergence guarantee for every possible selected model. Another approach to post-selection inference is to fix the model-selection method and then explicitly condition on the selection event [18, 26, 29, 40]. However, most decision tree splitting rules are highly non-convex and difficult to describe analytically, so analyzing random forests from the perspective of conditional inference would require new technical tools.

Our proofs are built on top of results by Walther [45] on the asymptotics of the multidimensional scan statistic [24], which give us a handle on the concentration of the empirical process over decision tree leaves. Earlier results with a somewhat similar flavor use arguments in the style of Vapnik and Chervonenkis [41] and study the complexity of the class of fixed-depth decision trees to provide global generalization bounds [2, 10]. These methods, however, cannot say anything about the local properties of the predictive surface and do not extend to random forests; thus, they are not directly comparable to our analysis.

Notable extensions of the original random forest algorithm include online random forests [15, 34], random survival forests [23] and Bayesian alternatives to random forests [13]. It would be an interesting avenue for further work to see if our results can be extended to these methods. Finally, we emphasize that the main goal of this paper is to understand why random forests work, and not necessarily to provide generalization bounds that should be used to evaluate random forest classifiers in practice. Out-of-bag error [7], an adaptation of cross-validation, is a popular way estimating the practical accuracy of a random forest. Methods for doing statistical inference on random forest predictions are also available [31, 38, 43, 44].

2. Main Results.

To give adaptive concentration bounds, we first need to disambiguate the hierarchy of concepts used to build forest predictors: a *forest* is an ensemble of *trees*, each of which relies on a *partition* of the data generated by a splitting rule. We begin by providing formal definitions of these quantities below; we state our main results in Section 2.2. Throughout our analysis, we assume that we have a set of training examples $(x^{(i)},\, y^{(i)})$ of size $n_{obs} \sim \text{Poisson}\,(n)$ with features $x^{(i)} \in [0,\, 1]^d$ and bounded responses $y^{(i)} \in [-M/2,\, M/2]$.

2.1. *A Review of Recursive Partitioning.*

The first concept underlying a regression tree is the splitting rule itself, which induces a partition V of $[0,\, 1]^d$ into rectangles. We use the short-hand $V(x)$ to denote the unique element of V containing x. For our purposes, we are interested in those partitions that can be obtained by *recursive partitioning* of the feature space [10]. Starting from a parent node $L = [0,\, 1]^d$, recursive partitioning operates by repeatedly selecting a leaf $L \subseteq \mathbb{R}^d$ of the current working tree, a splitting variable $j \in \{1,\, ...,\, d\}$ and a threshold $\tau \in \mathbb{R}$, and then splitting L into two children $L_- = P \cap \{x : x_j \leq \tau\}$ and $L_+ = P \cap \{x : x_j > \tau\}$. Given our training set $\{(x^{(i)},\, y^{(i)})\}$, we require the partition to be valid in the sense of Definition 1.

Definition 1 (Valid partition). A partition V is $\{\alpha,\, k\}$-*valid* if it can by generated by generated by recursive partitioning such each child-node contains at least a fraction α of the data points in its parent-node for some $0 < \alpha < 0.5$, and each terminal node of V contains at least k training examples for some $k \in \mathbb{N}$. Given a dataset $\mathcal{X}$, we denote the set of $\{\alpha,\, k\}$-valid partitions by $\mathcal{V}_{\alpha,\, k}\,(\mathcal{X})$.

The constraint that each terminal node must have at least k observations is implemented by default in, e.g., `randomForest`; meanwhile, the requirement that each child node must incorporate at least a fraction α of the data in its parent is standard in theoretical analyses [30], and is usually satisfied in practice even if it is not explicitly enforced.

A partition V can then be used to induce a tree predictor by averaging the responses y_i over the leaves of V. In our adaptive concentration analysis, we consider two kinds of trees: valid trees that can be fitted to the training data, and partition-optimal trees that would arise if we had infinitely much data to train a tree supported on the partition V.

Definition 2 (Valid and partition-optimal trees). A valid partition induces a *valid tree*

$$T_V : [0,\, 1]^d \to \mathbb{R}, \quad T_V\,(x) = \frac{1}{\left|\left\{x^{(i)} : x^{(i)} \in V(x)\right\}\right|} \sum_{\left\{i : x^{(i)} \in V(x)\right\}} y^{(i)}. \tag{1}$$

We denote the set of all $\{\alpha,\, k\}$-valid trees T_V with $V \in \mathcal{V}_{\alpha,\, k}\,(\mathcal{X})$ by $\mathcal{T}_{\alpha,\, k}\,(\mathcal{X})$. Given a partition V, we also define the *partition-optimal tree* as

$$T_V^* : [0,\, 1]^d \to \mathbb{R}, \quad T_V^*\,(x) = \mathbb{E}\left[Y \middle| X \in V(x)\right]. \tag{2}$$

Forests are, as their name suggests, ensembles of regression trees. Generating a regression forest involves growing multiple trees; then, the forest prediction is the average of all the tree predictions. In general, the choice of splitting splitting variables j is randomized to ensure that the different trees comprising the forest are not too correlated with each others. As shown by Breiman [8], the variance reduction of a forest in comparison with its constituent trees improves as the correlation between individual trees decreases.

Definition 3 (Valid and partition-optimal forests). For any $B \in \mathbb{N}$, let $T_{V^{(1)}},\, ...,\, T_{V^{(B)}} \in \mathcal{T}_{\alpha,\, k}\,(\mathcal{X})$. Then, the average

$$H_{\{V\}_1^B} : [0,\, 1]^d \to \mathbb{R}, \quad H_{\{V\}_1^B}(x) = \frac{1}{B} \sum_{b=1}^{B} T_{V^{(b)}}(x) \tag{3}$$

is a *valid forest*; we denote the set of $\{\alpha,\, k\}$-valid forests by $\mathcal{H}_{\alpha,\, k}\,(\mathcal{X})$. The corresponding *partition-optimal forest* is defined as

$$H^*_{\{V\}_1^B} : [0,\, 1]^d \to \mathbb{R}, \quad H^*_{\{V\}_1^B}(x) = \frac{1}{B} \sum_{b=1}^{B} T^*_{V^{(b)}}(x). \tag{4}$$

When there is no risk of ambiguity, we write $H := H_{\{V\}_1^B}$ and $H^* = H^*_{\{V\}_1^B}$.

There are many proposals for how to choose the splitting variables j and the thresholds τ for trees. Our theoretical results, however, do not depend on the specific splitting rules used, and only rely on the generic structure of recursive partitioning; thus, we will not discuss specific splitting rules in this paper. For a review of how trees and forests are implemented in practice, we recommend Hastie et al. [21] (see Chapters 9.2 and 15).

Remark: Bootstrapping. There is one way in which our forests from Definition 3 differ from the original proposal of Breiman [8]: we do not allow the individual trees to be evaluated on bootstrap samples.[1] It seems plausible, however, that all our results should still hold even if we allow for bootstrapping, since the bootstrap is thought to have a regularizing effect on the forest and should thus reduce its ability to overfit the training data [6, 11]. Studying the effect of the bootstrap on our adaptive concentration bounds and perhaps showing how it can improve adaptive concentration guarantees presents a promising avenue for further work.

2.2. *Adaptive Concentration Bounds.* We are now ready to state our main results on the adaptive concentration of regression trees and forests. Unlike existing analyses that seek to establish the consistency of random forests [e.g., 5, 30, 36, 43], our results require no assumptions on the conditional mean function $\mathbb{E}\left[Y \,\middle|\, X = x\right]$. However, we do still need for the features x_i to be uniformly distributed on the unit cube. This assumption could be relaxed to a requirement that the distribution of X has a density f such that $f_{low} \leq f(x) \leq f_{high}$ for all $x \in [0,\, 1]^d$, for some fixed strictly positive constants f_{low} and f_{high}. In the interest of simplicity, however, we prove our results with Assumption 1.

Assumption 1 (Uniformly distributed features). The features X are uniformly distributed: $X \sim U\left([0,\, 1]^d\right)$.

Moreover, we need the minimum leaf-size k to grow at least logarithmically fast in n; this bound ensures that we can control the magnitude of discreteness effects.

Assumption 2 (Minimum leaf size). The minimum leaf-size k grows with n at a rate bounded from below by

$$\lim_{n \to \infty} \frac{\log{(n)}^2}{k} = 0,. \tag{5}$$

The principal difficulty in establishing adaptive concentration bounds is that the splits defining the constituent regression tree are chosen after seeing the training data. Thus, we cannot directly apply standard methods such as the Hoeffding bound to tree leaves, and must instead bound the amount by which the recursive partitioning process can overfit the training data.

[1] Technically, we could use a bootstrap sample to pick the partition V, but would then need to use the whole training set to turn V into a tree T_V.

The result following theorem is our main result on the adaptive concentration of decision trees. As explained in Section 6, the convergence rate given in (7) is tight up to constant factors.

Theorem 1. *Suppose that we have $n_{obs} \sim \text{Poisson}\,(n)$ training examples $(x_i,\, y_i) \in [0,\, 1]^d \times [-M/2,\, M/2]$ satisfying Assumption 1, and that we have a sequence of problems with parameters $(n,\, d,\, k)$ satisfying Assumption 2 and*[2]

$$\log\left(d\right) = \Theta\left(\log\left(n\right)\right). \tag{6}$$

Then, sample averages over all possible valid partitions concentrate around their expectations with high probability:

$$\lim_{n,\, d,\, k \to \infty} \mathbb{P}\left[\sup_{x \in [0,\, 1]^d,\, V \in \mathcal{V}_{\alpha,\, k}} \left|T_V\left(x\right) - T_V^*\left(x\right)\right| \right. \\ \left. \leq 6M \sqrt{\frac{\log\left(n\right)\log\left(d\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}\frac{1}{\sqrt{k}}}\right] = 1. \tag{7}$$

This result implies that, in practical data analysis, we can treat the fitted prediction function T_V as a good approximation to the optimal tree T_V^* supported on the partition V. As shown below, we can also use Theorem 1 to induce a generalization bound for valid regression trees.

Corollary 2. *Under the conditions of Theorem 1,*

$$\lim_{n,\, d,\, k \to \infty} \mathbb{P}\left[\sup_{T \in \mathcal{T}_{\alpha,\, k}} \frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} \left(y_i - T(x_i)\right)^2 - \mathbb{E}\left[\left(Y - T(X)\right)^2\right] \right. \\ \left. \leq 2M^2 \sqrt{\frac{\log\left(n\right)\log\left(d\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}\frac{1}{\sqrt{k}}}\right] = 1. \tag{8}$$

We also prove a related bound for random forests.

Theorem 3. *Under the conditions of Theorem 1, let $\{V_i\}_1^B$ be a set of valid partitions. Then, using notation from Definition 3, (7) holds with T_V*

[2] The condition (6) that d must scale polynomially in n is not strictly necessary, but allows us to ignore several second order terms; see Lemma 6 for details.

and T^*_V replaced with H and H^*. Moreover, the generalization error of H is bounded by

$$\lim_{n,\, d,\, k \to \infty} \mathbb{P}\Bigg[\sup_{H \in \mathcal{H}_{\alpha,\, k}} \frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} \left(y_i - H(x_i)\right)^2 - \mathbb{E}\left[\left(Y - H(X)\right)^2\right] \\ \leq 11M^2 \sqrt{\frac{\log\left(n\right) \log\left(d\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}} \frac{1}{\sqrt{k}} \Bigg] = 1. \tag{9}$$

Despite their deceptive similarity, (9) is a much stronger result than (8). The generalization bound from Corollary 2 does not require our full adaptive concentration machinery developed in Theorem 1: in fact, it is possible to prove an analogue to Corollary 2 directly by computing the Vapnik-Chervonenkis dimension of fixed-depth decision trees [2, 10]. On the other hand, the bound in (9) uses Theorem 1 to show that the correlation between different decision trees concentrates; we are not aware of any other methods for proving an analogue to (9).

2.3. *Discussion.* Our generalization bounds have several implications that may help guide future work on random forests. First, the role of the CART splitting rule in the success of regression trees has been a focus of much discussion: Is CART good just because it provides a good approximation to the mathematically intractable problem of maximum likelihood estimation over all possible trees, or does the CART splitting rule do something special in itself. Our result helps bring some clarity to this issue. Defining the empirical risk minimizing valid partition $\widehat{V}$ and the optimal valid partition V^* as

$$\widehat{V} = \operatorname{argmin}\left\{ \frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} \left(y_i - T_V(x_i)\right)^2 : V \in \mathcal{V}_{\alpha,\, k} \right\}, \tag{10}$$

$$V^* = \operatorname{argmin}\left\{ \mathbb{E}\left[\left(Y - T^*_V(X)\right)^2\right] : V \in \mathcal{V}_{\alpha,\, k} \right\}, \tag{11}$$

Corollary 2 immediately implies that, asymptotically and with high probability,

$$\mathbb{E}\left[\left(Y - T_{\widehat{V}}(X)\right)^2\right] - \mathbb{E}\left[\left(Y - T^*_{V^*}(X)\right)^2\right] \\ \leq 4M^2 \sqrt{\frac{\log\left(n\right) \log\left(d\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}} \frac{1}{\sqrt{k}}. \tag{12}$$

TABLE 1
Summary of notation.

d	Dimension of feature space: $x \in [0,\, 1]^d$
k	Minimum leaf-size
n	Expected number of training examples
n_{obs}	Actual number of training examples $n_{obs} \sim \text{Poisson}(n)$
M	Bound on the response size $\lvert y \rvert \leq M/2$
α	Bounds the allowable imbalance in recursive partitioning
$\mathcal{V}_{\alpha,\, k}$	Set of $\{\alpha,\, k\}$-valid partitions; see Definition 1
$\mathcal{L}_{\alpha,\, k}$	Set of all leaves generated by $\{\alpha,\, k\}$-valid partitions
$\mathcal{T}_{\alpha,\, k}$	Set of $\{\alpha,\, k\}$-valid trees; see Definition 2
$\mathcal{H}_{\alpha,\, k}$	Set of $\{\alpha,\, k\}$-valid forests; see Definition 3
$\mu(R)$	Lebesgue measure of a rectangle R
$\#R$	Number of training examples inside a rectangle R
$S(R)$	Support of a rectangle R; see (16)

In other words, our result implies that if we could compute the empirical risk minimizer over valid trees, then we would be guaranteed to do well in comparison with the best possible tree; by Theorem 3, a similar result also applies to forests. Thus, it appears that the details of greedy splitting rules are not so important in themselves, and developing better heuristics for approximating the empirical risk minimizing trees and forests may yield good results.

Another repeatedly studied question about random forests is why they handle high-dimensional data so well. Several authors such as Biau [4] and Scornet et al. [36] have discussed the performance of random forests under the assumption that the true signal is low-dimensional and that the tree splits are concentrated on the useful features; they also discuss regimes in which CART will in fact find the right dimensions to split on. These results, however, make strong assumptions about the shape of the conditional mean function $\mathbb{E}\left[Y \,\middle|\, X = x\right]$, and only hold in a regime where d is possibly large but fixed. In comparison, our result implies that the random forest predictive surface will be stable in the standard ($d \to \infty$) high-dimensional regime, regardless of the shape of the true conditional mean function. To our knowledge, our result provides the first practical theoretical evidence that trees and forests should yield stable statistical estimators when d is large.

3. Outline and Notation.

Our analysis is built around bounding large deviations of the empirical process

$$\frac{1}{\lvert\{i : x_i \in L\}\rvert} \sum_{\{i : x_i \in L\}} y_i - \mathbb{E}\left[Y \,\middle|\, X \in L\right], \tag{13}$$

where L ranges over the set $\mathcal{L}_{\alpha,\,k}$ of all possible leaves of a valid partition $V \in \mathcal{V}_{\alpha,\,k}$. To this end, we begin in Section 4 by generalizing ideas from Walther [45] to construct an economical set of rectangles $\mathcal{R}$ that can uniformly approximate all possible leaves $\mathcal{L}$. Then, in Section 5, we use Chernoff-Hoeffding style concentration arguments to control the tail-behavior of (13). These results together yield our first main result, Theorem 1.

In Section 6, we complement our concentration analysis with matching lower bounds. Thus, the result in Theorem 1 is the best rate of convergence guarantee we could hope for with generic classification trees. Finally, in Section 7 we extend our analysis to ensembles of trees and prove Theorem 3 for random forests. All proofs are given in the appendix.

Throughout our analysis, we assume that we have $n_{obs} \sim \text{Poisson}\,(n)$ labeled training examples $(x_i,\, y_i) \in [0,\, 1]^d \times [-M/2,\, M/2]$. We denote rectangles $R \in [0,\, 1]^d$ by

$$R = \bigotimes_{j=1}^{d} \left[r_j^-,\, r_j^+\right], \text{ where } 0 \leq r_j^- < r_j^+ \leq 1 \text{ for all } j = 1,\, ...,\, d, \tag{14}$$

writing the Lebesgue measure of R as $\mu\,(R)$ and and the number of training examples x_i inside R as $\#R$:

$$\mu\,(R) = \prod_{j=1}^{d} \left(r_j^+ - r_j^-\right), \quad \#R = \left|\{i : x_i \in L\}\right|. \tag{15}$$

Notice that, marginally, $\#R \sim \text{Poisson}\,(n\mu\,(R))$. For any rectangle R, we define its support as

$$S(R) = \left\{j \in 1,\, ...,\, d : r_j^- \neq 0 \text{ or } r_j^+ \neq 1\right\}; \tag{16}$$

these are the features used in defining R. Finally, we write $\mathcal{L}_{\alpha,\,k}$ for the set of all possible leaves associated with a valid partition $V \in \mathcal{V}_{\alpha,\,k}$. This notation is summarized in Table 1.

4. A Set of Approximating Rectangles.

Our first result effectively bounds the complexity of the space of rectangles over the unit cube by showing how all such rectangles can be well-approximated using an economical set of rectangles $\mathcal{R}$. We detail a constructive characterization of $\mathcal{R}$ in Section 4.1; this construction is a generalization of the one used by Walther [45] to study scan statistics.

Theorem 4. *Let $S \in \{1, ..., d\}$ be a set of size $|S| = s$, and let $w, \varepsilon \in (0, 1)$. Then, there exists a set of rectangles $\mathcal{R}_{S,\, w,\, \varepsilon}$ such that the following properties hold.*

- *Any rectangle R with support $S(R) \subseteq S$ and of volume $\mu(R) \geq w$ can be well approximated by elements in $\mathcal{R}_{S,\, w,\, \varepsilon}$ from both above and below. Specifically, there exist rectangles R_-, $R_+ \in \mathcal{R}_{S,\, w,\, \varepsilon}$ such that*

$$R_- \subseteq R \subseteq R_+, \text{ and} \tag{17}$$

$$e^{-\varepsilon}\mu\left(R_+\right) \leq \mu\left(R\right) \leq e^{\varepsilon}\mu\left(R_-\right). \tag{18}$$

- *The set $\mathcal{R}_{S,\, w,\, \varepsilon}$ has cardinality bounded by*

$$\left|\mathcal{R}_{S,\, w,\, \varepsilon}\right| \leq \frac{1}{w}\left(\frac{8s^2}{\varepsilon^2}\left(1 + \log_2\left\lfloor\frac{1}{w}\right\rfloor\right)\right)^s \cdot \left(1 + \mathcal{O}\left(\varepsilon\right)\right). \tag{19}$$

In order to approximate all possible s-sparse rectangles, we can use the set

$$\mathcal{R}_{s,\, w,\, \varepsilon} = \cup_{|S|=s}\mathcal{R}_{S,\, w,\, \varepsilon} \tag{20}$$

of size

$$\left|\mathcal{R}_{s,\, w,\, \varepsilon}\right| \leq \binom{d}{s}\frac{1}{w}\left(\frac{8s^2}{\varepsilon^2}\left(1 + \log_2\left\lfloor\frac{1}{w}\right\rfloor\right)\right)^s \cdot \left(1 + \mathcal{O}\left(\varepsilon\right)\right). \tag{21}$$

As shown in the result below, by setting

$$s_{n,\, k,\, \alpha} = \left\lfloor\frac{\log\left(n/k\right)}{\log\left(1/\left(1-\alpha\right)\right)}\right\rfloor + 1, \tag{22}$$

we can use $\mathcal{R}_{s,\, w,\, \varepsilon}$ to approximate all possible tree leaves to within error ε.

Corollary 5. *Suppose that α is fixed, that $n/k \to \infty$, and that we set s as in (22). Then, with probability tending to 1, every leaf $L \in \mathcal{L}_{\alpha,\, k}$ with Lebesgue-measure $\mu(L) \geq w$ can be ε-approximated from above and below by rectangles R_-, $R_+ \in \mathcal{R}_{s,\, w,\, \varepsilon}$ in the sense of (17) and (18).*

We end this section with a useful bound on the size of the approximating set $\mathcal{R}_{s,\, w,\, \varepsilon}$. Assuming that k, n, and d scale jointly in a way that makes the $\mathcal{O}(\cdot)$ term small, we recover the $\log(n)\log(d)$ scaling of Theorem 1.

Lemma 6. *Suppose that we set*

$$w = (1-\eta)\,\frac{k}{n}, \quad \varepsilon = \frac{1}{\sqrt{k}}, \quad \textit{and} \;\; s = \left\lfloor \frac{\log(n/k)}{\log\left((1-\alpha)^{-1}\right)} \right\rfloor + 1, \tag{23}$$

where $0 < \eta < 1$ *and* $0 < \alpha < 0.5$ *are fixed constants. Then,*

$$\log|\mathcal{R}_{s,\,w,\,\varepsilon}| = \frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)} + \mathcal{O}\left(\max\left\{\log(n)\log(\log(n))\,,\,\log(d)\,,\,\log(k)\log(n/d)\right\}\right). \tag{24}$$

4.1. *Constructing Approximating Rectangles.* Without loss of generality, we can take $S = \{1, ..., s\}$; thus, our job is to ε-approximate all rectangles $R \in [0,\,1]^s$ of volume at least w. When $s = 1$, it is easy to verify that we can construct an approximating set containing on the order of w^{-2} elements that ε-approximate all possible intervals of length greater than w: we can build such a set by, e.g., considering all rectangles of the form $[a \cdot w\varepsilon/2,\, b \cdot w\varepsilon/2]$ where a and b are integers.

A naive extrapolation of this idea may suggest that, as s grows, the number of required rectangles scales as w^{-2s}: this is what we would get by varying all the parameters r_j^- and r_j^+ freely. However, as shown by the construction below, this guess is much too pessimistic. The reason for this is that the volume constraint $\mu(R) = \prod_{j=1}^s (r_j^+ - r_j^-) \geq w$ becomes more and more stringent as the dimension s grows, because every dimension along which $r_j^- \not\approx 0$ or $r_j^+ \not\approx 1$ geometrically cuts the size of $\mu(R)$. For example, we can immediately verify that if $\mu(R) \geq w$, then $r_j^+ - r_j^- \leq 0.5$ can can hold for at most $\log_2(w^{-1})$ coordinates.

The construction below exploits the intuition that at most a few coordinates can be active on a small scale. Generalizing ideas from [45], we define $\mathcal{R}$ as the set of all rectangles of the form $R = \bigotimes_{j=1}^s [r_j^-,\, r_j^+]$, with

$$r_j^- = a_j 2^{\tau_j - 1}\,\frac{w\varepsilon}{s} \;\; \text{and} \;\; r_j^+ = \min\left\{1,\, r_j^- + w2^{\tau_j} + b_j 2^{\tau_j-1}\,\frac{w\varepsilon}{s}\right\}, \tag{25}$$

such that

$$a_j \in 0,\, 1,\, ..,\, \left\lfloor 2^{1-\tau_j}\,\frac{s}{w\varepsilon} \right\rfloor,\;\; b_j \,\in\, 0,\, 1,\, ..,\, \left\lceil \frac{2s}{\varepsilon} \right\rceil, \tag{26}$$

$$\tau_j \in 0,\, 1,\, ...,\, \lfloor \log_2 w^{-1} \rfloor, \;\; \text{and} \;\; \sum_{j=1}^s \tau_j \geq (s-1)\log_2\left(\frac{1}{w}\right) - s. \tag{27}$$

In this construction, the j-th interval $[r_j^-,\, r_j^+]$ is on the scale $w\, 2^{\tau_j}$. The observation that only a few coordinates j can be active on a small scale is encoded in the lower bound (27) on $\sum_{j=1}^s \tau_j$. The lemma below confirms that this this approximating set is valid.

Lemma 7. *Given any rectangle R with support $S(R) \subseteq S$ and volume $\mu(R) \geq w$, we can select rectangles R_- and R_+ satisfying* (17) *and* (18) *from the approximating set $\mathcal{R}$ defined above.*

To complete our characterization of the approximating set, it suffices to bound the cardinality of $\mathcal{R}$. This computation is carried out in Appendix A, in the proof of Theorem 4.

5. Rectangles and Poisson Processes. In the previous section, we showed how to ε-approximate all possible tree leaves under the Lebesgue measure on $[0,\, 1]^d$. However, to understand the behavior of decision trees, we do not want to approximate tree leaves in terms of Lebesgue measure, but rather in terms of the empirical measure induced by the training features $\{x_i\}_{i=1}^n$, which, given our assumptions, are drawn from a uniform Poisson process over $[0,\, 1]^d$.

The following result lets us get around this issue by showing that our Poisson process is concentrated enough that, with high probability, the set $\mathcal{R}_{s,\, w,\, \varepsilon}$ is also a good approximating set in terms of the empirical measure induced by the training data. The proof of the result below is built around a Chernoff-Hoeffding style concentration bound for Poisson random variables.

Theorem 8. *Suppose that Assumption 1 holds, and that we have a sequence of inputs n, d_n and k_n. Let $\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}$ be as defined in* (20) *with s as in* (22)*, and choose ε_n and w_n such that*

$$\lim_{n\to\infty} \varepsilon_n = 0,\ n\,\varepsilon_n = \mathcal{O}(1)\,, \quad \textit{and} \quad w_n = (1-\eta)\,\frac{k_n}{n} \tag{28}$$

for some fixed $\eta > 0$. Finally, suppose that

$$\lim_{n\to\infty} \frac{\log\left(\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right)}{k_n} = 0. \tag{29}$$

Then, for any $\delta > 0$, there exists a sequence δ_n with $\lim_{n\to\infty} \delta_n = \delta$, such that the following statement holds with probability at least $1 - \delta_n$. For every possible leaf $L \in \mathcal{L}_{\alpha,\, k_n}$, we can select a rectangle $R \in \mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}$ such that

$L \subseteq R$, $\mu(R) \leq e^{\varepsilon_n} \mu(L)$, and,

$$\#R \leq \#L + 3\,\varepsilon_n \#L + 3\sqrt{2\log\left(\frac{2\left|\mathcal{R}_{s_n,\,w_n,\,\varepsilon_n}\right|}{\delta}\right)\#L} + o\left(\sqrt{\log\left(\left|\mathcal{R}_{s_n,\,w_n,\,\varepsilon_n}\right|\right)\#L}\right). \tag{30}$$

We can then turn this result into a concentration bound on our empirical process of interest (13).

Corollary 9. *Suppose that the conditions of Theorem 8 hold, that the parameters ε_n and w_n are chosen as in (28) with a small enough value $\eta > 0$, and that $|y_i| \leq M/2$. Then,*

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{L\in\mathcal{L}}\left|\frac{1}{\#L}\sum_{\{i:x_i\in L\}} y_i - \mathbb{E}\left[Y \middle| X \in L\right]\right| \leq 6M\sqrt{\frac{\log\left(\left|\mathcal{R}_{s_n,\,w_n,\,\varepsilon_n}\right|\right)}{k_n}}\right] = 1. \tag{31}$$

We have now gathered all the ingredients required to prove Theorem 1, which follows from combining Corollary 5 and Lemma 6 with Corollary 9.

5.1. *Uniform Concentration over Rectangles.*

In this section, we present a series of technical results that lead up to our key concentration result for Poisson processes, Theorem 8. We begin with an adaptation of the classical Chernoff-Hoeffding bound to Poisson random variables.

Lemma 10. *Let Z be a Poisson random variable with mean-m, and let $0.6\,m \leq \Delta$. Then,*

$$\mathbb{P}\left[Z \geq m + \Delta\right] \leq \exp\left[-\frac{\Delta^2}{2m}\left(1 - \frac{\Delta}{m}\right)\right],$$
$$\mathbb{P}\left[Z \leq m - \Delta\right] \leq \exp\left[-\frac{\Delta^2}{2m}\right].$$

In the context of our proof, the most useful consequence of this lemma is the corollary below.

Corollary 11. *Fix $\delta > 0$, and define the event*

$$\mathcal{A}^{\delta}_{n,\,s,\,w,\,\varepsilon}\ :\ \sup\left\{\frac{\left|\#R - n\,\mu(R)\right|}{\sqrt{n\,\mu\left(R\right)}} : R \in \mathcal{R}_{s,\,w,\,\varepsilon},\ \mu\left(R\right) \geq e^{-\varepsilon} w\right\} \leq \sqrt{2\log\left(\frac{2\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|}{\delta}\right)}. \tag{32}$$

Then, letting n, s_n, w_n, ε_n be sequences such that

$$\lim_{n\to\infty} \frac{e^{\varepsilon_n}\log\left(\left|\mathcal{R}_{s_n,\,w_n,\,\varepsilon_n}\right|\right)}{n\,w_n} = 0, \tag{33}$$

we have that

$$\liminf_{n\to\infty} \mathbb{P}\left[\mathcal{A}^{\delta}_{n,\ s_n,\,w_n,\,\varepsilon_n}\right] \geq 1 - \delta.$$

Now, the relation (32) is only valid for rectangles R contained in our finite approximating set $\mathcal{R}_{s,\,w,\,\varepsilon}$. In general, of course, the leaves $L \in \mathcal{L}_{\alpha,\,k}$ we want to study will not be in $\mathcal{R}_{s,\,w,\,\varepsilon}$. The following result gets us around this issue by providing a bound that is valid for all rectangles R, not just those in $\mathcal{R}_{s,\,w,\,\varepsilon}$.

Lemma 12. *Suppose that the event $\mathcal{A}^{\delta}_{n,\,s,\,w,\,\varepsilon}$ defined in Corollary 11 has occurred. Then, for all rectangles R with $\mu\left(R\right) \geq w$,*

$$\begin{aligned} e^{-\varepsilon} n\mu\left(R\right) - \sqrt{2n\mu\left(R\right)\log\left(\frac{2\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|}{\delta}\right)} & \\ \leq \#R & \\ \leq e^{\varepsilon} n\mu\left(R\right) + e^{\varepsilon/2}\sqrt{2n\mu\left(R\right)\log\left(\frac{2\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|}{\delta}\right)}. & \end{aligned}$$

In order for this result to be useful for understanding the leaves of decision trees, we need to show that all possible leaves L will satisfy the condition $\mu\left(L\right) \geq w$; the result below gives us such a guarantee.

Corollary 13. *Under the conditions of Theorem 8,*

$$\lim_{n\to\infty} \mathbb{P}\left[\inf_{L\in\mathcal{L}} \mu\left(L\right) \geq w_n\right] = 1. \tag{34}$$

With these results in hand, proving Theorem 8 reduces to algebra. The details are worked out in Appendix A.

6. Lower Bounds.

In this section, we show that the convergence rate given in (7) cannot be improved. Lin and Jeon [28] have also studied lower bounds for forest convergence; however, they only consider non-adaptive forests and so their lower bounds are substantially weaker. Our main result is given below.

Theorem 14. *Suppose that the assumptions of Theorem 1 hold and that $\alpha \leq 0.2$. Then, there exists a conditional distribution of Y given X such that*

$$\lim_{n,\,d,\,k\to\infty} \mathbb{P}\left[\sup_{x\in[0,\,1]^d,\,V\in\mathcal{V}_{\alpha,\,k}} \left|T_V\left(x\right)-T_V^*\left(x\right)\right| \geq \frac{M}{10}\sqrt{\frac{\log\left(n\right)\log\left(d\right)}{k}}\right]=1. \tag{35}$$

To establish this result, we take the Y_i to be i.i.d. and independent of $\mathcal{X}$ and n_{obs} with $\mathbb{P}\left[Y_1 = M/2\right] = \mathbb{P}\left[Y_1 = -M/2\right] = 1/2$. We will construct $N = N(n)$ nodes $L_1, \dots, L_N \in \mathcal{V}_{\alpha,\,k}\left(\mathcal{X}\right)$ and then consider for $j = 1, \dots, N$:

$$T_j \;:=\; \frac{1}{|L_j|}\sum_{\{i:X_i\in L_j\}} Y_i \tag{36}$$

as well as the approximations

$$\widetilde{T}_j \;:=\; \frac{1}{|L_j|}\sum_{\{i:X_i\in L_j\}} \widetilde{Y}_i \tag{37}$$

where the $\widetilde{Y}_i$ are given by

$$\widetilde{Y}_i = Y_i\,\left|Z_i\right| \text{ where } Z_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\,1\right); \tag{38}$$

notice in particular that the $\widetilde{Y}_i$ are jointly distributed as independent Gaussian random variables with variance $M^2/4$.

The idea of the proof is to construct $N \sim \log(n)\log(d)$ nodes L_j whose pairwise intersections are small enough that the multivariate normal distribution of the standardized $\widetilde{T}_j$ has correlations that are bounded away from unity; the construction of the L_j is detailed in the proof of Lemma 15. A normal approximation lemma [25] then allows us to stochastically lower-bound the distribution of $\max_j \widetilde{T}_j$ in terms of the distribution of a correlated multivariate normal that can be constructed in a simple way from an i.i.d. normal sequence. Specifically, we show establish the following lower bound of the tail of the $\widetilde{T}_j$.

Lemma 15. *Suppose that $\alpha \leq 0.2$ and that that the dimension d satisfies $d = d_n := n^r$ for some $r > 0$. Then, there exists a set of α-valid nodes L_j of size $N = \log(n) \log(d) \,/\, \log(5)$ chosen independently of the Y_i and $\widetilde{Y}_i$ for which*

$$\lim_{n \to \infty} \mathbb{P}\left[\max_{j=1,\ldots,N} \widetilde{T}_j \,\geq\, (1 - \varepsilon_n)\, M \sqrt{\frac{2}{5 \log 5}} \sqrt{\frac{\log(n) \log(d)}{k}}\right] \,=\, 1 \tag{39}$$

whenever $\varepsilon_n \frac{\log n}{\log \log n} \to \infty$.

In the second step, we establish a coupling between T_i and the $\widetilde{T}_i$ that is tight enough to guarantee that the approximation error $\max_j\{\widetilde{T}_j - T_j\}$ is smaller than $\max_j \widetilde{T}_j$. To get this coupling, we use the following bound on the moment-generating function of $Y_i - \widetilde{Y}_i$.

Lemma 16. *Let $\mathbb{P}[Y = 1] = \mathbb{P}[Y = -1] = 1/2$ and $Z \sim N(0, 1)$ independent of Y. Then*

$$\mathbb{E}\left[\exp\left\{t\left(Y - Y|Z|\right)\right\}\right] \,\leq\, \exp\left\{\left((1 - \sqrt{\frac{2}{\pi}}\right) t^2\right\} \tag{40}$$

for t in a neighborhood of zero.

This lemma implies the bound on $\max_j\{\widetilde{T}_j - T_j\}$ given below. The lower bound in Theorem 14 then follows immediately from Lemma 15 and Corollary 17, together with the observation that $\sqrt{2/(5 \log 5)} - 0.5/\sqrt{\log 5} \geq 1/10$.

Corollary 17. *Suppose that the statistics T_j and $\widetilde{T}_j$ are constructed as in (36 - 38), with leaf nodes L_j chosen independently of the Y_i and $\widetilde{Y}_i$. Then,*

$$\lim_{n \to \infty} \mathbb{P}\left[\max_{j=1,\ldots,N} (\widetilde{T}_j - T_j) \,\leq\, \frac{M}{2} \sqrt{\frac{\log(N)}{k}}\right] \,=\, 0. \tag{41}$$

7. From Trees to Forests. We finish our theoretical work by extending the generalization result from Corollary 2 to forests. Recall that a forest H is defined as

$$H(x) = \frac{1}{B} \sum_{b=1}^{B} T^{(b)}(x) \text{ for some trees } T^{(b)} \in \mathcal{T}_{\alpha,\,k},$$

where the $T^{(b)}$ are valid trees. We can then verify that, for any x and y,

$$\frac{1}{B}\sum_{b=1}^{B}\left(y - T^{(b)}(x)\right)^2 = (y - H(x))^2 + \frac{1}{2}\binom{B}{2}^{-1}\sum_{b<b'}\left(T^{(b)}(x) - T^{(b')}(x)\right)^2. \tag{42}$$

Thus, to understand the error rate of a forest, it suffices to understand the error rates of individual trees and their correlations. Corollary 2 already bounds the former; we bound the latter below.

Lemma 18. *Under the conditions of Theorem 1,*

$$\lim_{n,\,d,\,k\to\infty}\mathbb{P}\left[\sup_{T^{(1)},\,T^{(2)}\in\mathcal{T}_{\alpha,\,k}}\left|\frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\left(T^{(1)}\left(x_i\right) - T^{(2)}\left(x_i\right)\right)^2 - \mathbb{E}\left[\left(T^{(1)}\left(x\right) - T^{(2)}\left(x\right)\right)^2\right]\right| \leq 17\sqrt{\frac{\log\left(n\right)\log\left(d\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}\frac{1}{\sqrt{k}}}\right] = 1. \tag{43}$$

We note that our proof strategy here is dependent on the bias-variance decomposition (42), which lets us concisely describe the error rate of an ensemble of trees. Controlling the error rate of ensembles with other loss functions can be much more difficult. For example, Freund et al. [19] need to allow for abstentions in order to analyze the 0-1 error of ensembles. Finding conditions under which we can use Theorem 1 to bound the classification error rate of random forests could yield further insights about their behavior.

8. Conclusion.

We introduced *adaptive concentration*, a theoretical framework that lets us prove rate-optimal uniform convergence bounds for tree and forest predictive surfaces. Motivated by the post-selection inference results of Berk et al. [3], our approach treats split selection as a model selection phase, and then shows that the fitted tree converges uniformly to the best possible tree supported on the same splits.

The fact that our method does not seek convergence guarantees to the true conditional mean function $m(x) = \mathbb{E}\left[Y \,\middle|\, X = x\right]$ can be seen as both its main weakness and strength. On the one hand, some statisticians may feel uncomfortable using estimators without consistency guarantees of the form $\lim_{n\to\infty}\hat{y}(x) = m(x)$. On the other hand, such consistency results invariably require strong regularity conditions on the unknown function $m(\cdot)$ that are unverifiable in practice, e.g., that $m(\cdot)$ is Lipschitz. Such assumptions are especially problematic with machine learning methods like random forests, which are popular exactly when we do not know anything about the true shape of $m(\cdot)$; if we had prior insight about the conditional mean function, then we would in general not use random forests and instead prefer an estimator that takes into account our knowledge of its shape. By taking the optimal forest with the same splits instead of $m(\cdot)$ as the reference point, adaptive concentration allows us to give strong stability guarantees for tree and forest predictive surfaces without needing any unverifiable assumptions on the true conditional mean function.

References.

[1] Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

[2] Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3: 463–482, 2003.

[3] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

[4] Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

[5] Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008.

[6] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[7] Leo Breiman. Out-of-bag estimation. Technical report, Statistics Department, University of California, Berkeley, 1996.

[8] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[9] Leo Breiman. Consistency for a simple model of random forests. *Statistical Department, University of California at Berkeley. Technical Report*, (670), 2004.

[10] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC press, 1984.

[11] Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4): 927–961, 2002.

[12] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

[13] Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[14] D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[15] Misha Denil, David Matheson, and Nando de Freitas. Consistency of online random forests. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1256–1264, 2013.

[16] Misha Denil, David Matheson, and Nando de Freitas. Narrowing the gap: Random

forests in theory and in practice. In *Proceedings of The 31st International Conference on Machine Learning*, pages 665–673, 2014.

[17] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[18] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

[19] Yoav Freund, Yishay Mansour, and Robert E Schapire. Generalization bounds for averaged classifiers. *Annals of Statistics*, pages 1698–1722, 2004.

[20] Jisoo Ham, Yangchi Chen, Melba M Crawford, and Joydeep Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(3):492–501, 2005.

[21] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. New York: Springer, 2nd edition, 2009.

[22] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[23] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

[24] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

[25] Malcolm R Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and related properties of random sequences and processes*, volume 21. Springer-Verlag New York, 1983.

[26] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference with the lasso. *arXiv preprint arXiv:1311.6238*, 2013.

[27] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002. URL http://CRAN.R-project.org/doc/Rnews/.

[28] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

[29] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso (with discussion). *The Annals of Statistics*, 42(2):413–468, 2014.

[30] Nicolai Meinshausen. Quantile regression forests. *The Journal of Machine Learning Research*, 7:983–999, 2006.

[31] Lucas Mentch and Giles Hooker. Ensemble trees and CLTs: Statistical inference for supervised learning. *arXiv preprint arXiv:1404.6473*, 2014.

[32] M Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.

[33] Anantha M Prasad, Louis R Iverson, and Andy Liaw. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, 9(2):181–199, 2006.

[34] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. On-line random forests. In *IEEE International Conference on Computer Vision*, pages 1393–1400. IEEE, 2009.

[35] Erwan Scornet. On the asymptotics of random forests. *arXiv preprint arXiv:1409.2090*, 2014.

[36] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics, in press*, 2015.

[37] Clayton Scott and Robert D Nowak. Minimax-optimal classification with dyadic decision trees. *Information Theory, IEEE Transactions on*, 52(4):1335–1353, 2006.

[38] Joseph Sexton and Petter Laake. Standard errors for bagged and random forest

estimators. *Computational Statistics & Data Analysis*, 53(3):801–811, 2009.

[39] Tao Shi, David Seligson, Arie S Belldegrun, Aarno Palotie, and Steve Horvath. Tumor classification by tissue microarray profiling: Random forest clustering applied to renal cell carcinoma. *Modern Pathology*, 18(4):547–557, 2004.

[40] Jonathan Taylor, Joshua Loftus, and Ryan Tibshirani. Tests in adaptive regression via the Kac-Rice formula. *arXiv preprint arXiv:1308.3020*, 2013.

[41] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[42] William N Venables and Brian D Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

[43] Stefan Wager. Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*, 2014.

[44] Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651, 2014.

[45] Guenther Walther. Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics*, 38(2):1010–1033, 2010.

APPENDIX A: PROOFS

Proof of Theorem 1. In order to apply Corollary 9, set $\varepsilon = 1/\sqrt{k}$, and $w = (1-\eta)k/n$ for a small enough choice of $\eta > 0$. Thanks to (6), we can use Lemma 6 to bound

$$\log\left(\mathcal{R}_{s,\,n,\,\varepsilon}\right) = \frac{\log\left(n\right)\log\left(d\right)}{\log\left(1-\alpha\right)^{-1}} + \mathcal{O}\left(\log\left(n\right)\log\left(\log\left(n\right)\right)\right).$$

Given this bound, Assumption 2 also implies that the condition (29) holds. Thus, given our choices of w and ε we can apply Corollary 9 directly, and (31) is equivalent to the desired conclusion.

Proof of Corollary 2. Let $\mathcal{L}(S)$ denote the set of leaves of the partition underlying T, and write $\widehat{\mathbb{E}}$ and $\widehat{\mathbb{P}}$ for plug-in estimates for $\mathbb{E}$ and $\mathbb{P}$ on the training sample. With this notation

$$\begin{aligned}
&\frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\left(y_i - T(x_i)\right)^2 \\
&\quad= \sum_{L\in\mathcal{L}(S)} \widehat{\operatorname{Var}}\left[Y \,\middle|\, X\in L\right]\widehat{\mathbb{P}}\left[X\in L\right], \\
&\mathbb{E}\left[\left(Y - T\left(X\right)\right)^2\right] \\
&\quad= \sum_{L\in\mathcal{L}(S)}\left(\widehat{\mathbb{E}}\left[Y \,\middle|\, X\in L\right] - \mathbb{E}\left[Y \,\middle|\, X\in L\right]\right)^2\mathbb{P}\left[X\in L\right] \\
&\qquad+ \sum_{L\in\mathcal{L}(S)}\operatorname{Var}\left[Y \,\middle|\, X\in L\right]\mathbb{P}\left[X\in L\right].
\end{aligned}$$

First, as a direct consequence of Theorem 1, we see that the bias term

$$\begin{aligned}
&\sum_{L\in\mathcal{L}(S)}\left(\widehat{\mathbb{E}}\left[Y \,\middle|\, X\in L\right] - \mathbb{E}\left[Y \,\middle|\, X\in L\right]\right)^2\mathbb{P}\left[X\in L\right] \\
&\qquad= \mathcal{O}_p\left(\frac{\log(d)\log(n)}{k}\right),
\end{aligned}$$

meaning that it decays faster than the main component of our error bound. Now, setting $\varepsilon = 1/\sqrt{k}$ in the bound from Lemma 12, we find that, uniformly over all leaves L,

$$\frac{\left|\mathbb{P}\left[X\in L\right] - \widehat{\mathbb{P}}\left[X\in L\right]\right|}{\mathbb{P}\left[X\in L\right]} \leq \sqrt{2}\sqrt{\frac{\log\left(n\right)\log\left(d\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}}\,\frac{1}{\sqrt{k}}\left(1 + o_P(1)\right).$$

And, by Theorem 1 and the proof of Corollary 9, for any $\eta > 0$,

$$\begin{aligned}&\left|\widehat{\mathbb{E}}\left[Y \,\middle|\, X \in L\right] - \mathbb{E}\left[Y \,\middle|\, X \in L\right]\right| \\ &\qquad \leq 4\sqrt{2}M\sqrt{\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}}\,\frac{1}{\sqrt{k}}\left(1 + \eta + o_P(1)\right), \\ &\left|\widehat{\mathbb{E}}\left[Y^2 \,\middle|\, X \in L\right] - \mathbb{E}\left[Y^2 \,\middle|\, X \in L\right]\right| \\ &\qquad \leq \sqrt{2}M^2\sqrt{\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}}\,\frac{1}{\sqrt{k}}\left(1 + \eta + o_P(1)\right);\end{aligned}$$

thus, because $\lim_{n\to\infty}\log(n)\log(d)\,/\log(1/(1-\alpha)) = 0$ we see that

$$\begin{aligned}&\left|\widehat{\operatorname{Var}}\left[Y \,\middle|\, X \in L\right] - \operatorname{Var}\left[Y \,\middle|\, X \in L\right]\right| \\ &\qquad \leq \sqrt{2}M^2\sqrt{\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}}\,\frac{1}{\sqrt{k}}\left(1 + \eta + o_P(1)\right).\end{aligned}$$

Finally, using the relation

$$\begin{aligned}&\widehat{\operatorname{Var}}\left[Y \,\middle|\, X \in L\right]\widehat{\mathbb{P}}\left[Y \,\middle|\, X \in L\right] - \operatorname{Var}\left[Y \,\middle|\, X \in L\right]\mathbb{P}\left[X \in L\right] \\ &\quad = \left(\widehat{\operatorname{Var}}\left[Y \,\middle|\, X \in L\right] - \operatorname{Var}\left[Y \,\middle|\, X \in L\right]\right)\mathbb{P}\left[X \in L\right] \\ &\qquad + \frac{\widehat{\mathbb{P}}\left[Y \,\middle|\, X \in L\right] - \mathbb{P}\left[Y \,\middle|\, X \in L\right]}{\mathbb{P}\left[X \in L\right]}\,\widehat{\operatorname{Var}}\left[Y \,\middle|\, X \in L\right]\mathbb{P}\left[X \in L\right],\end{aligned}$$

and the fact that

$$\sum_{L \in \mathcal{L}} \mathbb{P}\left[X \in L\right] = 1, \text{ and}$$
$$\widehat{\operatorname{Var}}\left[Y \,\middle|\, X \in L\right] \leq \frac{M^2}{4},$$

we can put everything together and find that

$$\begin{aligned}&\frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\left(y_i - T(x_i)\right)^2 - \mathbb{E}\left[\left(Y - T(X)\right)^2\right] \\ &\quad \leq \left(\sqrt{2} + \frac{\sqrt{2}}{4}\right)M^2\sqrt{\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}}\,\frac{1}{\sqrt{k}}\left(1 + \eta + o_P(1)\right).\end{aligned}$$

Making the constant $\eta > 0$ small enough, we recover the stated result because $1.25\sqrt{2} < 2$.

Proof of Theorem 3. Because (7) holds simultaneously for every valid tree T_V, we immediately also see that

$$\lim_{n,\, d,\, k \to \infty} \mathbb{P}\left[\sup_{x \in [0,\,1]^d,\, V_1,\, \ldots,\, V_B \in \mathcal{V}_{\alpha,\, k}} \left| H_{\{V\}_1^B}(x) - H^*_{\{V\}_1^B}(x) \right| \right. \\ \left. \leq 6M \sqrt{\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}\, \frac{1}{\sqrt{k}}} \right] = 1.$$

Thus, it only remains to establish (9). Recall that by the bias-variance decomposition (42), for any x and y,

$$\frac{1}{B}\sum_{b=1}^{B}\left(y - T^{(b)}(x)\right)^2 = (y - H(x))^2 \\ + \frac{1}{2}\binom{B}{2}^{-1}\sum_{b<b'}\left(T^{(b)}(x) - T^{(b')}(x)\right)^2.$$

Now, Corollary 2 implies that, with probability tending to 1, for any $\eta > 0$,

$$\frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\frac{1}{B}\sum_{b=1}^{B}\left(y_i - T^{(b)}(x_i)\right)^2 - \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\left[\left(Y - T^{(b)}(X)\right)^2\right] \\ \leq 1.25\sqrt{2}\sqrt{\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}\,\frac{1}{\sqrt{k}}}\left(1 + \eta + o(1)\right).$$

Meanwhile, Lemma 18 implies that, again with probability tending to 1, for any $\eta > 0$,

$$-\frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\binom{B}{2}^{-1}\sum_{b<b'}^{B}\left(T^{(b)}(x_i) - T^{(b')}(x_i)\right)^2 \\ + \binom{B}{2}^{-1}\sum_{b=1}^{B}\mathbb{E}\left[\left(T^{(b)}(X) - T^{(b')}(X)\right)^2\right] \\ \leq 12\sqrt{2}\sqrt{\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}\,\frac{1}{\sqrt{k}}}\left(1 + o(1)\right).$$

Combining these bounds and noting that $(1.25 + 12/2)\sqrt{2} < 11$ leads to the desired conclusion.

Proof of Theorem 4. Given Lemma 7, in order to complete the proof of Theorem 4 it suffices to bound the cardinality of the approximating set defined in Section 4.1. To do so, we first observe that for fixed values of $\{\tau_j\}$, the number of possible choices for the $\{a_j\}$ and $\{b_j\}$ is bounded by

$$\begin{aligned}\prod_{j=1}^{s}&\left(1+\left\lfloor 2^{1-\tau_j}\frac{s}{w\varepsilon}\right\rfloor\right)\left(1+\left\lceil\frac{2s}{\varepsilon}\right\rceil\right)\\ &=\left(\frac{4s^2}{w\varepsilon^2}\right)^s 2^{-\sum_{j=1}^s\tau_j}\cdot\left(1+\mathcal{O}\left(\varepsilon\right)\right)\\ &\leq\left(\frac{4s^2}{w\varepsilon^2}\right)^s 2^s\left(\prod_{j=1}^{s}\frac{r_j^+-r_j^-}{w}\right)^{-1}\cdot\left(1+\mathcal{O}\left(\varepsilon\right)\right)\\ &=\frac{1}{w}\left(\frac{8s^2}{\varepsilon^2}\right)^s\cdot\left(1+\mathcal{O}\left(\varepsilon\right)\right),\end{aligned}$$

because $\prod_{j=1}^s\left(r_j^+-r_j^-\right)\geq w$. Now, we can loosely bound the number of possible choices for $\{\tau_j\}$ by $\left(1+\log_2 w^{-1}\right)^s$, yielding the desired bound.

Proof of Corollary 5. To show the desired result, it suffices to show that

$$\lim_{n\to\infty}\mathbb{P}\left[\sup_{L\in\mathcal{L}_{\alpha,\,k}}|S(L)|\leq\frac{\log\left(n/k\right)}{\log\left(1/\left(1-\alpha\right)\right)}+1\right]=1; \tag{44}$$

the conclusion then follows directly from Theorem 4. Let n denote the number of observed data points; recall that $n\sim\text{Poisson}\left(N\right)$. Since each child node must be smaller than its parent by at least a factor $1-\alpha$, we must have

$$\#L\leq\left(1-\alpha\right)^{|S(L)|}n,$$

and so, for any $L\in\mathcal{L}$,

$$\begin{aligned}|S(L)|&\leq\frac{\log\left(n/\#L\right)}{\log\left(1/\left(1-\alpha\right)\right)}\\ &\leq\frac{\log\left(n/k\right)}{\log\left(1/\left(1-\alpha\right)\right)}.\end{aligned}$$

Meanwhile,

$$\lim_{n\to\infty}\mathbb{P}\left[\log\left(n/k\right)\leq\log\left(N/k\right)+\frac{1}{\log\left(1/\left(1-\alpha\right)\right)}\right]=1,$$

and so we recover the desired conclusion.

Proof of Lemma 6. Given the parameter choices (23), we can verify that

$$\begin{aligned}\log\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right| &\leq \log\left(\left(\binom{d}{s}\frac{1}{w}\left(\frac{8s^2}{\varepsilon^2}\left(1+\log_2\left\lfloor\frac{1}{w}\right\rfloor\right)\right)\right)^s\cdot\left(1+\mathcal{O}\left(\varepsilon\right)\right)\right)\\ &= \log\binom{d}{s}+2s\log\left(\varepsilon^{-1}\right)+\mathcal{O}\left(\log\left(n\right)\log\left(\log\left(n\right)\right)\right).\end{aligned}$$

Meanwhile,

$$\begin{aligned}\log\binom{d}{s} &\leq s\log\left(d\right)=\frac{\log\left(n/k\right)\log\left(d\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}+\mathcal{O}\left(\log\left(d\right)\right), \text{ and}\\ 2s\log\left(\varepsilon^{-1}\right) &= \frac{\log\left(n/k\right)\log\left(k\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}+\mathcal{O}\left(\log\left(n\right)\right).\end{aligned}$$

Combining these results, we find that

$$\begin{aligned}&\log\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|\\ &\quad\leq\frac{\log\left(n/k\right)\log\left(dk\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}+\mathcal{O}\left(\max\left\{\log\left(n\right)\log\left(\log\left(n\right)\right),\,\log\left(d\right)\right\}\right)\\ &\quad=\frac{\log\left(n\right)\log\left(d\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)}\\ &\qquad\quad+\mathcal{O}\left(\max\left\{\log\left(n\right)\log\left(\log\left(n\right)\right),\,\log\left(d\right),\,\log\left(k\right)\log\left(n/d\right)\right\}\right),\end{aligned}$$

thus completing the proof.

Proof of Lemma 7. We focus on showing how to construct R_+; the construction of R_- is analogous. Recall that, given a rectangle

$$R=\bigotimes_{j=1}^{s}\left[r_j^-,\,r_j^+\right],\text{ our goal is to select a rectangle } R_+=\bigotimes_{j=1}^{s}\left[q_j^-,\,q_j^+\right]$$

from $\mathcal{R}$ such that $R\subseteq R_+$ and $\mu\left(R_+\right)\leq e^{\varepsilon}\mu\left(R\right)$. In order to guarantee this, it is sufficient to check that, for all j,

$$q_j^-\leq r_j^-,\, r_j^+\leq q_j^+,\text{ and } q_j^+-q_j^-\leq e^{\varepsilon/s}\left(r_j^+-r_j^-\right).$$

Now, for each j, define

$$\tau_j=\left\lfloor\log_2\frac{r_j^+-r_j^-}{w}\right\rfloor,$$

let q_j^- be the largest choice of the form (26) such that $q_j^- \leq r_j^-$, and pick q_j^+ analogously. These choices define a rectangle R_+ in $\mathcal{R}$ such that $R \subseteq R_+$. By construction, we immediately see that

$$2^{\sum_{j=1}^s \tau_j} \geq 2^{\sum_{j=1}^s \left(\log_2 \frac{r_j^+ - r_j^-}{w} - 1\right)} \geq 2^{-s} w^{-(s-1)}.$$

Moreover, by definition of τ_j

$$2^{\tau_j} w \leq r_j^+ - r_j^- \leq 2^{\tau_j + 1} w,$$

thus, we can verify that

$$q_j^+ - r_j^+,\, r_j^- - q_j^- \leq 2^{\tau_j - 1} \frac{w\varepsilon}{s} \leq \frac{1}{2} \frac{\varepsilon}{s} \left(r_j^+ - r_j^-\right).$$

This implies that

$$q_j^+ - q_j^- \leq \left(r_j^+ - r_j^-\right)\left(1 + \frac{\varepsilon}{s}\right),$$

and so $|R_+| \leq e^{\varepsilon} |R|$.

Proof of Theorem 8. With probability converging to 1, we know from Corollary 13 that, for all $L \in \mathcal{L}$, $\mu\left(L\right) \geq w$. Thus, by Theorem 4, for each possible leaf $L \in \mathcal{L}_{\alpha,\,k}$ we can select a rectangle $R \in \mathcal{R}_{s_n,\,w_n,\,\varepsilon_n}$ such that $L \subseteq R$ and

$$\mu\left(R\right) \leq e^{\varepsilon_n} \mu\left(L\right).$$

Moreover, on the event $\mathcal{A}^{\delta}_{n,\,s_n,\,w_n,\,\varepsilon_n}$ defined in Corollary 11, we know from Lemma 12 that

$$e^{-\varepsilon_n} n\mu\left(L\right) - \sqrt{2\log\left(\frac{2\left|\mathcal{R}_{s_n,\,w_n,\,\varepsilon_n}\right|}{\delta}\right)} \sqrt{n\mu\left(L\right)} - \#L \leq 0,$$

or equivalently that

$$\begin{aligned} n\mu\left(L\right) &\leq \frac{e^{2\varepsilon_n}}{4}\left(\sqrt{2\log\left(\frac{2\left|\mathcal{R}_{s_n,\,w_n,\,\varepsilon_n}\right|}{\delta}\right)}\right. \\ &\quad \left. + \sqrt{2\log\left(\frac{2\left|\mathcal{R}_{s_n,\,w_n,\,\varepsilon_n}\right|}{\delta}\right) + 4e^{-\varepsilon_n}\#L}\right)^2 \\ &\leq \frac{e^{2\varepsilon_n}}{4}\left(\sqrt{8\log\left(\frac{2\left|\mathcal{R}_{s_n,\,w_n,\,\varepsilon_n}\right|}{\delta}\right)} + \sqrt{4e^{-\varepsilon_n}\#L}\right)^2, \end{aligned}$$

and so

$$n\mu\left(L\right) \leq e^{\varepsilon_n} \#L + \sqrt{8 \log\left(\frac{2\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|}{\delta}\right)} e^{\frac{3\varepsilon_n}{2}} \sqrt{\#L}$$
$$+ 2e^{2\varepsilon_n} \log\left(\frac{2\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|}{\delta}\right).$$

Again by Lemma 12, we know that on the event $\mathcal{A}^{\delta}_{n,\, s_n,\, w_n,\, \varepsilon_n}$,

$$\#R \leq e^{\varepsilon_n} n\mu\left(R\right) + \sqrt{2e^{\varepsilon_n} n\mu\left(R\right) \log\left(\frac{2\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|}{\delta}\right)}$$
$$\leq e^{2\varepsilon_n} n\mu\left(L\right) + e^{\varepsilon_n} \sqrt{2n\mu\left(L\right) \log\left(\frac{2\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|}{\delta}\right)}.$$

Now, by hypothesis, we know that $\#L \geq k_n$ and

$$\lim_{n\to\infty} \frac{1}{k_n} \log\left(\frac{2\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|}{\delta}\right) = 0.$$

Thus, by combining these inequalities, we find that on $\mathcal{A}^{\delta}_{n,\, s_n,\, w_n,\, \varepsilon_n}$

$$\#R \leq e^{3\varepsilon_n} \#L + \sqrt{8 \log\left(\frac{2\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|}{\delta}\right)} e^{\frac{7\varepsilon_n}{2}} \sqrt{\#L}$$
$$+ e^{\frac{3\varepsilon_n}{2}} \sqrt{2 \log\left(\frac{2\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|}{\delta}\right) \#L}$$
$$+ o\left(\sqrt{\log\left(\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|\right) \#L}\right).$$

Since ε_n is converging to 0, this expression simplifies to

$$\#R - \#L \leq 3\varepsilon_n \#L + 3\sqrt{2 \log\left(\frac{2\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|}{\delta}\right) \#L}$$
$$+ o\left(\sqrt{\log\left(\left|\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}\right|\right) \#L}\right),$$

which is what we set out to show.

Proof of Corollary 9. For any leaf L generated by a valid tree, let $R(L) \in \mathcal{R}_{s_n,\, w_n,\, \varepsilon_n}$ be the upper approximation for L constructed in Theorem 8. By the triangle inequality,

$$\begin{aligned}
&\sup\left\{\left|\frac{1}{\#L}\sum_{\{i:x_i\in L\}} y_i - \mathbb{E}\left[Y\middle|X\in L\right]\right| : L\in\mathcal{L}\right\}\\
&\quad\le \sup\left\{\left|\frac{1}{\#L}\sum_{\{i:x_i\in L\}} y_i - \frac{1}{\#R(L)}\sum_{\{i:x_i\in R(L)\}} y_i\right| : L\in\mathcal{L}\right\}\\
&\qquad + \sup\left\{\left|\frac{1}{\#R}\sum_{\{i:x_i\in R\}} y_i - \mathbb{E}\left[Y\middle|X\in R\right]\right| : R\in\mathcal{R}_{s_n,\, w_n,\, \varepsilon_n},\ \#R\ge k\right\}\\
&\qquad + \sup\left\{\left|\mathbb{E}\left[Y\middle|X\in R(L)\right] - \mathbb{E}\left[Y\middle|X\in L\right]\right| : L\in\mathcal{L}\right\}
\end{aligned}$$

We can now proceed to bound each term individually. Starting with the last one, we note that because $L \subseteq R$ and $y_i \in [-M/2,\, M/2]$,

$$\begin{aligned}
&\left|\mathbb{E}\left[Y\middle|X\in R(L)\right] - \mathbb{E}\left[Y\middle|X\in L\right]\right|\\
&\quad=\left|\frac{\mu\left(R(L)\right)-\mu\left(L\right)}{\mu\left(R(L)\right)}\cdot\left(\mathbb{E}\left[Y\middle|X\in\{R(L)-L\}\right]-\mathbb{E}\left[Y\middle|X\in L\right]\right)\right|\\
&\quad\le M\,\frac{\mu\left(R(L)\right)-\mu\left(L\right)}{\mu\left(R(L)\right)},
\end{aligned}$$

and by Theorem 8, we know that with probability tending to 1

$$\sup\left\{\frac{\mu\left(R(L)\right)-\mu\left(L\right)}{\mu\left(R(L)\right)} : L\in\mathcal{L}\right\}\le\left(e^{\varepsilon_n}-1\right).$$

We can bound the first term similarly:

$$\left|\frac{1}{\#L}\sum_{\{i:x_i\in L\}} y_i - \frac{1}{\#R}\sum_{\{i:x_i\in R\}} y_i\right|\le M\,\frac{\#R(L)-\#L}{\#R(L)},$$

and by Theorem 8,

$$\sup\left\{\frac{\#R(L)-\#L}{\#R(L)} : L\in\mathcal{L}\right\}\le M\left(1+\eta\right)\left(3\varepsilon_n+3\sqrt{\frac{2\log\left(\mathcal{R}\right)}{\#L}}\right)$$

with probability tending to 1. Finally, conditionally on $\#R$, then mean of the y_i over R is sub-Gaussian with parameter $\sigma^2 = M/\#R$, and so by

Hoeffding's inequality, with probability at least $1-\delta$,

$$\sup\left\{\left|\frac{1}{\#R}\sum_{\{i:x_i\in R\}} y_i - \mathbb{E}\left[Y\middle|X\in R\right]\right| R\in\mathcal{R}_{s_n,\,w_n,\,\varepsilon_n},\ \#R\geq k\right\}$$
$$\leq M\sqrt{\frac{2\log\left(2|\mathcal{R}|/\delta\right)}{k}}.$$

Combining all these bounds together and setting $\eta = 6/\left(4\sqrt{2}\right) - 1$ yields (31).

Proof of Lemma 10. This bound is an adaptation of a result due to Chernoff [12] and Hoeffding [22], stated below for convenience.

PROPOSITION (Chernoff-Hoeffding). *Let Q be a binomial $(n,\,p)$ random variable. Then*

$$\mathbb{P}\left[\frac{Q}{n}\geq p+\varepsilon\right]\leq\left(\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right)^{n},$$
$$\mathbb{P}\left[\frac{Q}{n}\leq p-\varepsilon\right]\leq\left(\left(\frac{p}{p-\varepsilon}\right)^{p-\varepsilon}\left(\frac{1-p}{1-p+\varepsilon}\right)^{1-p+\varepsilon}\right)^{n}.$$

Let Q_n be a binomial $(n, m/n)$ random variable; it is well known that Q_n converges in distribution to Z as $n\to\infty$. The desired result follows by applying the Chernoff-Hoeffding bound to the limit of Q_n:

$$\mathbb{P}\left[Q_n\geq m+\Delta\right]\leq\left(\frac{m}{m+\Delta}\right)^{m+\Delta}\left(\frac{n-m}{n-m-\Delta}\right)^{n-m-\Delta}.$$

Now, we can verify by calculus that

$$\frac{1}{1+x}\leq e^{-x+\frac{x^2}{2}}$$

for all $x\geq 0$; thus

$$\left(\frac{m}{m+\Delta}\right)^{m+\Delta}\leq\exp\left[-\left(\frac{\Delta}{m}+\frac{\Delta^2}{2m^2}\right)(m+\Delta)\right].$$

Meanwhile, we can then verify that

$$\lim_{n\to\infty}\left(\frac{n-m}{n-m-\Delta}\right)^{n-m-\Delta}=e^{\Delta},$$

and so

$$\begin{aligned}\mathbb{P}\left[Z \geq m + \Delta\right] &= \lim_{n\to\infty} \mathbb{P}\left[Q_n \geq m + \Delta\right] \\ &\leq \exp\left[-\frac{\Delta^2}{2m}\left(1 - \frac{\Delta}{m}\right)\right].\end{aligned}$$

As to the second inequality, we know that

$$\mathbb{P}\left[Q_n \geq m - \Delta\right] \leq \left(\frac{m}{m-\Delta}\right)^{m-\Delta}\left(\frac{n-m}{n-m+\Delta}\right)^{n-m+\Delta},$$

and can check that, for all $-0.4 \leq x \leq 0$,

$$\frac{1}{1+x} \leq e^{-x+\frac{x^2}{2}-\frac{x^3}{3}}.$$

By the same argument as above, we conclude that

$$\begin{aligned}\mathbb{P}\left[Z \geq m - \Delta\right] &\leq \exp\left[\left(\frac{\Delta}{m} + \frac{\Delta^2}{2m^2} + \frac{\Delta^3}{3m^3}\right)(m-\Delta) - \Delta\right] \\ &\leq \exp\left[-\frac{\Delta^2}{2m} - \frac{\Delta^3}{6m^2} - \frac{\Delta^4}{3m^2}\right],\end{aligned}$$

which is in fact stronger than the desired result.

Proof of Corollary 11. Since $\#R \sim \text{Poisson}\left(n\mu\left(R\right)\right)$, we can apply Lemma 10 to both the left and right tails of the distribution and applying a union bound over all R with $R \in \mathcal{R}_{s,\,w,\,\varepsilon}$ and $\mu\left(R\right) \geq e^{-\varepsilon}w$ to show that

$$\mathbb{P}\left[\mathcal{A}^{\delta}_{n,\ s,\,w,\,\varepsilon}\right] \geq 1 - \delta\left(1 + \frac{1}{2}\exp\left[\sqrt{\frac{2\log\left(\frac{2\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|}{\delta}\right)^3}{n\mu\left(R\right)}}\right]\right).$$

Thanks to the condition (33), the multiplicative error term is asymptotically negligible, and so we recover the desired conclusion.

Proof of Lemma 12. Because $\mu\left(R\right) \geq w$, we know by Theorem 4 that there are rectangles

$$R_-,\ R_+ \in \mathcal{R}_{s,\,w,\,\varepsilon}$$

such that

$$\begin{aligned}&R_- \subseteq R \subseteq R_+, \text{ and} \\ &e^{-\varepsilon}\mu\left(R_+\right) \leq \mu\left(R\right) \leq e^{\varepsilon}\mu\left(R_-\right).\end{aligned}$$

Because $R \subseteq R_+$ we find that

$$\begin{aligned} \#R &\leq \#R_+ \\ &\leq n\mu\left(R_+\right) + \sqrt{2n\mu\left(R_+\right)\log\left(\frac{2\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|}{\delta}\right)} \\ &\leq e^{\varepsilon}\mu\left(R\right) + e^{\varepsilon/2}\sqrt{2n\mu\left(R\right)\log\left(\frac{2\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|}{\delta}\right)}, \end{aligned}$$

where the second inequality followed by Corollary 11. We also get an analogous lower bound

$$\begin{aligned} \#R &\geq \#R_- \\ &\geq n\mu\left(R_-\right) - \sqrt{2n\mu\left(R_-\right)\log\left(\frac{2\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|}{\delta}\right)} \\ &\geq e^{-\varepsilon}\mu\left(R\right) - \sqrt{2n\mu\left(R\right)\log\left(\frac{2\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|}{\delta}\right)}. \end{aligned}$$

Notice that, by construction, $\mu\left(R_-\right) \geq e^{-\varepsilon}s$, and so the hypotheses required to apply Corollary 11 are met.

Proof of Corollary 13. Combining Lemma 12 with our assumptions (28) and (29), we find that,

$$\lim_{n\to\infty}\mathbb{P}\left[\frac{\inf\left\{\#R : \mu\left(R\right) = w_n\right\}}{k_n} > 1 - \frac{\eta}{2}\right] = 0,$$

or, in other words, all rectangles of size w_n must have at least $(1-\eta/2)k_n$ points in them. Thus, we conclude that, with probability converging to 1, any rectangle with k_n points must have size at least w_n.

Proof of Lemma 15. We begin by showing how to construct leaf sets $\{L_j\}$ for which the desired conclusion holds. At a high level, to get such leaves L_j it suffices to recursively select a random feature j, do a 20%–80% split along that feature, and then recurse on the smaller leaf.

More specifically, for any $0 < \alpha < 0.5$, we study α-random partitions generated as follows. Given $d_n = n^r$ as assumed in Lemma 15, we set

$$s = s_n := \left\lfloor \log\left(\frac{\log^3(n)}{n}\right)(\log\alpha)^{-1}\right\rfloor \text{ and } k = k_n := \left\lfloor n\alpha^{s_n}\left(1 - \frac{\log n}{2\sqrt{n}}\right)\right\rfloor.$$

Then $k \geq \log^3(n)(1 - \frac{\log n}{2\sqrt{n}}) - 1$, so Assumption 2 is met. Given an index set $S \subset \{1, \ldots, d\}$ with $|S| = s$, define the partition V_S as follows: recursively partition $[0,1]^d$ with splits on those axes having indices in S, using some splitting rule such that each child-node contains at least a fraction α of the data points in its parent-node, with one child-node containing exactly a fraction α up to rounding.

Given this construction, each terminal node L has at least

$$|L| \;\geq\; \alpha^s n_{obs} \;\geq\; \frac{k}{1 - \frac{\log n}{2\sqrt{n}}} \frac{n_{obs}}{n}$$

data points, so $\mathbb{E}\left[|L|\right] \geq k$. Moreover, on the event

$$\mathcal{B}_n \;:=\; \Big\{ |n_{obs} - n| \leq \log(n)\sqrt{n}/2 \Big\}$$

V_S is an $\{\alpha, k\}$-valid partition since $\mathcal{B}_n$ implies $|L| \geq k$. The above construction provides for one terminal node L that further satisfies

$$\begin{aligned}|L| &\leq \alpha^s n_{obs} + \sum_{i=0}^{s-1} \alpha^i \\ &\leq \frac{k+1}{1 - \frac{\log n}{2\sqrt{n}}} \frac{n_{obs}}{n} + \frac{1}{1-\alpha} \\ &\leq (k+1)\Big(1 + \frac{\log n}{\sqrt{n}}\Big)\frac{n_{obs}}{n} + 2;\end{aligned}$$

hence $\mathbb{E}\left[|L|\right] \leq k + 4$ and $L \leq k + 4$ on $\mathcal{B}_n$, for n large enough.

For each s-combination S of $\{1, \ldots, d\}$ construct V_S as described above. Then for each of the $N := \binom{d}{s}$ s-combinations the resulting partition has one terminal node with the above properties. Denote these N terminal nodes by $L_1, \ldots, L_N$. So for $i = 1, \ldots, N$:

$$\begin{aligned}&k \leq \mathbb{E}\left[|L_i|\right] \leq k + 4 \quad \text{and} \\ &k \leq |L_i| \leq k + 4 \quad \text{on } \mathcal{B}_n.\end{aligned} \tag{45}$$

Moreover, if $i \neq j$ then the splits in L_i and L_j occur on axes that differ in at least one index. Since X has independent marginals we get

$$\mathbb{E}\left[|L_i \cap L_j|\right] \;\leq \alpha \mathbb{E}\left[|L_i|\right] + 1 \;\leq\; \alpha k + 3. \tag{46}$$

Since the overlap between the L_j is not very large, we might hope that the maximum of the $\widetilde{T}_j$ would be of comparable size to the maximum of N independent Gaussian random variables with variance $|L_j|^{-1}M^2/4$. The following sub-result shows that this intuition is valid, at least to within a factor $\sqrt{1-\alpha}$.

PROPOSITION. *Assume that $d_n = n^r$ for some $r > 0$. Then, given the nodes L_j and statistics $\widetilde{T}_j$ as constructed above,*

$$\lim_{n\to\infty} \mathbb{P}\left[\max_{j=1,\ldots,N} \widetilde{T}_j \;\geq\; (1-\varepsilon_n)\, M \sqrt{\frac{1-\alpha}{2}} \sqrt{\frac{\log(n)\log(d)}{\log\left(\alpha^{-1}\right)}}\, \frac{1}{\sqrt{k}}\right] \;=\; 1 \tag{47}$$

whenever $\varepsilon_n \frac{\log n}{\log\log n} \to \infty$.

The claim from Lemma 15 follows directly from (47) by noting that, following our construction, an $\tilde{\alpha}$-random partition is always an $\{\alpha,\, k\}$-valid partition for any $\tilde{\alpha} \geq \alpha$. Thus, if we know that $\alpha \leq 0.2$, we can apply (47) for 0.2-random partitions, thus yielding the desired bound.

We now proceed to proving the sub-result. For simplicity, we take $M = 2$. Independence of X and $\widetilde{Y}$ implies that the $\sqrt{|L_j|}\, \widetilde{T}_j$ are standard normal. For $j \neq m$:

$$\begin{aligned}
&\operatorname{Cov}\left[\sqrt{|L_j|}\, \widetilde{T}_j,\, \sqrt{|L_m|}\, \widetilde{T}_m\right] \\
&= \mathbb{E}\left[\frac{1}{\sqrt{|L_j||L_m|}}\right] \sum_{(i,i'):X_i\in L_j, X_{i'}\in L_m} \mathbb{E}\left[\widetilde{Y}_i \widetilde{Y}_{i'} | X\right] \\
&= \mathbb{E}\left[\frac{|L_j \cap L_m|}{\sqrt{|L_j||L_m|}}\right] \\
&\leq \mathbb{E}\left[\frac{|L_j \cap L_m|}{k} 1_{\mathcal{B}_n}\right] \;+\; \mathbb{E}\left[1_{\mathcal{B}_n^c}\right] \qquad \text{by (45)} \\
&\leq \alpha + \frac{3}{k_n} \;+\; \mathbb{P}\left[\mathcal{B}_n^c\right] \qquad \text{by (46)} \\
&\leq \alpha_n := \alpha + \frac{4}{k_n} \qquad \text{by Lemma 10.}
\end{aligned}$$

Now let $Z_1, \ldots, Z_{N+1}$ be i.i.d. $N(0,1)$ and for $j = 1, \ldots, N$ set

$$\widetilde{Z}_j \;:=\; \sqrt{1-\alpha_n}\, Z_j + \sqrt{\alpha_n}\, Z_{N+1}.$$

The $\widetilde{Z}_j$ are standard normal with $\operatorname{Cov}[\widetilde{Z}_j,\, \widetilde{Z}_m] = \alpha_n$ if $j \neq m$. Using $\sqrt{|L_j|/k} \leq 1 + 2/k$ on $\mathcal{B}_n$ by (45) and employing Corollary 4.2.3 to the normal approximation lemma of Leadbetter et al. [25], we get for every

$u > 0$:

$$\begin{aligned}\mathbb{P}\left[\max_{j=1,\ldots,N} \sqrt{k}\,\widetilde{T}_j \le u\right] &\le \mathbb{P}\left[\max_{j=1,\ldots,N} \sqrt{|L_j|}\,\widetilde{T}_j \le (1+\frac{2}{k})u\right] + \mathbb{P}\left[\mathcal{B}_n^c\right]\\ &\le \mathbb{P}\left[\max_{j=1,\ldots,N} \widetilde{Z}_j \le (1+\frac{2}{k})u\right] + \mathbb{P}\left[\mathcal{B}_n^c\right]\\ &= \mathbb{P}\left[\max_{j=1,\ldots,N} Z_j \le \frac{(1+\frac{2}{k})u - \sqrt{\alpha_n}Z_{N+1}}{\sqrt{1-\alpha_n}}\right] + o(1)\end{aligned}$$

Setting $u = u_n := \sqrt{2(1-\alpha)}(1-2\varepsilon_n)\sqrt{\frac{\log(n)\log(d)}{\log(\alpha^{-1})}}$ with $\varepsilon_n \to 0$, $\varepsilon_n \frac{\log n}{\log\log n} \to \infty$, observing

$$\mathbb{P}\left[\mathcal{C}_n := \left\{\frac{2u_n}{k_n} - \sqrt{\alpha_n}Z_{N+1} \le \sqrt{2(1-\alpha)}\,\varepsilon_n\sqrt{\frac{\log(n)\log(d)}{\log(\alpha^{-1})}}\right\}\right] \to 1,$$

and using the normal tail bound shows that the above expression is not larger than

$$\begin{aligned}&\left(\mathbb{P}\left[Z_1 \le \sqrt{2\frac{1-\alpha}{1-\alpha_n}}(1-\varepsilon_n)\sqrt{\frac{\log(n)\log(d)}{\log(\alpha^{-1})}}\right]\right)^N + \mathbb{P}\left[\mathcal{C}_n^c\right] + o(1)\\ &\le \left(1 - \frac{\exp\Big\{-\frac{1}{2}2\big(1-\frac{\varepsilon_n}{2}\big)^2\frac{\log(n)\log(d)}{\log(\alpha^{-1})}\Big\}}{2\sqrt{\frac{\log(n)\log(d)}{\log(\alpha^{-1})}}}\right)^N + o(1)\\ &\qquad \text{since } \sqrt{\frac{1-\alpha}{1-\alpha_n}} \le 1 + \frac{4}{k_n} \quad\text{and}\quad \frac{4}{k_n} \le \frac{\varepsilon_n}{2}\\ &\le \exp\left\{-N\frac{\exp\Big\{\big(-1+\frac{\varepsilon_n}{2}\big)\frac{\log(n)\log(d)}{\log(\alpha^{-1})}\Big\}}{2\sqrt{\log(n)\log(d)}}\sqrt{\log(\alpha^{-1})}\right\} + o(1)\\ &\le \exp\left\{-\frac{\exp\Big\{\frac{\log n}{\log\alpha^{-1}}\Big(\frac{\varepsilon_n}{2}\log(d) - 3(r+1)\log\log n\Big)\Big\}}{2\sqrt{\log(n)\log(d)}}\sqrt{\log(\alpha^{-1})}\right\} + o(1)\\ &= o(1)\end{aligned}$$

by (48) and since $\frac{\varepsilon_n}{2}\log(d) - 3(r+1)\log\log n \to \infty$.

We used

$$\begin{aligned}\log N = \log\binom{d}{s} &\geq \log\Big(\frac{d-s}{s}\Big)^s \\ &\geq s\log d - s\log(2s) \qquad \text{once } \frac{d}{2} \geq s \\ &\geq \frac{\log(n)\log(d)}{\log(\alpha^{-1})} - \frac{3(r+1)}{\log(\alpha^{-1})}\log(n)\log\log n \qquad (48)\end{aligned}$$

and for later use we note

$$\log N \leq \log d^s \leq \frac{\log(n)\log(d)}{\log(\alpha^{-1})} = O(\log^2 n). \qquad (49)$$

Proof of Lemma 16. One readily checks that

$$\mathbb{E}\Big[\exp\Big\{t\big(Y - Y|Z|\big)\Big\}\Big] = \exp\Big(\frac{1}{2}t^2 + t\Big)\Phi(-t) + \exp\Big(\frac{1}{2}t^2 - t\Big)\Phi(t),$$

where Φ is the cdf of Z. Using the expansion

$$\Phi(t) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\sum_{k=0}^{\infty}\frac{(-1)^k t^{2k+1}}{k!2^k(2k+1)},$$

we find that

$$\begin{aligned}&e^t\Phi(-t) + e^{-t}\Phi(t) \\ &= \cosh(t) - \sqrt{\frac{2}{\pi}}\sinh(t)\sum_{k=0}^{\infty}\frac{(-1)^k t^{2k+1}}{k!2^k(2k+1)} \\ &= \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} - \sqrt{\frac{2}{\pi}}\left(\sum_{k=0}^{\infty}\frac{t^{2k+1}}{(2k+1)!}\right)\left(\sum_{k=0}^{\infty}\frac{(-1)^k t^{2k+1}}{k!2^k(2k+1)}\right) \\ &= 1 + \left(\frac{1}{2} - \sqrt{\frac{2}{\pi}}\right)t^2 + \frac{1}{4\,!}t^4 + O(t^6).\end{aligned}$$

Lemma 16 follows since

$$\frac{1}{4\,!} < \left(\frac{1}{2} - \sqrt{\frac{2}{\pi}}\right)^2 / \, 2\,!.$$

Proof of Corollary 17. For simplicity, we take $M = 2$, and so $Y_i \in \pm 1$ and $\mathrm{Var}[\widetilde{Y}_i] = 1$. At a high level, Lemma 16 implies that the maximum of the N standardized differences $\widetilde{T}_j - T_j$ will be of smaller magnitude than that of N i.i.d. normal random variables with variance $2\big(1 - \sqrt{2/\pi}\big)$, and since the latter expression is smaller than $4/5$ it follows that the approximation error is smaller than the lower bound given in (39). In more detail, a standard argument using Markov's inequality gives for any $t, v > 0$:

$$\begin{aligned}&\mathbb{P}\left[\sqrt{|L_j|}p\widetilde{T}_j - T_j > v\middle|X\right]\\&\qquad\le \left(\mathbb{E}\left[\exp\left\{t\left(Y_1|Z_1| - Y_1\right)\right\}\right]\right)^{|L_j|}\exp\left\{-\sqrt{|L_j|}\,tv\right\}\\&\qquad\le \exp\left\{-\frac{v^2}{4\left(1-\sqrt{\frac{2}{\pi}}\right)}\right\}\end{aligned}$$

by Lemma 16, provided that $t := v/\big(2\sqrt{|L_j|}\big(1 - \sqrt{2/\pi}\big)\big)$ is small enough. Setting $v = v_n := \varepsilon\sqrt{\log N}$ for some fixed $\varepsilon > 0$ and recalling $\min_j |L_j| \ge k_n \ge \log^3 n$ on $\mathcal{B}_n$ by (45), we see with (49) that

$$\max_j \frac{v_n}{2\sqrt{|L_j|}\left(1-\sqrt{2/\pi}\right)} \to 0 \quad \text{on } \mathcal{B}_n.$$

Together with the fact that $\mathcal{B}_n$ and X are independent we get

$$\begin{aligned}&\mathbb{P}\left[\max_{j=1,\dots,N}\sqrt{k_n}\left(\widetilde{T}_j - T_j\right) > \varepsilon\sqrt{\log N}\right]\\&\qquad\le N\max_j \mathbb{E}\left[\mathbb{P}\left[\sqrt{|L_j|}\left(\widetilde{T}_j - T_j\right) > v_n\middle|X\right]1(\mathcal{B}_n)\right] \;+\; \mathbb{P}\left[\mathcal{B}_n^c\right]\\&\qquad\le \exp\left\{(\log N)\left(1 - \frac{\varepsilon^2}{4\left(1-\sqrt{2/\pi}\right)}\right)\right\} \;+\; o(1),\end{aligned}$$

which converges to zero provided that $\varepsilon^2 > 4\big(1 - \sqrt{2/\pi}\big) \approx 0.81$; thus, (41) follows from (49).

Proof of Lemma 18. Recall for a tree T with corresponding partition S, we write $\mathcal{L}(S)$ for the set of leaves determining T. To establish (43), we need to define an approximation T_+ to T, defined as follows. For every leaf $L \in \mathcal{L}(S)$, let $L_+ \in \mathcal{R}_{s,\,w,\,\varepsilon}$ be a super-set approximation to L as constructed in Theorem 4; for convenience, write $\mathcal{L}_+(S)$ for the set of all these L_+. The tree approximation is then

$$\begin{aligned} &T_+ : [0,\, 1]^d \to [0,\, 1], \\ &T_+(x) = \frac{1}{\left|\{L_+ \in \mathcal{L}_+ : x_i \in L_+(S)\}\right|} \sum_{\{L_+ \in \mathcal{L}_+ : x_i \in L_+(S)\}} \widehat{\mathbb{E}}\left[Y \,\middle|\, X \in L_+\right]. \end{aligned}$$

The reason we need the extra denominator is that the fattened leaves L_+ no longer form a partition, so some test points may fall into multiple leaves. Now, let

$$\widehat{\Gamma}(T) = \{i = 1,\, ...,\, n_{obs} : \left|\{L_+ \in \mathcal{L}_+ : x \in L_+(S)\}\right| \geq 2\}$$

be the set of training examples that end up in many leaves because of the fattening of the leaves. By Theorem 8, we can verify that, with probability tending to 1, simultaneously for all valid trees T,

$$\begin{aligned} \frac{1}{n_{obs}} \left|\widehat{\Gamma}(T)\right| &\leq \frac{1}{n_{obs}} \sum_{L \in \mathcal{L}(S)} \frac{\#L_+ - \#L}{\#L} \#L \\ &\leq \sup_{L \in \mathcal{L}(S)} \frac{\#L_+ - \#L}{\#L} \\ &= 3 \sqrt{2 \frac{\log(n) \log(d)}{\log\left((1-\alpha)^{-1}\right)}} \frac{1}{\sqrt{k}} \left(1 + o(1)\right). \end{aligned}$$

Meanwhile, by the proof of Corollary 11, we find that, again with probability tending to 1 and simultaneously for all valid trees T,

$$\begin{aligned} &\sup\left\{\left|T_+(x_i) - T(x_i)\right| : i = 1,\, ...,\, n_{obs},\, i \not\in \widehat{\Gamma}(T)\right\} \\ &\leq 3M \sqrt{2 \frac{\log(n) \log(d)}{\log\left((1-\alpha)^{-1}\right)}} \frac{1}{\sqrt{k}} \left(1 + \eta + o(1)\right), \end{aligned}$$

where η is an arbitrary positive constant. Combining these arguments and applying them simultaneously to the two trees $T^{(1)}$ and $T^{(2)}$ defined in the hypothesis, we find that with probability tending to 1 for all possible choices of $T^{(1)}$ and $T^{(2)}$,

$$\begin{aligned}
&\left|\frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\left(T^{(1)}(x_i)-T^{(2)}(x_i)\right)^2-\frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\left(T^{(1)}_{+}(x_i)-T^{(2)}_{+}(x_i)\right)^2\right| \\
&\quad\le \frac{M^2}{n_{obs}}\left|\widehat{\Gamma}\left(T^{(1)}\right)\right|+\frac{M^2}{n_{obs}}\left|\widehat{\Gamma}\left(T^{(2)}\right)\right| \\
&\qquad\quad + 2\cdot 3M^2\sqrt{2\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}}\frac{1}{\sqrt{k}}\left(1+\eta+o(1)\right) \\
&\quad\le 12\sqrt{2\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}}\frac{1}{\sqrt{k}}\left(1+\eta+o(1)\right).
\end{aligned}$$

Meanwhile, by a similar argument, we can establish that, under the same high-probability conditions,

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(T^{(1)}\left(x\right)-T^{(2)}\left(x\right)\right)^2\right]-\mathbb{E}\left[\left(T^{(1)}_{+}\left(x\right)-T^{(2)}_{+}\left(x\right)\right)^2\right]\right| \\
&\qquad = o\left(M^2\sqrt{\frac{\log(n)\log(d)}{k}}\right).
\end{aligned}$$

Finally, we consider the quantity

$$\frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\left(T^{(1)}_{+}\left(x_i\right)-T^{(2)}_{+}\left(x_i\right)\right)^2-\mathbb{E}\left[\left(T^{(1)}_{+}\left(x\right)-T^{(2)}_{+}\left(x\right)\right)^2\right]$$

Now, for any fixed choices $T^{(1)}_{+}$ and $T^{(2)}_{+}$, conditionally on n_{obs},

$$\sum_{i=1}^{n_{obs}}\left(T^{(1)}_{+}\left(x_i\right)-T^{(2)}_{+}\left(x_i\right)\right)^2$$

is a sub-Gaussian random variable with parameter $1/n_{obs}$. Moreover, there are only $\mathcal{R}^2_{s,\,w,\,\varepsilon}$ possible distinct choices for $T^{(1)}_{+}$ and $T^{(2)}_{+}$. Thus, by Hoeffding's inequality, we find that with probability tending to 1,

$$\begin{aligned}
&\frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}}\left(T^{(1)}_{+}\left(x_i\right)-T^{(2)}_{+}\left(x_i\right)\right)^2-\mathbb{E}\left[\left(T^{(1)}_{+}\left(x\right)-T^{(2)}_{+}\left(x\right)\right)^2\right] \\
&\qquad = M^2\sqrt{2\frac{\log\left(\mathcal{R}^2_{s,\,w,\,\varepsilon}\right)}{n_{obs}}}\left(1+o(1)\right) \\
&\qquad = o\left(M^2\sqrt{\frac{\log(n)\log(d)}{k}}\right).
\end{aligned}$$

simultaneously for all possible choices of $T_+^{(1)}$ and $T_+^{(2)}$. Thus, the desired bound holds, noting that $12\sqrt{2} < 17$.

Department of Statistics
Stanford University
Stanford, CA-94305
E-mail: swager@stanford.edu
E-mail: gwalther@stanford.edu