Stability of Random Forests and Coverage of Random-Forest Prediction Intervals

Yan Wang
Department of Mathematics
Wayne State University
Detroit, MI 48202
`wangyan@wayne.edu`

Huaiqing Wu, Dan Nettleton
Department of Statistics
Iowa State University
Ames, IA 50011
`{isuhwu,dnett}@iastate.edu`

Abstract

We establish stability of random forests under the mild condition that the squared response (Y^2) does not have a heavy tail. In particular, our analysis holds for the practical version of random forests that is implemented in popular packages like `randomForest` in R. Empirical results show that stability may persist even beyond our assumption and hold for heavy-tailed Y^2. Using the stability property, we prove a non-asymptotic lower bound for the coverage probability of prediction intervals constructed from the out-of-bag error of random forests. With another mild condition that is typically satisfied when Y is continuous, we also establish a complementary upper bound, which can be similarly established for the jack-knife prediction interval constructed from an arbitrary stable algorithm. We also discuss the asymptotic coverage probability under assumptions weaker than those considered in previous literature. Our work implies that random forests, with its stability property, is an effective machine learning method that can provide not only satisfactory point prediction but also justified interval prediction at almost no extra computational cost.

1 Introduction

Random forests (RFs) is a successful machine learning method that serves as a standard approach to tabular data analysis and has good predictive performance [10, 7]. However, there is a big gap between the empirical effectiveness of RFs and the limited understanding of its properties. Most known theoretical results are established for variants of RFs not necessarily used in practice [5, 25, 22, 14, 30]. For the RF version implemented in packages like `randomForest` in R [21], little is known without strong assumptions [6, 26, 35]; RFs is notoriously difficult to analyze as a greedy algorithm. Here we show an important property for the RF used in practice (as well as for other variants) under realistic conditions.

1.1 Stability of random forests

The first main contribution of this work establishes the stability condition for the RF.

Theorem 1 (Stability of random forests, informal). *For independent and identically distributed (iid) training data points $(X_i, Y_i), i \in \{1, \ldots, n\} \equiv [n]$ and a test point (X, Y), if the squared response Y^2 does not have a heavy tail, then the RF predictor RF_B and any out-of-bag (OOB) predictor $\mathsf{RF}_B^{\setminus i}$ predict similar values, i.e.,*

$$\mathbb{P}\left(\left| \mathsf{RF}_B(X) - \mathsf{RF}_B^{\setminus i}(X) \right| > \varepsilon_{n,B} \right) \leqslant \nu_{n,B}, \tag{1}$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

where RF_B *results from the aggregation of all* B *base tree predictors, while* $\mathsf{RF}_B^{\setminus i}$ *only those with the point* (X_i, Y_i) *excluded in training;* $\varepsilon_{n,B}$ *and* $\nu_{n,B}$ *are small numbers depending on* n *and* B.

This result is referred to as the stability of the RF because it indicates that no single training point is extremely important in determining RF_B in a probabilistic sense. Theorem 1 relies on a recent important work that establishes the absolute stability (see below for a precise definition) of general bagged algorithms with bounded outputs [28]. We take advantage of the fact that the range of the RF output is conditionally dependent upon the maximal and minimal values of Y in the training set, and then we show in theory that the stability property of the RF is possible even if Y is marginally unbounded. To our knowledge, this is the first stability result established for the RF.

The technique used in our analysis requires that Y^2 not have a heavy tail (to make $\varepsilon_{n,B}$ and $\nu_{n,B}$ small). Though arguably already mild, we conjecture that this condition might be further relaxed. As shown below, numerical evidence suggests that the light-tail assumption may not be necessary for RF stability, which could hold even when Y follows a heavy-tail distribution like the Cauchy distribution.

1.2   Random-forest prediction intervals

Stability is a crucial property of a learning algorithm. For example, stability has a deep connection with the generalization error of an algorithm [9, 19, 23]. Moreover, stability also turns out to be important in distribution-free predictive inference. In particular, an algorithm being stable justifies the jackknife prediction interval (PI), which otherwise has no coverage guarantee [3].

In this work, we show that stability makes it possible to construct a PI with guaranteed coverage from the OOB error of the RF. The OOB error is defined as $R_i = |Y_i - \mathsf{RF}_B^{\setminus i}(X_i)|$, $i \in [n]$. A main reason why such a PI is appealing is that R_i can be obtained almost without extra effort. For example, a one-shot training using the R package `randomForest` gives us an RF predictor RF_B and all n OOB predictions $\mathsf{RF}_B^{\setminus i}(X_i)$. So, from the computational point of view, a convenient way to construct a PI for a test point (X, Y) is of the form "$\mathsf{RF}_B(X) \pm$ proper quantile of $\{R_i\}$" [17, 35].

The second main contribution of this work constructs such PIs and theoretically proves, under mild conditions, the non-asymptotic lower and upper bounds for the coverage probability.

Theorem 2 (Coverage lower bound, informal)**.** *Under the same assumptions as in Theorem 1, and for* $\alpha \in (0, 1)$ [1]*, we have the following lower bound of coverage probability:*

$$\mathbb{P}\left(|Y - \mathsf{RF}_B(X)| \leqslant \text{the } \lceil(n+1)(1-\alpha)\rceil\text{-th smallest } R_i + \varepsilon_{n,B}\right) \geqslant 1 - \alpha - O(\sqrt{\nu_{n,B}}),$$

where $\lceil \cdot \rceil$ *is the ceiling function. Big O and other related notations are used in the usual way.*

Theorem 3 (Coverage upper bound, informal)**.** *If we further assume that* Y *is continuous, resulting in distinct prediction errors, then we also have the following upper bound:*

$$\mathbb{P}\left(|Y - \mathsf{RF}_B(X)| \leqslant \text{the } \lceil(n+1)(1-\alpha)\rceil\text{-th smallest } R_i - \varepsilon_{n,B}\right) \leqslant 1 - \alpha + \frac{1}{n+1} + O(\sqrt{\nu_{n,B}}).$$

As we detail below, the PIs we provide coverage guarantees for are neither the jackknife-with-stability interval discussed in [3], nor the jackknife+-after-bootstrap interval established in [18]. In our context, constructing the former needs n leave-one-out (LOO) predictors (rather than n OOB predictors), i.e., n additional RFs with each built on a training set of size $n - 1$. Constructing the latter needs the explicit information of each $\mathsf{RF}_B^{\setminus i}(\cdot)$ rather than the OOB prediction $\mathsf{RF}_B^{\setminus i}(X_i)$ for each X_i only. Both these methods require additional, sometimes extensive, computation given current popular packages. In contrast, our results are operationally more convenient. After one-shot training, we obtain not only a point predictor $\mathsf{RF}_B(\cdot)$, but also a valid interval predictor at almost no extra cost. Under reasonable conditions, our results indicate that by slightly inflating (or deflating) the PI constructed from the $\lceil(n+1)(1-\alpha)\rceil$-th smallest R_i, the coverage probability is guaranteed not to decrease (or increase) too much from the desired level of $1 - \alpha$. In fact, many numerical results, such as those in [17, 35], suggest that

$$\mathbb{P}\left(|Y - \mathsf{RF}_B(X)| \leqslant \text{the } \lceil(n+1)(1-\alpha)\rceil\text{-th smallest } R_i\right) \approx 1 - \alpha.$$

Motivated by this fact, we further establish an asymptotic result of coverage for such PIs.

---

[1] When $\alpha \in (0, 1/(n+1))$, we follow the convention that the $(n+1)$-th smallest R_i is ∞.

2

Theorem 4. *(Asymptotic coverage, informal) In addition to the conditions in the above theorems, also suppose the prediction error $|Y - \mathsf{RF}_B(X)|$ is continuous, and its cumulative distribution function (CDF) does not change too drastically for all sufficiently large n. Then*

$$\mathbb{P}\left(|Y - \mathsf{RF}_B(X)| \leqslant \textit{the } \lceil (n+1)(1-\alpha) \rceil \textit{-th smallest } R_i \right) \to 1 - \alpha \textit{ as } n \to \infty.$$

In [35], this asymptotic coverage was proved based on stronger assumptions. In particular, the true model is assumed to be additive such that "$Y = f_0(X) + $ noise" with the zero-mean noise independent of X, and $\mathsf{RF}_B(X)$ is assumed to converge to $f_0(X)$ in probability. We do not require RF_B to converge to anything in any sense when $n \to \infty$. Technically, we need the family of prediction error CDFs be uniformly equicontinuous.

Based on our results, the RF seems to be the only one, among existing popular machine learning algorithms, that can provide both point and interval predictors with justification in such a convenient way. This makes the RF appealing, especially for tasks where the computational cost is a concern.

It is also worth noting that the upper-bound result is of interest in its own right. It can be generalized to jackknife PIs that are constructed from any stable algorithm; the result serves as a complement to the lower bounds established previously [3, 18].

Summarizing, we

- theoretically prove that the (greedy) RF algorithm is stable when Y^2 does not have a heavy tail;
- numerically show that RF stability may hold beyond the above light-tail assumption;
- construct PIs based on the OOB error with finite-sample coverage guarantees: the lower bound of coverage does not need any additional assumption beyond stability; the upper bound needs an additional assumption, which is usually satisfied when Y is continuous;
- provide the upper bound of coverage for jackknife PIs constructed from general stable algorithms, assuming distinct LOO errors; and
- prove asymptotically exact coverage for RF-based PIs under weaker assumptions than those previously considered in published work.

2 Concepts of algorithmic stability

Stability stands at the core of this work. There are different types of stability, each of which is used to assess quantitatively how stable (in some certain sense) an algorithm is with respect to small variations in training data [9, 28, 4]. In a recent work [4], robust optimization is used to enhance the stability of algorithms in classification tasks. In [28], bagging is proved to be an efficient mechanism to stabilize algorithms in regression tasks. We focus on regression here. As will be made clear, the technique used in this work relies on the fact that the RF predictor in regression results from averaging tree predictors. However, the majority vote of tree predictors is used in classification, and new ideas are needed to analyze the RF stability in this setting. For our purposes, we introduce three levels of stability from strongest to weakest. The strongest version of stability, introduced in [28], does *not* depend on the data distribution, and may be referred to as "absolute stability."

Definition 1 (Absolute stability of algorithms)**.** For any dataset consisting of $n \geqslant 2$ training points $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and any test point (X, Y), an algorithm $\mathcal{A}$ is defined to be (ε, δ)-absolutely-stable if

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_\xi \left(\left| \hat{f}(X) - \hat{f}^{-i}(X) \right| > \varepsilon \right) \leqslant \delta$$

for some $\varepsilon, \delta \geqslant 0$, where ξ denotes the possible innate randomness in the algorithm (such as the node splitting procedure in the RF) and can be seen as a random variable uniformly distributed in $[0, 1]$, $\hat{f} = \mathcal{A}(D; \xi)$ is the predictor trained on D, and $\hat{f}^{-i} = \mathcal{A}(D^{-i}; \xi)$ is the ith LOO predictor trained on D^{-i}, i.e., D without the ith point (X_i, Y_i). We might refer to the RF as both an algorithm (the learning procedure) and a predictor (the learned function) for simplicity.

Many bagged algorithms, in particular those with bounded predicted values, can achieve absolute stability with both ε and δ converging to 0, as long as n and the number of bags B go to infinity.

3

However, the predicted value of the RF is in general unbounded (for regression tasks considered in this work), and we are more interested in another type of stability, investigated in [9], and called out-of-sample stability [3]. For simplicity, we name it "stability." This notion of stability turns out to be important in validating a jackknife prediction interval.

Definition 2 (Stability of algorithms). For iid training and test data, algorithm $\mathcal{A}$ is (ε, δ)-stable if

$$\mathbb{P}_{D,X,\xi}\left(\left|\hat{f}(X) - \hat{f}^{-i}(X)\right| > \varepsilon\right) \leqslant \delta$$

for some $\varepsilon, \delta \geqslant 0$, where $D, X, \hat{f}, \hat{f}^{-i}$ are as defined above.

We will establish this type of stability for the derandomized RF defined below, where the data-generating distribution is involved. To this end, we will use the methods in [28], which aim to provide absolute stability for bagged algorithms. Technically, we use such methods to first establish the "conditional stability" of an algorithm with respect to given data.

Definition 3 (Conditional stability of algorithms). Conditional on D and X, an algorithm $\mathcal{A}$ is defined to be (ε, δ)-conditionally-stable if

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\xi|D,X}\left(\left|\hat{f}(X) - \hat{f}^{-i}(X)\right| > \varepsilon \Big| D, X\right) \leqslant \delta$$

for some $\varepsilon, \delta \geqslant 0$, where $D, X, \hat{f}, \hat{f}^{-i}$ are as defined above.

Once conditional stability is established for the derandomized RF algorithm, its stability can be consequently established by invoking

$$\mathbb{P}_{D,X,\xi}(\cdot) = \mathbb{E}_{D,X}\left[\mathbb{P}_{\xi|D,X}(\cdot|D,X)\right].$$

Stability of the derandomized RF provides the most essential ingredient for that of the practical RF, although the latter involves another type of stability, known as ensemble stability [18]. Ensemble stability justifies replacing the LOO predictor with the OOB predictor in (1). We may abuse the term "stability" in the following when the OOB, rather than the LOO, predictor is used.

3 Stability of random forests

3.1 Basics of random forests

This work mainly considers using the RF to perform regression tasks, where the response $Y \in \mathbb{R}$ can be unbounded. By construction, the RF predictor with B bags, denoted by RF_B, is a bagged algorithm with the base algorithm being a tree, and $\mathsf{RF}_B = \frac{1}{B}\sum_{b=1}^{B}\mathsf{TREE}_b$, where TREE_b is the bth tree predictor, trained on the bth bag r_b, a bootstrapped sample of the training set D. The randomness in the tree predictor TREE originates from two independent sources: innate randomness ξ in the node splitting process and resampling randomness from the bag r. For the ith point, one can define the OOB RF predictor as $\mathsf{RF}_B^{\setminus i} = \frac{1}{B_i}\sum_{b=1}^{B}\mathsf{TREE}_b \times \mathbb{I}\{i \notin r_b\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function, and $B_i = \sum_{b=1}^{B}\mathbb{I}\{i \notin r_b\}$. Define $p \equiv \mathbb{P}(i \in r)$ as the probability that the ith point is included in bag r. Then it is clear that $B_i \sim \mathrm{Binomial}(B, 1-p)$ for all i. We also denote rf and $\mathsf{rf}^{\setminus i}$ as the *derandomized* versions of RF_B and $\mathsf{RF}_B^{\setminus i}$, respectively. Precisely, $\mathsf{rf} = \mathbb{E}_{\xi,r}[\mathsf{TREE}]$ and $\mathsf{rf}^{\setminus i} = \mathbb{E}_{\xi,r}[\mathsf{TREE}|i \notin r]$. It is worth noting that, by definition, $\mathsf{RF}_B^{\setminus i} \neq \mathsf{RF}_B^{-i}$ for finite B, while $\mathsf{rf}^{\setminus i} = \mathsf{rf}^{-i}$ as the derandomized RF results from the aggregation of an infinite number of trees. Since RF predictors are averages over tree predictors, the predicted values they output, given training set D, are bounded in $[Y_{(1)}, Y_{(n)}]$, where $Y_{(1)}$ and $Y_{(n)}$ are the minimum and maximum of $\{Y_1, \ldots, Y_n\}$, respectively. We also let $Z_i = |Y_i|$ for all i, and denote the maximum as $Z_{(n)}$. As a result, we have that

$$|\mathsf{rf}^{\setminus i} - \mathsf{rf}| \leqslant Y_{(n)} - Y_{(1)} \leqslant 2Z_{(n)}. \tag{2}$$

Remark 1. This is also true for $\mathsf{RF}_B^{\setminus i}$ and RF_B for any finite B. In fact, this is a distinctive feature of the RF, *irrespective* of the node splitting rule. Other regression methods do not necessarily have such a data-dependence bound. This observation helps to establish the conditional stability of the RF.

4

Remark 2. Practically, when n is large, one might think that the bound (2) is crude. On one hand, if we look for a bound valid for any finite $n \geqslant 2$, then there is not much room for improvement for small n. On the other hand, we do expect that the *typical* stability of the RF can go beyond the finite-sample guarantee provided by (2) when n is big, which is consistent with the numerical results shown below. A more informative bound for large n is worth future investigation.

There are several quantities that are useful in establishing the RF stability; they can be calculated explicitly and are listed below. First, it is well known that

$$p \equiv \mathbb{P}(i \in r) = 1 - (1 - 1/n)^n = 1 - 1/e + O(1/n). \tag{3}$$

Actually, p is monotonically decreasing for $n \geqslant 1$. Second,

$$q \equiv -\mathrm{Cov}(\mathbb{I}\{i \in r\}, \mathbb{I}\{j \in r\}) = (1 - 1/n)^{2n} - (1 - 2/n)^n = O(1/n), \tag{4}$$

as can be directly checked. Third, the moment generating function of B_i is

$$\mathbb{E}\left[e^{sB_i}\right] = (p + (1-p)e^s)^B. \tag{5}$$

In the following, we first perform the stability analysis for the derandomized RF (consisting of an infinite number of trees) and then extend the results to the practical finite-B case.

3.2 Derandomized random forests

The following theorem formalizes the conditional stability property for the derandomized RF, the proof of which is a direct result of Theorem 8 in [28], and is omitted here.

Theorem 5 (Conditional stability of derandomized random forests)**.** *Conditional on training set D and test point (X, Y), for the derandomized random forest predictor* rf *we have that*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{\left|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)\right| > \varepsilon \Big| D, X\right\} \leqslant \delta(D, X) \equiv \frac{Z_{(n)}^2}{\varepsilon^2 n}\left(\frac{p}{1-p} + \frac{q}{(1-p)^2}\right). \tag{6}$$

If $\delta(D, X) \geqslant 1$, the statement is trivial, and we will focus on the case that $\delta(D, X) \in (0, 1)$. We can now establish the stability property for the derandomized RF.

Theorem 6 (Stability of derandomized random forests)**.** *For iid training and test data and $\varepsilon > 0$, the derandomized random forest predictor* rf *is stable with*

$$\mathbb{P}_{D,X}\left(\left|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)\right| > \varepsilon\right) \leqslant \frac{\mathbb{E}[Z_{(n)}^2]}{\varepsilon^2 n}\left(\frac{p}{1-p} + \frac{q}{(1-p)^2}\right) \equiv \nu. \tag{7}$$

This result follows directly from the conditional stability (6) by averaging over D and X. There is some freedom in choosing the dependence of ε on n. On one hand, in order to make sense of the word "stability," we do expect ε and ν to be small for large n. From (3) and (4), it is clear that the asymptotic behavior of ν is dominated by that of $\mathbb{E}[Z_{(n)}^2]/(\varepsilon^2 n)$, which can be tuned by manipulating ε. For example, a matching convergence rate to 0 between ε and ν might be desirable, and one can then set $\varepsilon = O((\mathbb{E}[Z_{(n)}^2]/n)^{1/3})$ if the scaling of $\mathbb{E}[Z_{(n)}^2] = o(n)$ is known or can be inferred. On the other hand, we can fix ε to further investigate the relation between stability and the convergence-in-probability property of the RF. By (7), under the condition that $\mathbb{E}[Z_{(n)}^2]/n \to 0$ as $n \to \infty$, one immediately comes to the conclusion that $\mathsf{rf}^{\backslash i}(X) - \mathsf{rf}(X)$ converges to 0 in probability. Actually, a stronger conclusion can be drawn under the same condition.

Corollary 1. *For iid training and test data, we have*

$$\mathbb{E}_{D,X}\left[|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)|\right] < 2\sqrt{\frac{\mathbb{E}[Z_{(n)}^2]}{n}\left(\frac{p}{1-p} + \frac{q}{(1-p)^2}\right)}. \tag{8}$$

Further assume that $\mathbb{E}[Y^2] < \infty$. Then we have

$$\mathbb{E}_{D,X}\left[|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)|\right] \to 0 \text{ and } \mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X) \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty. \tag{9}$$

5

Remark 3. The additional assumption that $\mathbb{E}[Y^2] < \infty$ is mild. Many commonly encountered random variables have a light tail and thus a finite second moment, irrespective of the detailed information of the distribution in question. Note that the bound (2) itself can be crude, and our result is expected to be valid even beyond this mild condition.

Remark 4. This result indicates that the *difference* is diminishing between rf and $\mathrm{rf}^{\backslash i}$, built on n and $n-1$ training data points, respectively. However, there is no indication that the derandomized $\mathrm{rf}(X)$ itself will converge to anything. This idea inspires the proposal of Theorem 11.

The proof of this result, as well as others below, will be deferred to the Appendix. So far, we have investigated the derandomized version of the RF, which is a limiting case and can be seen as consisting of an infinite number of trees, averaging out all kinds of possible randomness in the predictor construction process. In order to make the results more relevant to applied machine learning, the finite-B analysis for the RF is conducted below.

3.3 Finite-B random forests

We now consider the difference between RF_B and $\mathsf{RF}_B^{\backslash i}$. We denote $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_B)$ and $\boldsymbol{r} = (r_1, \ldots, r_B)$ as the corresponding sources of randomness in B trees. We also consider conditional stability first and then move to the stability of RF_B.

Theorem 7 (Conditional stability of finite-B random forests). *Conditional on training set D and test point (X, Y), for a random forest predictor RF_B that consists of B trees, we have for $\varepsilon > 0$ that*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{P}_{\boldsymbol{\xi},\boldsymbol{r}|D,X}\left(\left|\mathsf{RF}_B(X) - \mathsf{RF}_B^{\backslash i}(X)\right| > \varepsilon + 2\sqrt{\frac{2Z_{(n)}^2}{B}\ln\left(\frac{1}{\delta}\right)}\middle|D,X\right) \leqslant 3\delta + g(p,\delta,B),$$

where δ is short for $\delta(D, X)$ as defined in (6) and $g(p, \delta, B) = 2(p + (1-p)\delta^{\frac{1}{B}})^B$.

Next, we consider the case of iid data and investigate the RF stability by averaging out the randomness in data. Note that $Z_{(n)}$ and δ are random and depend on the data distribution, while we are interested in a probability bound for $|\mathsf{RF}_B(X) - \mathsf{RF}_B^{\backslash i}(X)|$ greater than a deterministic quantity, which is only a function of B and n. In this finite-B case, the stability of RF_B cannot be directly obtained from its conditional stability as in the derandomized situation.

Theorem 8 (Stability of finite-B random forests). *Assume training points in set D and the test point (X, Y) are iid, drawn from a fixed distribution. For the random forest predictor RF_B consisting of B trees and trained on D, we have*

$$\mathbb{P}_{D,X,\boldsymbol{\xi},\boldsymbol{r}}\left(\left|\mathsf{RF}_B(X) - \mathsf{RF}_B^{\backslash i}(X)\right| > \varepsilon_{n,B}\right) \leqslant \nu_{n,B}, \tag{10}$$

where $\varepsilon_{n,B} = \sum_{i=1}^3 \varepsilon_i$, and $\nu_{n,B} = \sum_{i=1}^3 \nu_i$. The pair of $(\varepsilon_2, \nu_2/\lambda)$ satisfies the derandomized RF stability condition (7) with $\lambda > 1$. Moreover, $\varepsilon_1 = \varepsilon_3 = \sqrt{2\lambda\mathbb{E}[Z_{(n)}^2]\ln(\frac{1}{\nu_2})/B}$, $\nu_1 = 2\nu_2 + 2\mathbb{P}(Z_{(n)}^2 > \lambda\mathbb{E}[Z_{(n)}^2])$, and $\nu_3 = g(p, \nu_2, B) + 2\mathbb{P}(Z_{(n)}^2 > \lambda\mathbb{E}[Z_{(n)}^2])$.

On a high level, the establishment of this theorem relies on two observations: (i) the stability of the derandomized RF, so that the difference $|\mathrm{rf}(X) - \mathrm{rf}^{\backslash i}(X)|$ is controlled, and (ii) the concentration of measure, so that the differences $|\mathsf{RF}_B(X) - \mathrm{rf}(X)|$ and $|\mathsf{RF}_B^{\backslash i}(X) - \mathrm{rf}^{\backslash i}(X)|$ are controlled. In order to make full sense of the word "stability," it is desirable that $\varepsilon_{n,B}$ and $\nu_{n,B}$ can converge to 0. It is known that $\mathbb{E}[Y^2] < \infty$ suffices to ensure $\mathbb{E}[Z_{(n)}^2] = o(n)$ [15, 13], and hence the stability of the derandomized RF. Now in the finite-B case, we need an additional distributional assumption to control the tail probability $\mathbb{P}(Z_{(n)}^2 > \lambda\mathbb{E}[Z_{(n)}^2])$. It turns out that for typical light-tailed Y^2, such a tail probability will converge to 0 as $n \to \infty$. Technically, we can assume Y^2 to be sub-gamma [8]. Note that bounded and sub-Gaussian random variables are sub-gamma. Hence the sub-gamma assumption is not strong and can be satisfied by distributions underlying many real datasets.

Definition 4 (Sub-gamma random variables [8]). A random variable W is said to be sub-gamma (on the right tail) with parameters (σ^2, c) where $c \geqslant 0$, if $\ln\mathbb{E}[e^{s(W-\mathbb{E}[W])}] \leqslant \frac{s^2\sigma^2}{2(1-cs)}$ for all $s \in (0, 1/c)$.

6
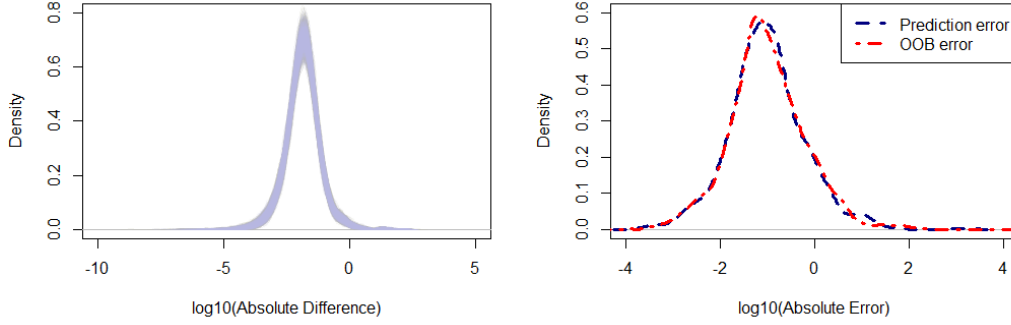
Figure 1: Left: Density plots of the $\log_{10}$ absolute difference $|\mathsf{RF}_B(X) - \mathsf{RF}_B^{\setminus i}(X)|$ for 3000 OOB predictors $\mathsf{RF}_B^{\setminus i}$ on 1000 test points. We let $B = 1000$. The RF stability (10) seems to persist, even though Y follows the (heavy-tailed) standard Cauchy distribution. Numerically, we set $\hat{\nu}_{n,B} = 0.05$ and calculated the maximum of the 0.95 quantile of the 3000 empirical distributions to have $\hat{\varepsilon}_{n,B} = 0.237$. Right: Density plots of 1000 $\log_{10}$ absolute prediction errors $|Y - \mathsf{RF}_B(X)|$ and of 3000 $\log_{10}$ absolute OOB errors $|Y_i - \mathsf{RF}_B^{\setminus i}(X_i)|$. The similarity between the plots supports the idea that the OOB errors can be used to construct PIs.

Lemma 1. *Suppose Y^2 is sub-gamma with parameters (σ^2, c) with $c > 0$, and $\mathbb{E}[Z_{(n)}^2] \sim a \ln n$ with $a \leqslant c$. For $\lambda > c/a$, we have $\lim_{n \to \infty} \mathbb{P}(Z_{(n)}^2 > \lambda \mathbb{E}[Z_{(n)}^2]) = 0$.*

Remark 5. We have set $c > 0$ above. If $c = 0$, then Y^2 is in fact sub-Gaussian, and the tail probability can be controlled similarly. If Y^2 is upper bounded by some constant M^2, the stability analysis is even simpler, and there is no need to consider the tail probability at all, as we can use M^2 in place of $Z_{(n)}^2$ in the conditional stability of the RF and then take expectation with respect to data.

Example 1. *Consider $Y^2 \sim \mathsf{Exp}(1)$, the exponential distribution with scale parameter 1. It is known that Y^2 is sub-gamma with $(\sigma^2, c) = (1,1)$ [8], and $\mathbb{E}[Z_{(n)}^2] = \sum_{i=1}^{n} \frac{1}{i} \equiv H_n$ with $H_n \in (\gamma + \ln n, \gamma + \ln(n+1))$, where $\gamma \approx 0.577$ is Euler's constant. Hence $H_n = \ln n + o(\ln n)$, and a straightforward calculation reveals that $\lim_{n \to \infty} \mathbb{P}(Z_{(n)}^2 > \lambda \mathbb{E}[Z_{(n)}^2]) = 0$ as long as $\lambda > 1$.*

From such results, one can see that the vanishing tail probability is not a stringent condition. By taking this additional assumption, it is indeed possible that both $\varepsilon_{n,B}$ and $\nu_{n,B}$ converge to 0.

Corollary 2. *For the same setting as in Theorem 8, suppose Y^2 is sub-gamma with parameters (σ^2, c) with $c > 0$ and $\mathbb{E}[Z_{(n)}^2] \sim a \ln n$ with $a \leqslant c$. Let $\lambda > c/a$ be a fixed number, and let B depend on n. Then for ε_2 that satisfies both $\varepsilon_2 = \omega(\sqrt{\ln n/n})$ and $\varepsilon_2 = o(1)$, and $B = \Omega(\ln^2 n)$, we have $\lim_{n \to \infty} \varepsilon_{n,B} = \lim_{n \to \infty} \nu_{n,B} = 0$.*

It is worth noting that there are multiple ways to let $\varepsilon_{n,B}$ and $\nu_{n,B}$ approach 0, as the dependence of ε_2, B, and even λ on n can all be manipulated. The point is that, theoretically, even the greedy RF can be stable with vanishing parameters. In practice, however, the stability of RF_B seems to hold in broader situations where both the moment and tail assumptions on Y^2 can be relaxed.

3.4 Stability in practice and limitations of theory

We created a virtual dataset consisting of $n = 4000$ points. We let Y be a standard Cauchy random variable, which is even without a well-defined mean. The feature vector $X \in \mathbb{R}^3$ is determined as $X = [0.5Y + \sin(Y), Y^2 - 0.2Y^3, \mathbb{I}\{Y > 0\} + \zeta]^T$ where ζ is a standard normal random variable. We used 3000 of the points for training and 1000 of them as test points. Using the `randomForest` package with default setting (except letting $B = 1000$), we had an output RF predictor RF_B. We also aggregated corresponding tree predictors to have 3000 OOB predictors $\mathsf{RF}_B^{\setminus i}$. For each $i \in [3000]$, we calculated the absolute difference $|\mathsf{RF}_B(X) - \mathsf{RF}_B^{\setminus i}(X)|$ on 1000 test points to come up with a

7

density plot for such a difference, shown in Fig. 1. We also calculated 1000 absolute prediction errors $|Y - \mathsf{RF}_B(X)|$ that are incurred by RF_B on test points, and 3000 OOB errors $|Y_i - \mathsf{RF}_B^{\backslash i}(X_i)|$, each incurred by an OOB predictor $\mathsf{RF}_B^{\backslash i}$ on its OOB point (X_i, Y_i). The computation can be done within a few minutes on a laptop. The density plots of these two kinds of errors are also shown in Fig. 1. This example shows that the RF stability can be present beyond the realm guaranteed by the light-tail assumption. As mentioned above, this is because the bound (2) can be conservative when n is large. We hope our results can inspire future study towards a more informative bound. Also, the similarity between the prediction error and the OOB error in this heavy-tail case indicates that the RF-based PIs analyzed below can find more applications in practice than justified by the current theory.

4 Random-forest prediction intervals

4.1 Comparison with related methods

With the stability property of the RF, it is possible to construct PIs with finite-sample guaranteed coverage. Recent years have witnessed the development of distribution-free predictive inference [1] with the full [33, 27], split [24, 31, 20], and jackknife+ [3, 32] conformal prediction methods being three milestones. The full conformal method is computationally prohibitive when used in practice. The split method greatly reduces the computational cost but fails to thoroughly extract the available information of training data. The jackknife+ (**J+**) method maximizes the usage of data at a computational cost in between those of full and split methods. In [18], jackknife+-after-bootstrap (**J+aB**) was proposed for bagged algorithms to achieve the same goal as in J+, while the training cost can be further reduced. However, the number of bags B is required to be a Binomial random variable, which might seem unnatural. It turns out that by further imposing the assumption of *ensemble* stability (which is essentially the concentration of resampling measure), J+aB can still have guaranteed coverage with a fixed B. Ensemble stability is defined for bagged algorithms. It measures how typical a bootstrap sample is, and is different from the algorithmic stability that quantifies the influence of removing one training point. If algorithmic stability is also imposed, then not only J+aB, but also jackknife can provide guaranteed coverage, which is otherwise impossible [29, 3].

Conceptually, the J+ approach and its variants under stability conditions are particularly relevant to this work. As the stability we establish for the RF contains both ensemble and algorithmic components, we will generally refer to the J+aB method with both ensemble and algorithmic stability as **J+aBS** and the jackknife method with algorithmic stability as **JS**. Our method might be best described as "jackknife-after-bootstrap-with-stability (**JaBS**)" tailored for the RF, which is different from both JS and J+aBS. Our method requires the least effort of computing as only one output predictor is needed, while all others require at least n output predictors.

There also exist RF-based PIs [17, 35] that are essentially of the jackknife-after-bootstrap (**JaB**) type and almost identical to ours practically when ε is small and n equals the size of a typical dataset. However, without stability, there is, in general, no guarantee for the coverage of such PIs, although the asymptotic coverage $1 - \alpha$ can be established based on strong assumptions [35]. We take advantage of the stability of the RF algorithm to establish the lower bound of coverage in Theorem 9 below. An upper bound is established in Theorem 10 with an additional mild assumption. We also propose a weaker assumption for asymptotic coverage in Theorem 11.

We compare these relevant methods to ours in Table 1 and Table 2, where the RF is set as the working algorithm for all methods and (ε, ν) is a general pair of stability parameters. We define $q_{n,\alpha}\{R_i\}$, $q_{n,\alpha}^+\{R_i\}$, $q_{n,\alpha}^-\{R_i\}$, and $q_{n,\alpha}'\{R_i\}$ as follows. Given $\{a_1, \ldots, a_n\}$,

$$q_{n,\alpha}\{a_i\} = q_{n,\alpha}^+\{a_i\} \equiv \text{the } \lceil (1-\alpha)(n+1) \rceil \text{-th smallest value of } \{a_1, \ldots, a_n\},$$

$$q_{n,\alpha}'\{a_i\} \equiv \text{the } \lceil (1-\alpha)n \rceil \text{-th smallest value of } \{a_1, \ldots, a_n\},$$

$$q_{n,\alpha}^-\{a_i\} \equiv \text{the } \lfloor \alpha(n+1) \rfloor \text{-th smallest value of } \{a_1, \ldots, a_n\},$$

where $\lfloor \cdot \rfloor$ is the floor function. Let $R_i^{\mathrm{LOO}} = |Y_i - \mathsf{RF}_B^{-i}(X_i)|$ be the LOO error, where RF_B^{-i} is trained without the ith training point, and by definition $\mathsf{RF}_B^{-i} \neq \mathsf{RF}_B^{\backslash i}$.

In Table 1, we list the corresponding PI constructed from each method and the output predictors of each method. The number of output predictors directly reflects the computational cost. It is

8

Table 1: Methods to construct prediction intervals using random forests: computational cost

Method	Output predictors	Prediction interval for future Y
J+ [3]	$\mathsf{RF}_B^{-i}, i \in [n]$	$[q_{n,\alpha}^-\{\mathsf{RF}_B^{-i}(X) - R_i^{\mathrm{LOO}}\}, q_{n,\alpha}^+\{\mathsf{RF}_B^{-i}(X) + R_i^{\mathrm{LOO}}\}]$
J+aB [18]	$\mathsf{RF}_B^{\backslash i}, i \in [n]$	$[q_{n,\alpha}^-\{\mathsf{RF}_B^{\backslash i}(X) - R_i\}, q_{n,\alpha}^+\{\mathsf{RF}_B^{\backslash i}(X) + R_i\}]$
JS [3]	RF_B and $\mathsf{RF}_B^{-i}, i \in [n]$	$\mathsf{RF}_B(X) \pm q_{n,\alpha}\{R_i^{\mathrm{LOO}} + \varepsilon\}$
J+aBS [18]	$\mathsf{RF}_B^{\backslash i}, i \in [n]$	$[q_{n,\alpha}^-\{\mathsf{RF}_B^{\backslash i}(X) - R_i\} - \varepsilon, q_{n,\alpha}^+\{\mathsf{RF}_B^{\backslash i}(X) + R_i\} + \varepsilon]$
JaB	RF_B	$\mathsf{RF}_B(X) \pm q_{n,\alpha}\{R_i\}$ [17]
		$\mathsf{RF}_B(X) \pm q_{n,\alpha}'\{R_i\}$ [35]
Ours (JaBS)	RF_B	$\mathsf{RF}_B(X) \pm q_{n,\alpha}\{R_i + \varepsilon\}$ (Theorem 9)
		$\mathsf{RF}_B(X) \pm q_{n,\alpha}\{R_i - \varepsilon\}$ (Theorem 10)
		$\mathsf{RF}_B(X) \pm q_{n,\alpha}\{R_i\}$ (Theorem 11)

Table 2: Methods to construct prediction intervals using random forests: theoretical coverage

Method	Theoretical coverage	Additional conditions
J+ [3]	$\geqslant 1 - 2\alpha$	None
J+aB [18]	$\geqslant 1 - 2\alpha$	Binomial B
JS [3]	$\geqslant 1 - \alpha - O(\sqrt{\nu})$	Stability (algorithmic)
J+aBS [18]	$\geqslant 1 - \alpha - O(\sqrt{\nu})$	Stability (ensemble + algorithmic)
JaB	No guarantee [17]	-
	$\to 1 - \alpha$ [35]	Strong (additive model, consistency of RF predictor)
Ours (JaBS)	$\geqslant 1 - \alpha - O(\sqrt{\nu})$	Stability (Theorem 9)
	$\leqslant 1 - \alpha + \frac{1}{n+1} + O(\sqrt{\nu})$	+ Distinct residuals (Theorem 10)
	$\to 1 - \alpha$	+ Uniformly equicontinuous CDF of $\lvert Y - \mathsf{RF}_B(X)\rvert$ and vanishing ε, ν (Theorem 11)

worth noting that acquiring the LOO predictor RF_B^{-i} needs substantial computation. In packages like `randomForest`, aggregating tree predictors to obtain the OOB predictor $\mathsf{RF}_B^{\backslash i}$ also needs extra computation. However, the predicted value $\mathsf{RF}_B^{\backslash i}(X_i)$ can be obtained immediately by calling the predict() function. The fact that the value of $\mathsf{RF}_B^{\backslash i}(X)$ on a test point is *not* needed further reduces the computational cost of JaB and our method, which only need one output RF predictor, and are more favorable computationally.

In Table 2, we list the coverage of the PI constructed from each method, as well as the additional conditions (beyond iid data) needed to achieve the coverage. Note that J+ does not require any additional conditions to achieve the coverage lower bound $1 - 2\alpha$, but J+aB requires that the number of trees B be a Binomial random variable. For JS, J+aBS, and our method, stability is needed to achieve the coverage lower bound $1 - \alpha - O(\sqrt{\nu})$. With additional mild assumptions, the coverage upper bound and asymptotic coverage of our method can be established. However, there is no guarantee of coverage for JaB without strong assumptions.

In summary, our theoretical work provides a series of coverage guarantees to a computationally feasible method for constructing PIs based on the RF algorithm. In the following, we will establish the lower and upper bound of coverage, as well as the asymptotic coverage.

4.2 Non-asymptotic coverage guarantees

Theorem 9 (Coverage lower bound). *Suppose the RF predictor* RF_B *satisfies the stability condition as in Theorem 8. Then we have for a test point* (X, Y) *that*

$$\mathbb{P}(Y \in \mathsf{RF}_B(X) \pm q_{n,\alpha}\{R_i + \varepsilon_{n,B}\}) \geqslant 1 - \alpha - \nu_1 - 2\sqrt{\nu_2} - 2\sqrt{\nu_3}. \tag{11}$$

This result is established by starting from the analysis of an imaginary extended dataset $\overline{D} = D \cup \{(X, Y)\}$, where the test point is *assumed* to be known. We denote (X, Y) as (X_{n+1}, Y_{n+1}) for convenience. For all points in $\overline{D}$, consider the derandomized RF predictor $\widetilde{\mathsf{rf}}^{\backslash i}$ that is built on n data

9

points without the ith point in $\overline{D}$, $i \in [n+1]$. One can then define the OOB error $\widetilde{r}_i \equiv |Y_i - \widetilde{\mathsf{rf}}^{\setminus i}|$. Since all data are iid, we have that $\mathbb{P}(\widetilde{r}_{n+1} \leqslant q_{n,\alpha}\{\widetilde{r}_i\}) \geqslant 1 - \alpha$, where $q_{n,\alpha}\{\widetilde{r}_i\}$ is the $\lceil(1-\alpha)(n+1)\rceil$-th smallest value of $\{\widetilde{r}_1, \ldots, \widetilde{r}_n\}$. Next, notice $\widetilde{r}_{n+1} = |Y_{n+1} - \widetilde{\mathsf{rf}}^{\setminus(n+1)}(X_{n+1})| = |Y_{n+1} - \mathsf{rf}(X_{n+1})|$ by the definitions of $\widetilde{\mathsf{rf}}^{\setminus(n+1)}$ and rf. By concentration of measure, $\mathsf{rf}(X_{n+1})$ can be approximated by $\mathsf{RF}_B(X_{n+1})$, and thus $\widetilde{r}_{n+1}$ can be roughly replaced with $|Y_{n+1} - \mathsf{RF}_B(X_{n+1})|$, which is desired. Then by stability of rf, $\{\widetilde{r}_i\}$ can be approximated by $\{r_i \equiv |Y_i - \mathsf{rf}^{\setminus i}(X_i)|\}$. Although $\{r_i\}$ is still unavailable in practice, by applying the idea of concentration of measure again, $\{r_i\}$ can be further approximated by $\{R_i\}$, which is accessible given D. Eventually, we can bound $|Y_{n+1} - \mathsf{RF}_B(X_{n+1})|$ in terms of $\{R_i\}$. The approximations are accounted for by the stability parameters in Theorem 8.

If we further assume that there are no ties among $\{\widetilde{r}_i\}$, $i \in [n+1]$, a typical case when Y is continuous, then we can also establish the upper bound of coverage.

Theorem 10 (Coverage upper bound). *Suppose there are no ties in $\{\widetilde{r}_i\}$, $i \in [n+1]$, and the RF predictor RF_B satisfies the stability condition as in Theorem 8. Then*

$$\mathbb{P}(Y \in \mathsf{RF}_B(X) \pm q_{n,\alpha}\{R_i - \varepsilon_{n,B}\}) \leqslant 1 - \alpha + \frac{1}{n+1} + \nu_1 + 2\sqrt{\nu_2} + 2\sqrt{\nu_3}. \tag{12}$$

The upper bound can be established because if there are no ties among $\widetilde{r}_1, \ldots, \widetilde{r}_{n+1}$, then $\mathbb{P}(\widetilde{r}_{n+1} \leqslant q_{n,\alpha}\{\widetilde{r}_i\}) \leqslant 1 - \alpha + \frac{1}{n+1}$. The apparent symmetry between the lower and upper bound originates from the fact that they both are established by using the RF stability once and the concentration of measure twice. Note that this idea can be applied to JS intervals for an arbitrary stable algorithm in exactly the same way, providing a complement to the lower bound for JS intervals established in [3].

Corollary 3 (Coverage upper bound for jackknife-with-stability intervals). *Let $\hat{f}$ be a predictor trained on n iid data points and $\hat{f}^{-i}$ be the LOO predictor without the ith point. Suppose $\hat{f}$ is stable with $\mathbb{P}(|\hat{f}(X) - \hat{f}^{-i}(X)| > \varepsilon) \leqslant \nu$, and the LOO errors are distinct on the extended training set that includes an iid test point (X, Y). Then we have $\mathbb{P}\left(|Y - \hat{f}(X)| \leqslant q_{n,\alpha}\{r_i - \varepsilon\}\right) \leqslant 1 - \alpha + \frac{1}{n+1} + 2\sqrt{\nu}$, where r_i are the LOO errors on the original training set.*

4.3 Asymptotic coverage guarantee

As shown above, the stability parameters $(\varepsilon_{n,B}, \nu_{n,B})$ can vanish when $n \to \infty$. It is reasonable to expect that $\mathbb{P}(Y \in \mathsf{RF}_B(X) \pm q_{n,\alpha}\{R_i\}) \to 1 - \alpha$ in this limit, as is consistent with numerous empirical observations [17, 35]. However, to achieve this goal, it seems that more assumptions are unavoidable. In [35], the guaranteed coverage of the JaB method is established by assuming that $\mathsf{RF}_B(X)$ converges to some $f_0(X)$ in probability as $n \to \infty$, where f_0 is the true regression function of an additive model that generates the data. We show that this can be done under weaker conditions.

Theorem 11 (Asymptotic coverage). *Denote F_n as the CDF of $|Y - \mathsf{RF}_B(X)|$. Suppose $\{F_n\}_{n \geqslant n_0}$ is uniformly equicontinuous for some n_0. Then $\mathbb{P}(Y \in \mathsf{RF}_B(X) \pm q_{n,\alpha}\{R_i\}) \to 1 - \alpha$ as $n \to \infty$ when conditions in Theorem 9, Theorem 10, and Corollary 2 are satisfied.*

Remark 6. Intuitively, using errors from $\mathsf{RF}_B^{\setminus i}$ that are trained on $n-1$ points to approximate those from RF_B, trained on n points, we only need this approximation to be exact asymptotically. There is no need for RF_B itself to converge to anything. This is one major conceptual difference between our work and [35], and it is in this sense that our assumption is weaker. Practically, this kind of PI is recommended as it does not involve $(\varepsilon_{n,B}, \nu_{n,B})$, and has great performance on numerous datasets.

5 Conclusion

In this work, for the first time, we theoretically establish the stability property of the greedy version of random forests, which is implemented in popular packages. The theoretical guarantee is based on a light-tail assumption of the marginal distribution of the squared response Y^2. However, numerical evidence suggests that this stability could persist in much broader situations. Based on the stability property and some mild conditions, we also establish finite-sample lower and upper bounds of coverage, as well as the exact coverage asymptotically, for prediction intervals constructed from the out-of-bag error of random forests, justifying random forests as an appealing method to provide both point and interval prediction simultaneously.

10

Acknowledgments and Disclosure of Funding

Much of the work was completed while Yan Wang was a PhD student in the Department of Statistics at Iowa State University. This research was supported in part by the US National Science Foundation under grant HDR:TRIPODS 19-34884.

References

[1] Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

[2] Arthur Asuncion and David Newman. UCI Machine Learning Repository, 2007.

[3] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

[4] Dimitris Bertsimas, Jack Dunn, and Ivan Paskov. Stable classification. *The Journal of Machine Learning Research*, 23(1):13401–13453, 2022.

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

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

[7] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

[8] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[9] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

[11] Thomas F Brooks, D Stuart Pope, and Michael A Marcolini. *Airfoil Self-Noise and Prediction*. National Aeronautics and Space Administration, Office of Management, 1989.

[12] Andrea Coraddu, Luca Oneto, Aessandro Ghio, Stefano Savio, Davide Anguita, and Massimo Figari. Machine learning approaches for improving condition-based maintenance of naval propulsion plants. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 230(1):136–153, 2016.

[13] José R Correa and Matías Romero. On the asymptotic behavior of the expectation of the maximum of iid random variables. *Operations Research Letters*, 49(5):785–786, 2021.

[14] Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *International Conference on Machine Learning*, pages 665–673. PMLR, 2014.

[15] Peter J Downey. Distribution-free bounds on the expectation of the maximum with scheduling applications. *Operations Research Letters*, 9(3):189–201, 1990.

[16] Francesca Grisoni, Viviana Consonni, Marco Vighi, Sara Villa, and Roberto Todeschini. Investigating the mechanisms of bioconcentration through QSAR classification trees. *Environment International*, 88:198–205, 2016.

[17] Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine Learning*, 97:155–176, 2014.

[18] Byol Kim, Chen Xu, and Rina Barber. Predictive inference is free with the jackknife+-after-bootstrap. *Advances in Neural Information Processing Systems*, 33:4138–4149, 2020.

[19] Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36. PMLR, 2014.

11

[20] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[21] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

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

[23] Andreas Maurer and Massimiliano Pontil. Concentration inequalities under sub-Gaussian and sub-exponential conditions. *Advances in Neural Information Processing Systems*, 34:7588–7597, 2021.

[24] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

[25] Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016.

[26] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

[27] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[28] Jake A Soloff, Rina Foygel Barber, and Rebecca Willett. Bagging provides assumption-free stability. *arXiv preprint arXiv:2301.12600*, 2023.

[29] Lukas Steinberger and Hannes Leeb. Conditional predictive inference for stable algorithms. *The Annals of Statistics*, 51(1):290–311, 2023.

[30] Cheng Tang, Damien Garreau, and Ulrike von Luxburg. When do random forests fail? *Advances in Neural Information Processing Systems*, 31, 2018.

[31] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pages 475–490. PMLR, 2012.

[32] Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28, 2015.

[33] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

[34] I-C Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808, 1998.

[35] Haozhe Zhang, Joshua Zimmerman, Dan Nettleton, and Daniel J Nordman. Random forest prediction intervals. *The American Statistician*, 74(4):392–406, 2020.

12

A Proof of Corollary 1

Let

$$\eta \equiv \frac{p}{1-p} + \frac{q}{(1-p)^2}. \tag{13}$$

Then $\delta(D, X) = Z_{(n)}^2 \eta/(\varepsilon^2 n)$. Also, by (3) and (4), η is upper bounded as

$$
\begin{aligned}
\eta &= \frac{1}{1-p} - 1 + \frac{\left(1-\frac{1}{n}\right)^{2n} - \left(1-\frac{2}{n}\right)^n}{\left(1-\frac{1}{n}\right)^{2n}} \\
&= \frac{1}{\left(1-\frac{1}{n}\right)^n} - \frac{\left(1-\frac{2}{n}\right)^n}{\left(1-\frac{1}{n}\right)^{2n}} = \frac{\left(1-\frac{1}{n}\right)^n - \left(1-\frac{2}{n}\right)^n}{\left(1-\frac{1}{n}\right)^{2n}} \\
&\leqslant \frac{\frac{1}{e} - 0}{1/16} \text{ (since } n \geqslant 2 \text{ and } (1-1/n)^n \text{ is monotonically increasing in } n) \\
&= \frac{16}{e} < 6.
\end{aligned}
$$

As a result, for all $n \geqslant 2$, we have $\eta/n < 3$.

Then, note that for a non-negative random variable W, it holds that $\mathbb{E}[W] = \int_0^\infty \mathbb{P}(W > t)dt$. By (2) and (6), we have

$$
\begin{aligned}
\mathbb{E}\left[\left|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)\right|\right] &= \mathbb{E}_{D,X}\left[\int_0^\infty \mathbb{P}_{\xi,r|D,X}\left(\left|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)\right| > \varepsilon \middle| D, X\right) d\varepsilon\right] \\
&= \mathbb{E}_{D,X}\left[\int_0^{2Z_{(n)}} \mathbb{P}_{\xi,r|D,X}\left(\left|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)\right| > \varepsilon \middle| D, X\right) d\varepsilon\right] \\
&\leqslant \mathbb{E}_{D,X}\left[\int_0^{2Z_{(n)}} \min\{\delta(D, X), 1\} d\varepsilon\right] \\
&= \mathbb{E}_{D,X}\left[\int_0^{\sqrt{Z_{(n)}^2\eta/n}} 1 d\varepsilon + \int_{\sqrt{Z_{(n)}^2\eta/n}}^{2Z_{(n)}} \frac{Z_{(n)}^2\eta}{\varepsilon^2 n} d\varepsilon\right] \\
&\quad \text{(note } 2Z_{(n)} > \sqrt{Z_{(n)}^2\eta/n} \text{ for } n \geqslant 2) \\
&< \mathbb{E}\left[2\sqrt{Z_{(n)}^2\eta/n}\right] \\
&\leqslant 2\sqrt{\mathbb{E}[Z_{(n)}^2]\eta/n} \text{ (by Jensen's inequality)},
\end{aligned}
$$

which completes the first part of the corollary. For the second part, we need the following lemma, which is established in [15, 13].

Lemma 2. *For iid random variables $W_i, i \in [n]$, if $\mathbb{E}[|W_i|] < \infty$, then $\lim_{n\to\infty} \mathbb{E}[W_{(n)}]/n = 0$.*

By this lemma, $\mathbb{E}[Y^2] < \infty$ implies $\mathbb{E}[Z_{(n)}^2]/n \to 0$ as $n \to \infty$. Since η is bounded, it is clear that $|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)|$ converges to 0 in mean, which further implies convergence in probability, completing the proof.

Remark 7. We can actually have a slightly tighter bound in (6), and thus in Theorem 6 and Corollary 1. A closer look at the method in [28] indicates that the key point is to bound the conditional variance of TREE, $\mathbb{E}_{\xi,r|D,X}[(\mathsf{rf} - \mathsf{TREE})^2]$. Note that $\mathsf{rf} = p \times \mathbb{E}_{\xi,r|D,X}[\mathsf{TREE}|i \in r] + (1-p) \times \mathbb{E}_{\xi,r|D,X}[\mathsf{TREE}|i \notin r] \equiv p \times \mathsf{rf}^i + (1-p) \times \mathsf{rf}^{\backslash i}$, but $|\mathsf{rf}^i - \mathsf{rf}^{\backslash i}|$ is also bounded by $2Z_{(n)}$, and we have $|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)| \leqslant 2p \times Z_{(n)}$. By (3), p converges to a constant as $n \to \infty$. Hence this tighter bound does not provide any qualitative difference as n increases, and we ignore this minor improvement in this work.

13

B   Proof of Theorem 7 and discussion

By the triangle inequality and the union bound, we have for some $t > 0$ that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\xi},\boldsymbol{r}|D,X}\left(\left|\mathsf{RF}_B(X) - \mathsf{RF}_B^{\backslash i}(X)\right| > \varepsilon + 2t\Big|D,X\right)$$

$$\leqslant \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\xi},\boldsymbol{r}|D,X}\left(|\mathsf{RF}_B(X) - \mathsf{rf}(X)| > t|D,X\right)$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\xi},\boldsymbol{r}|D,X}\left(\left|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)\right| > \varepsilon\Big|D,X\right)$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\boldsymbol{\xi},\boldsymbol{r}|D,X}\left(\left|\mathsf{rf}^{\backslash i}(X) - \mathsf{RF}_B^{\backslash i}(X)\right| > t\Big|D,X\right)$$

$$\equiv T_1 + T_2 + T_3. \tag{14}$$

In T_2, rf and $\mathsf{rf}^{\backslash i}$ are derandomized predictors, and by (6), we have

$$T_2 \leqslant \delta.$$

Of course, this result is nontrivial when $\delta \in (0,1)$. Note that rf is the RF with an infinite number of trees, where each tree is characterized by a pair of independent random variables (ξ, r), while RF_B consists of a finite number of B trees. Since the value of $\mathsf{RF}_B(X)$ is bounded in $\left[-Z_{(n)}, Z_{(n)}\right]$, T_1 can be bounded by Hoeffding's inequality as

$$T_1 \leqslant 2\exp\left(-\frac{Bt^2}{2Z_{(n)}^2}\right).$$

The analysis of the first two terms T_1 and T_2 is similar to that in [28], but the bound for T_3 to be developed is specific to the RF predictor, and is nontrivially different from the case in [28] where the number of bags for each LOO predictor is fixed to B. In our setting, $\mathsf{RF}_B^{\backslash i}(X) \in [-Z_{(n)}, Z_{(n)}]$ is the aggregation of B_i trees where the ith training point is not included. Hence by Hoeffding's inequality and (5), we have

$$T_3 \leqslant \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{B_i|D,X}\left[2\exp\left(-\frac{B_i t^2}{2Z_{(n)}^2}\right)\Bigg|D,X\right] = 2\left(p + (1-p)e^{-t^2/(2Z_{(n)}^2)}\right)^B.$$

Now, we choose t such that $(0,1) \ni \delta = \exp\left(-Bt^2/(2Z_{(n)}^2)\right)$, which then yields $t = \sqrt{\frac{2Z_{(n)}^2}{B}\ln\left(\frac{1}{\delta}\right)}$, and thus

$$T_3 \leqslant 2\left(p + (1-p)\delta^{\frac{1}{B}}\right)^B \equiv g(p,\delta,B). \tag{15}$$

Combining these results together completes the proof.

Discussion.   Note that $g(p,\delta,B)$ is monotonically decreasing in B, because

$$\frac{\partial \ln[g(p,\delta,B)/2]}{\partial B} = \ln\left(p + (1-p)\delta^{1/B}\right) - \frac{B(1-p)\delta^{1/B}}{p+(1-p)\delta^{1/B}}\frac{\ln\delta}{B^2}$$

$$\leqslant p\ln 1 + (1-p)\ln\delta^{1/B} - \frac{(1-p)\delta^{1/B}}{p+(1-p)\delta^{1/B}}\frac{\ln\delta}{B} \quad \text{(by Jensen's inequality)}$$

$$= (1-p)\left[1 - \frac{\delta^{1/B}}{p+(1-p)\delta^{1/B}}\right]\frac{\ln\delta}{B}$$

$$= \frac{1-p}{B}\frac{p(1-\delta^{1/B})}{p+(1-p)\delta^{1/B}}\ln\delta$$

$$< 0 \quad \text{(as } 0 < \delta < 1\text{)}.$$

14

Moreover, since $g \geqslant 0$ and monotonically decreasing in B, it has a limit when p and δ are fixed and $B \to \infty$:

$$
\begin{aligned}
\lim_{B \to \infty} \ln[g(p, \delta, B)/2] &= \lim_{B \to \infty} B \ln \left[p + (1-p)\delta^{1/B} \right] \\
&= \lim_{B \to \infty} B \ln \left[p + (1-p)e^{\ln \delta / B} \right] \\
&= \lim_{B \to \infty} B \ln \left[p + (1-p) \left(1 + \frac{\ln \delta}{B} \right) + o \left(\frac{\ln \delta}{B} \right) \right] \\
&= \lim_{B \to \infty} B \ln \left[1 + \frac{\ln \delta^{1-p}}{B} + o \left(\frac{\ln \delta}{B} \right) \right] \\
&= \lim_{B \to \infty} \ln \left[\left(1 + \frac{\ln \delta^{1-p}}{B} + o \left(\frac{\ln \delta}{B} \right) \right)^B \right] \\
&= \ln \delta^{1-p}. \quad (16)
\end{aligned}
$$

Hence we have that $\lim_{B \to \infty} g(p, \delta, B) = 2\delta^{1-p}$. This result implies that if B is allowed to increase in a way that is independent of n, then under the conditions of Theorem 7, we have

$$
\lim_{B \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_{\boldsymbol{\xi}, \boldsymbol{r}|D, X} \left(\left| \mathsf{RF}_B(X) - \mathsf{RF}_B^{\backslash i}(X) \right| > \varepsilon + 2\sqrt{\frac{2Z_{(n)}^2}{B} \ln \left(\frac{1}{\delta} \right)} \middle| D, X \right) \leqslant 5\delta^{1-p}.
$$

We will come back to this analysis in the proof of Corollary 2 below where B grows with n, but it still holds that $\ln \delta(n) = o(B(n))$.

C Proofs of Theorem 8, Lemma 1, and Corollary 2; and discussion

C.1 Proof of Theorem 8

Similar to the proof for conditional stability, we have for some $\varepsilon_2 \equiv \varepsilon, \varepsilon_1 = \varepsilon_3 \equiv t > 0$ that

$$
\begin{aligned}
\mathbb{P}_{D, X, \boldsymbol{\xi}, \boldsymbol{r}} &\left(\left| \mathsf{RF}_B(X) - \mathsf{RF}_B^{\backslash i}(X) \right| > \varepsilon + 2t \right) \\
&\leqslant \mathbb{P}_{D, X, \boldsymbol{\xi}, \boldsymbol{r}} \left(|\mathsf{RF}_B(X) - \mathsf{rf}(X)| > t \right) \\
&\quad + \mathbb{P}_{D, X} \left(\left| \mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X) \right| > \varepsilon \right) \\
&\quad + \mathbb{P}_{D, X, \boldsymbol{\xi}, \boldsymbol{r}} \left(\left| \mathsf{rf}^{\backslash i}(X) - \mathsf{RF}_B^{\backslash i}(X) \right| > t \right) \\
&\equiv T_1' + T_2' + T_3'.
\end{aligned}
$$

We first consider the T_2' term, which can be bounded when the stability condition (7) for rf is applied. Thus

$$
T_2' \leqslant \nu \equiv \nu_2/\lambda. \quad (17)
$$

Then, taking expectation with respect to data distribution for T_1 and T_3 in (14) yields

$$
T_1' \leqslant \mathbb{E} \left[2 \exp \left(-\frac{Bt^2}{2Z_{(n)}^2} \right) \right],
$$

and

$$
T_3' \leqslant \mathbb{E} \left[2 \left(p + (1-p)e^{-t^2/(2Z_{(n)}^2)} \right)^B \right].
$$

In order to obtain more informative bounds than the above ones, we introduce the following lemma.

Lemma 3. *Let W be a random variable with finite mean, and suppose that h is a generic monotonically increasing function of W and is bounded in $[0, 1]$. Then we have for some $\lambda > 1$ that*

$$
\mathbb{E}[h(W)] \leqslant h(\lambda \mathbb{E}[W]) + \mathbb{P}(W > \lambda \mathbb{E}[W]).
$$

15

Proof. This can be established by noting that

$$\mathbb{E}[h(W)] = \mathbb{E}[h(W)\mathbb{I}\{W \leqslant \lambda\mathbb{E}[W]\}] + \mathbb{E}[h(W)\mathbb{I}\{W > \lambda\mathbb{E}[W]\}]$$
$$\leqslant h(\lambda\mathbb{E}[W]) + \mathbb{P}(W > \lambda\mathbb{E}[W]).$$

The first term on the right-hand side results from h being monotonically increasing, and the second term from h being upper bounded by 1. $\qquad\square$

Applying this lemma to the upper bounds of T_1' and T_3' above yields

$$T_1' \leqslant 2\exp\left(-\frac{Bt^2}{2\lambda\mathbb{E}[Z_{(n)}^2]}\right) + 2\mathbb{P}\left(Z_{(n)}^2 > \lambda\mathbb{E}[Z_{(n)}^2]\right) \equiv \nu_1,$$

and

$$T_3' \leqslant 2\left(p + (1-p)e^{-t^2/(2\lambda\mathbb{E}[Z_{(n)}^2])}\right)^B + 2\mathbb{P}\left(Z_{(n)}^2 > \lambda\mathbb{E}[Z_{(n)}^2]\right) \equiv \nu_3.$$

Now we choose t such that $\nu_2 = \exp\left(-Bt^2/(2\lambda\mathbb{E}[Z_{(n)}^2])\right)$ to complete the proof of Theorem 8.

Remark 8. Unlike the conditional stability case, an additional tail probability term is introduced in the upper bounds for both T_1' and T_3'. This is because when $Z_{(n)}^2$ is unbounded and no detailed information of the data-generating distribution is provided, there seems no universal way to control terms like $\mathbb{E}[\exp(-1/Z_{(n)}^2)]$. We tackle this problem by introducing the tail probability. As shown below, it is not difficult to control this term for typical data distributions.

C.2   Proof of Lemma 1

Suppose Z_1^2 is sub-gamma on the right tail with parameters (σ^2, c), then it is known that $\mathbb{E}[Z_{(n)}^2]$ is at most on the order of $O(\ln n)$ by the maximal inequality for sub-gamma random variables [8]. More precisely, we have

$$\mathbb{E}[Z_{(n)}^2] \leqslant \sqrt{2\sigma^2 \ln n} + c\ln n.$$

When $c = 0$, this reduces to the maximal inequality for sub-Gaussian random variables. When $c > 0$, the leading term of the bound is $O(\ln n)$, meaning that the tail probability decays in an exponential, rather than Gaussian, way. The parameter c also sets an upper bound quantitatively, and for large n, we must have $\mathbb{E}[Z_{(n)}^2] \sim a\ln n$ with $a \leqslant c$.

Since data are iid, the tail probability for the order statistic $Z_{(n)}^2$ can be rewritten as

$$\mathbb{P}\left(Z_{(n)}^2 > \lambda\mathbb{E}[Z_{(n)}^2]\right) = 1 - \left[1 - \mathbb{P}\left(Z_1^2 > \lambda\mathbb{E}[Z_{(n)}^2]\right)\right]^n \leqslant n\mathbb{P}\left(Z_1^2 > \lambda\mathbb{E}[Z_{(n)}^2]\right).$$

Moreover, we have the concentration inequality that for every $t > 0$ [8],

$$\mathbb{P}\left(Z_1^2 > \sqrt{2\sigma^2 t} + ct + \mathbb{E}[Z_1^2]\right) \leqslant e^{-t}.$$

Setting $\lambda\mathbb{E}[Z_{(n)}^2] = \sqrt{2\sigma^2 t} + ct + \mathbb{E}[Z_1^2]$, and noticing $t = \frac{\lambda a}{c}\ln n + o(\ln n)$ for large n, we obtain for $\lambda > \frac{c}{a}$ that

$$\lim_{n\to\infty}\mathbb{P}\left(Z_{(n)}^2 > \lambda\mathbb{E}\left[Z_n^2\right]\right) \leqslant \lim_{n\to\infty} ne^{-\frac{\lambda a}{c}\ln n} = \lim_{n\to\infty}\left(\frac{1}{n}\right)^{\frac{\lambda a}{c}-1} = 0,$$

completing the proof.

C.3   Proof of Corollary 2

Consider each pair of (ε_i, ν_i) defined above. First, the $(\varepsilon_2, \nu_2/\lambda)$ pair satisfies the stability condition (7) of the derandomized RF. Hence for (ε_2, ν_2) to converge to 0, we require

$$\varepsilon_2(n) \to 0 \text{ as } n \to \infty,$$

16

and

$$\nu_2(n) = \frac{\lambda\mathbb{E}[Z_{(n)}^2]\eta(n)}{n\varepsilon_2^2(n)} \to 0 \text{ as } n \to \infty,$$

where $\eta(n)$ is defined in (13). It is clear that $\varepsilon_2(n) = o(1)$ is required. Since $\eta(n)$ converges to a positive constant, and $\mathbb{E}[Z_{(n)}^2] \sim a \ln n$ with $a \leqslant c$ by the sub-gamma assumption [2], we thus have

$$\nu_2(n) = \Theta\left(\frac{\ln n}{n\varepsilon_2^2(n)}\right).$$

So if $\nu_2(n) \to 0$ as $n \to \infty$, we must also have

$$\varepsilon_2(n) = \omega\left(\sqrt{\frac{\ln n}{n}}\right).$$

Second, consider the (ε_1, ν_1) pair. In ν_1, the additional tail probability term converges to 0 by the sub-gamma assumption, as proved in Lemma 1, so ν_1 converges to 0. For ε_1, by definition,

$$\varepsilon_1 = \sqrt{\frac{2\lambda\mathbb{E}[Z_{(n)}^2]}{B}\ln\left(\frac{1}{\nu_2}\right)} = O\left(\sqrt{\frac{\ln n}{B}\ln\left(\frac{n\varepsilon_2^2}{\ln n}\right)}\right).$$

Since ε_2 satisfies both $\varepsilon_2 = o(1)$ and $\varepsilon_2 = \omega(\sqrt{\ln n/n})$, letting $B = \Omega(\ln^2 n)$ suffices for ε_1 to converge to 0.

Last, consider the (ε_3, ν_3) pair. Since $\varepsilon_3 = \varepsilon_1$, hence ε_3 converges to 0. In ν_3, the tail probability term is proved to converge to 0 above, and it remains to show $(p + (1-p)\nu_2^{\frac{1}{B}})^B$ converges to 0. In (16), we have seen in the conditional stability case that $g(p, \delta, B) \to 2\delta^{1-p}$ as $B \to \infty$, as long as $\ln \delta = o(B)$. Similarly, when $B(n) = \Omega(\ln^2 n)$, then $\ln \nu_2 = o(B)$, and we have

$$\lim_{n\to\infty} B(n)\ln\left(p(n) + (1-p(n))\nu_2(n)^{\frac{1}{B(n)}}\right)$$
$$= \lim_{n\to\infty} B(n)\ln\left(p(n) + (1-p(n))\exp\left(\frac{\ln\nu_2(n)}{B(n)}\right)\right)$$
$$= \lim_{n\to\infty} B(n)\ln\left(p(n) + (1-p(n))\left(1 + \frac{\ln\nu_2(n)}{B(n)}\right)\right)$$
$$= \lim_{n\to\infty} B(n)\ln\left(1 + \frac{(1-p(n))\ln\nu_2(n)}{B(n)}\right)$$
$$= \lim_{n\to\infty} B(n)\ln\left(1 + \frac{(1/e + O(1/n))\ln\nu_2(n)}{B(n)}\right)$$
$$= \lim_{n\to\infty} B(n)\ln\left(1 + \frac{\ln\nu_2^{1/e}(n)}{B(n)}\right)$$
$$= \lim_{n\to\infty} \ln\left[\left(1 + \frac{\ln\nu_2^{1/e}(n)}{B(n)}\right)^{B(n)}\right],$$

which indicates

$$\lim_{n\to\infty} (p(n) + (1-p(n))\nu_2(n)^{\frac{1}{B(n)}})^{B(n)} = \lim_{n\to\infty}\left(1 + \frac{\ln\nu_2^{1/e}(n)}{B(n)}\right)^{B(n)} = \lim_{n\to\infty}\nu_2^{1/e}(n) = 0,$$

and the proof is completed.

---

[2]Strictly speaking, it is possible for a sub-gamma random variable to have the scaling of $\mathbb{E}[Z_{(n)}^2]$ in between $\sqrt{\ln n}$ and $\ln n$, but the point is that the heavier the tail (i.e., the faster growth of $\mathbb{E}[Z_{(n)}^2]$), the more difficult to achieve stability in theory. So we focus on the most heavy-tail case that $\mathbb{E}[Z_{(n)}^2]$ scales as $\ln n$.

17

C.4 Discussion: beyond the sub-gamma assumption

There is some subtlety in choosing the dependence of ε_2, B, and λ on n. We consider the case that λ is a fixed number above. However, if we make λ also depend on n, then the assumption that Y^2 is sub-gamma can be removed. To see this, consider $\lim_{n\to\infty} \lambda(n) = \infty$. Then by Markov's inequality,

$$\lim_{n\to\infty} \mathbb{P}(Z_{(n)}^2 > \lambda(n)\mathbb{E}[Z_{(n)}^2]) \leqslant \lim_{n\to\infty} \frac{1}{\lambda(n)} = 0.$$

So in this case, $\mathbb{E}[Z_{(n)}^2]$ can have a faster growth rate in n and the sub-gamma assumption of Y^2 is not needed to control the tail probability. Consider $\mathbb{E}[Z_{(n)}^2] = \omega(\ln n)$, so that Y^2 has a heavier tail than sub-gamma. Requiring $\nu_2(n)$ to converge to 0 then results in

$$\varepsilon_2(n) = \omega\left(\sqrt{\lambda(n)\mathbb{E}[Z_{(n)}^2]/n}\right),$$

but $\varepsilon_2(n)$ also has to converge to 0, i.e., $\varepsilon_2(n) = o(1)$. This introduces another constraint on $\mathbb{E}[Z_{(n)}^2]$ that $\mathbb{E}[Z_{(n)}^2] = o(n/\lambda(n))$. A straightforward analysis shows that

$$B(n) = \Omega(\lambda(n)\mathbb{E}[Z_{(n)}^2]\ln(1/\nu_2(n)))$$

suffices to guarantee the convergence of ε_1, ε_3, and ν_3.

Summing up, we find that by allowing $\lim_{n\to\infty} \lambda(n) = \infty$, one can generalize the results in Corollary 2 to the case that Y^2 follows a distribution beyond sub-gamma. The scaling of $\mathbb{E}[Z_{(n)}^2]$ is an indicator of the tail behavior of the underlying distribution of Y^2, and $\mathbb{E}[Z_{(n)}^2]$ that satisfies both

$$\mathbb{E}[Z_{(n)}^2] = \omega(\ln n) \text{ and } \mathbb{E}[Z_{(n)}^2] = o(n/\lambda(n))$$

represents a wide class of distributions with tails heavier than sub-gamma for proper dependence of $\lambda(n)$ on n.

Example 2. *Consider some random variable Y^2 with $\mathbb{E}[Z_{(n)}^2] = \Theta(n^{1/4})$, and pick $\lambda(n) = \Theta(n^{1/4})$, $\varepsilon_2 = \Theta(n^{-1/6})$, and $B = \Omega(n\ln n)$. Then consequently*

$$\nu_2 = \frac{\lambda\mathbb{E}[Z_{(n)}^2]\eta}{n\varepsilon_2^2} = \Theta(n^{-1/6}),$$

$$\varepsilon_1 = \varepsilon_3 = \sqrt{\frac{2\lambda\mathbb{E}[Z_{(n)}^2]}{B}\ln\left(\frac{1}{\nu_2}\right)} = O(n^{-1/4}),$$

$$\nu_1 = 2\nu_2 + 2\mathbb{P}(Z_{(n)}^2 > \lambda\mathbb{E}[Z_{(n)}^2]) = \Theta(n^{-1/6}) + O(n^{-1/4}) = \Theta(n^{-1/6}),$$

$$\nu_3 = 2(p + (1-p)\nu_2^{\frac{1}{B}})^B + 2\mathbb{P}(Z_{(n)}^2 > \lambda\mathbb{E}[Z_{(n)}^2]) = \Theta(n^{-1/(6e)}) + O(n^{-1/4}) = \Theta(n^{-1/(6e)}).$$

Hence $\varepsilon_{n,B}$ and $\nu_{n,B}$ converge to 0 for a heavy-tailed Y^2 beyond sub-gamma.

In Table 3, we summarize the limiting stability parameters that can be achieved based on our theory, with properly chosen ε_2, B, and λ. This result shows the wide applicability of the RF stability. Sub-Gaussian is a subset of sub-gamma random variables with $c = 0$, and the analysis is similar. For bounded Y^2, there is even no need to introduce the tail probability term because $|Y| \leqslant M$ implies $Z_{(n)}^2 \leqslant M^2$, and T_1' and T_3' are naturally bounded as

$$T_1' \leqslant 2\exp\left(-\frac{Bt^2}{2M^2}\right),$$

and

$$T_3' \leqslant 2\left(p + (1-p)e^{-t^2/(2M^2)}\right)^B,$$

respectively. Both of the upper bounds are deterministic, and the analysis is greatly simplified. All we need to do is replace $\lambda\mathbb{E}[Z_{(n)}^2]$ with M^2 and drop the tail probability term in Theorem 8.

18

Table 3: Summary of random forest stability conditions

Scaling of $\mathbb{E}[Z_{(n)}^2]$	Value of λ	Stability parameters
$O(1)$ (bounded Y^2)	Unnecessary	$\varepsilon_{n,B} \to 0, \nu_{n,B} \to 0$
$O(\sqrt{\ln n})$ (sub-Gaussian Y^2)	Constant	$\varepsilon_{n,B} \to 0, \nu_{n,B} \to 0$
$O(\ln n)$ (sub-gamma Y^2)	Constant	$\varepsilon_{n,B} \to 0, \nu_{n,B} \to 0$
Between $\omega(\ln n)$ and $o(n/\lambda(n))$	$\lambda \to \infty$	$\varepsilon_{n,B} \to 0, \nu_{n,B} \to 0$

Corollary 4. *Assume training points in set D and the test point (X, Y) are iid, and $|Y| \leqslant M$. For the RF predictor* RF_B *consisting of B trees and trained on D, we have*

$$\mathbb{P}_{D,X,\boldsymbol{\xi},\boldsymbol{r}} \left(\left| \mathsf{RF}_B(X) - \mathsf{RF}_B^{\backslash i}(X) \right| > \varepsilon_{n,B} \right) \leqslant \nu_{n,B}, \tag{18}$$

where $\varepsilon_{n,B} = \sum_{i=1}^3 \varepsilon_i$, and $\nu_{n,B} = \sum_{i=1}^3 \nu_i$. The pair of (ε_2, ν_2) satisfies the derandomized RF stability condition

$$\mathbb{P}_{D,X} \left(\left| \mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X) \right| > \varepsilon_2 \right) \leqslant \frac{M^2}{\varepsilon^2 n} \left(\frac{p}{1-p} + \frac{q}{(1-p)^2} \right) = \nu_2,$$

where $\varepsilon_1 = \varepsilon_3 = \sqrt{2M^2 \ln(\frac{1}{\nu_2})/B}$, $\nu_1 = 2\nu_2$, and $\nu_3 = g(p, \nu_2, B)$.

Based on our theory, for a random variable Y^2 with $\mathbb{E}[Z_{(n)}^2] = O(n)$, there is no hope to get both $\nu_{n,B}$ and $\varepsilon_{n,B}$ to converge to 0. But as we show in Fig. 1, there are hints that the RF stability persists beyond our theory. This is because (2) is used to provide the worst-case deviation bound between rf and $\mathsf{rf}^{\backslash i}$. If a more informative bound can be found to replace (2), and the dependence of such a bound on n is $o(\mathbb{E}[Z_{(n)}^2])$, then it is possible to find vanishing stability parameters even for strongly heavy-tailed random variables theoretically. Such an improved bound may also help boost the convergence rates of $\varepsilon_{n,B}$ and $\nu_{n,B}$. It might be a future research direction to look for a better bound.

C.5 More examples of RF stability

We consider the RF stability on four real datasets, publicly available at UCI Machine Learning Repository [2]. We name them as `Concrete` [34], `Airfoil` [11], `Bioconcentration` [16], and `Naval` [12]. For each dataset, we investigate three aspects:

Marginal distribution of Y. As shown in Fig. 2 (left column), all the density plots of Y seem to have a light tail. In fact, the response in many real datasets is bounded or narrowly distributed within a given interval, which has to do with the physical constraints. For example, the strength of some material is determined by the underlying chemical bond strength, and cannot be arbitrarily large.

Difference between the RF predictor and OOB predictors. To this end, we randomly split every dataset into two parts with equal size n. One is for training and the other for testing. Based on n training points, we have an RF predictor RF_B, as well as n OOB predictors $\mathsf{RF}_B^{\backslash i}$. We fix $B = 1000$ in all cases. For each of these predictors, we compute its error on n test points, and in Fig. 2 (middle column) we plot the density of the absolute difference $|\mathsf{RF}_B(X) - \mathsf{RF}_B^{\backslash i}(X)|$ for all n OOB predictors. Also, we calculate the 0.95 quantile of each empirical distribution of the difference, say, $\hat{\varepsilon}_{0.95,i}, i \in [n]$. We let $\hat{\varepsilon}_{n,B} = \max_i \hat{\varepsilon}_{0.95,i}$ as an estimate of $\varepsilon_{n,B}$ given $\nu_{n,B} = 0.05$.

Comparison between prediction error and OOB error. For the RF predictor, we calculate its prediction error on n test points to come up with a density plot of such prediction errors. For each OOB predictor, we calculate its corresponding OOB error, and we have a density plot based on n such OOB errors. In Fig. 2 (right column), we plot both of prediction and OOB error. The similarity between them, especially on the right tail, provides much credence to the idea of constructing PIs using the OOB error.

We train the RFs using default parameters except that the number of trees is fixed to be $B = 1000$ in all cases. The training can be done within a few minutes on a laptop.

19
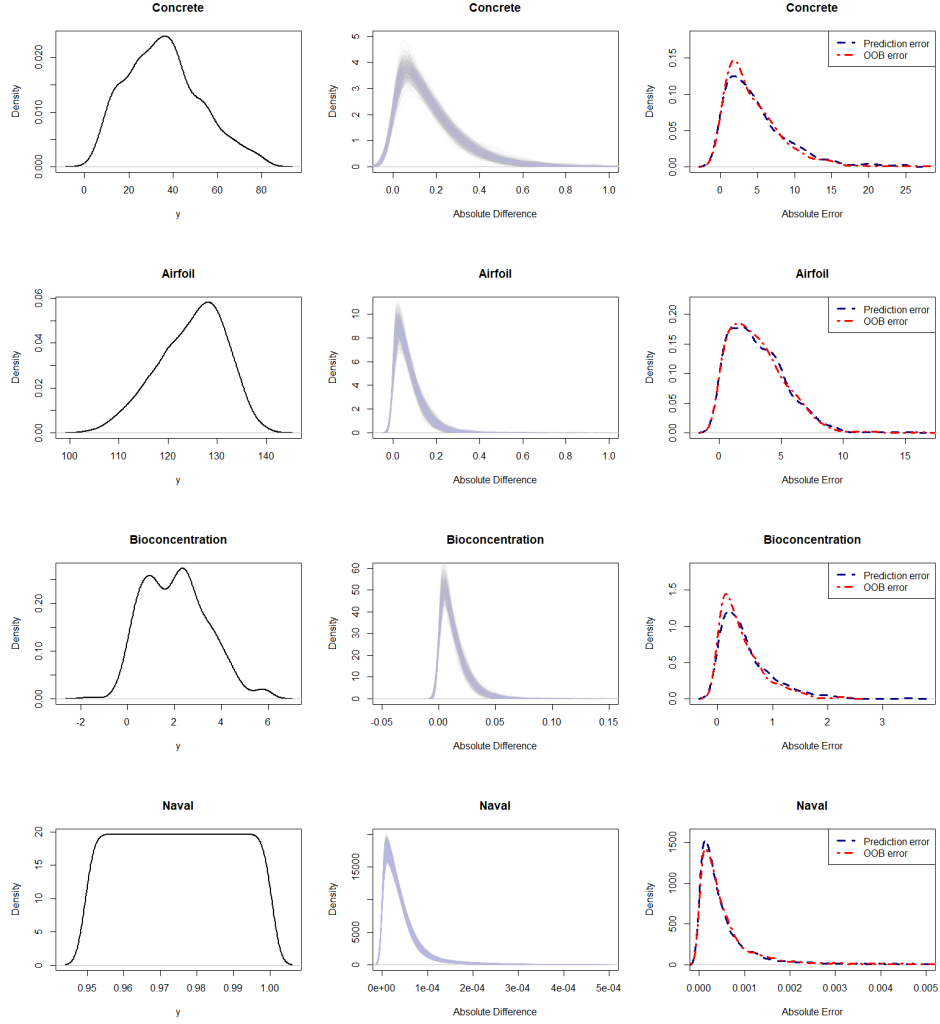
Figure 2: Left column: Density plots of Y. Middle column: Density plots of $|\mathsf{RF}_B(X) - \mathsf{RF}_B^{\backslash i}(X)|$. Right column: Density plots of $|Y - \mathsf{RF}_B(X)|$ and $|Y_i - \mathsf{RF}_B^{\backslash i}(X_i)|$. Row 1: `Concrete` dataset. Numerically, we find $\hat{\varepsilon}_{n,B} \approx 0.62, \hat{\nu}_{n,B} = 0.05$ for $n = 515$. Row 2: `Airfoil` dataset with $\hat{\varepsilon}_{n,B} \approx 0.25, \hat{\nu}_{n,B} = 0.05$ for $n = 751$. Row 3: `Bioconcentration` dataset with $\hat{\varepsilon}_{n,B} = 0.05, \hat{\nu}_{n,B} \approx 0.05$ for $n = 389$. Row 4: `Naval` dataset with $\hat{\varepsilon}_{n,B} \approx 0.00016, \hat{\nu}_{n,B} = 0.05$ for $n = 5967$.

20

D   Proof of Theorem 9

Recall that each of $\varepsilon_{n,B}$ and $\nu_{n,B}$ can be written as the sum of three terms:

$$\varepsilon_{n,B} = \varepsilon_1 + \varepsilon_2 + \varepsilon_3, \ \nu_{n,B} = \nu_1 + \nu_2 + \nu_3,$$

and we have established in the proof of Theorem 8 that

$$\mathbb{P}\left(|\mathsf{rf}(X) - \mathsf{RF}_B(X)| > \varepsilon_1\right) \leqslant \nu_1,$$
$$\mathbb{P}\left(|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)| > \varepsilon_2\right) \leqslant \nu_2,$$
$$\mathbb{P}\left(|\mathsf{RF}_B^{\backslash i}(X) - \mathsf{rf}^{\backslash i}(X)| > \varepsilon_3\right) \leqslant \nu_3.$$

Now *assume* the test point $(X, Y) \equiv (X_{n+1}, Y_{n+1})$ is accessible to us, and define $n+1$ derandomized RF predictors as follows:

$$\widetilde{\mathsf{rf}}^{\backslash i} \equiv \widetilde{\mathsf{rf}}^{\backslash i}((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_{n+1}, Y_{n+1})),$$

meaning that $\widetilde{\mathsf{rf}}^{\backslash i}$ is trained on n pairs of data points without (X_i, Y_i). Furthermore, define for $i \in [n+1]$ that

$$\widetilde{r}_i = |Y_i - \widetilde{\mathsf{rf}}^{\backslash i}(X_i)|.$$

Lemma 4. *For $\alpha \in (1/(n+1), 1)$, let $\widetilde{r}_i, i \in [n+1]$ be defined above. Then*

$$\mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha}\{\widetilde{r}_i\}\right) \geqslant 1 - \alpha,$$

where $q_{n,\alpha}\{\widetilde{r}_i\}$ denotes the $\lceil(1-\alpha)(n+1)\rceil$-th smallest value of $\{\widetilde{r}_1, \ldots, \widetilde{r}_n\}$. If we further assume there are no ties among $\widetilde{r}_i$, then we also have

$$\mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha}\{\widetilde{r}_i\}\right) \leqslant 1 - \alpha + \frac{1}{n+1}.$$

Proof. This proof is provided for completeness. For the first part, note that all $\widetilde{r}_i$ are exchangeable. Hence the rank of each $\widetilde{r}_i$ is uniformly distributed on $[n+1]$, and

$$
\begin{aligned}
\mathbb{P}\left(\widetilde{r}_{n+1} > q_{n,\alpha}\{\widetilde{r}_i\}\right) &\leqslant \frac{n+1 - \lceil(1-\alpha)(n+1)\rceil}{n+1} \\
&= \frac{\lfloor \alpha(n+1) \rfloor}{n+1} \qquad \text{(note } n+1 = \lceil(n+1)(1-\alpha)\rceil + \lfloor(n+1)\alpha\rfloor) \\
&\leqslant \frac{\alpha(n+1)}{n+1} \\
&= \alpha.
\end{aligned}
$$

For the second part, if there are no ties among all $\{\widetilde{r}_i\}$, then we have

$$\mathbb{P}(\widetilde{r}_{n+1} < \widetilde{r}_{(1)}) = \mathbb{P}\left(\widetilde{r}_{n+1} \text{ is the smallest among } \{\widetilde{r}_i\}_{i=1}^{n+1}\right) = \frac{1}{n+1},$$

and for each $i \in [n-1]$, we have

$$\mathbb{P}\left(\widetilde{r}_{(i)} < \widetilde{r}_{n+1} < \widetilde{r}_{(i+1)}\right) = \mathbb{P}\left(\widetilde{r}_{n+1} \text{ is the } (i+1)\text{-th smallest among } \{\widetilde{r}_i\}_{i=1}^{n+1}\right) = \frac{1}{n+1}.$$

As a result,

$$
\begin{aligned}
\mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha}\{\widetilde{r}_i\}\right) &= \mathbb{P}\left(\widetilde{r}_{n+1} \leqslant \widetilde{r}_{(\lceil(1-\alpha)(n+1)\rceil)}\right) \\
&= \mathbb{P}\left(\widetilde{r}_{n+1} < \widetilde{r}_{(\lceil(1-\alpha)(n+1)\rceil)}\right) \\
&= \mathbb{P}\left(\widetilde{r}_{n+1} < \widetilde{r}_{(1)}\right) + \sum_{i=1}^{\lceil(1-\alpha)(n+1)\rceil - 1} \mathbb{P}\left(\widetilde{r}_{(i)} < \widetilde{r}_{n+1} < \widetilde{r}_{(i+1)}\right) \\
&= \frac{\lceil(1-\alpha)(n+1)\rceil}{n+1}.
\end{aligned}
$$

21

By the definition of the ceiling function, we have
$$(1-\alpha)(n+1) \leqslant \lceil(1-\alpha)(n+1)\rceil \leqslant (1-\alpha)(n+1)+1.$$
Therefore, we conclude that if there are no ties between $\{\widetilde{r}_i\}$, then
$$1-\alpha \leqslant \mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha}\{\widetilde{r}_i\}\right) \leqslant 1-\alpha+\frac{1}{n+1},$$
which completes the proof. $\qquad\square$

By the first part of the above lemma and the definition of $\widetilde{r}_{n+1}$, we know that
$$\mathbb{P}\left(|Y_{n+1}-\widetilde{\mathsf{rf}}^{\backslash(n+1)}(X_{n+1})| \leqslant q_{n,\alpha}\{\widetilde{r}_i\}\right) \geqslant 1-\alpha.$$
But $\widetilde{\mathsf{rf}}^{\backslash(n+1)}$ is just rf, so we have
$$\mathbb{P}\left(|Y_{n+1}-\mathsf{rf}(X_{n+1})| \leqslant q_{n,\alpha}\{\widetilde{r}_i\}\right) \geqslant 1-\alpha.$$
Of course, $\widetilde{r}_i$ are unknown, and we eventually will need to replace them with R_i, but before that, we use r_i instead. Before proceeding, a useful lemma that connects $\{\widetilde{r}_i\}$, $\{r_i\}$, and $\{R_i\}$ is given below.

Lemma 5. *Suppose there are n pairs of real numbers (a_i, b_i), $i \in \{1, \ldots, n\}$. Let $a_{(1)} \leqslant \ldots \leqslant a_{(n)}$ and $b_{(1)} \leqslant \ldots \leqslant b_{(n)}$. For any $\varepsilon \in \mathbb{R}$, if $b_{(k)} > a_{(j)} + \varepsilon$ for some j and k, then there are at least $j-k+1$ pairs of (a_i, b_i) such that $b_i > a_i + \varepsilon$. In particular, if $k = \lceil(n+1)(1-\alpha)\rceil$ and $j = \lceil(n+1)(1-\alpha')\rceil$ with $\alpha' = \alpha - \delta_\alpha$ and $\delta_\alpha \geqslant 0$, then $j-k+1 \geqslant (n+1)\delta_\alpha$.*

Remark 9. It might be more convenient to state the lemma as that if $q_{n,\alpha}\{b_i\} > q_{n,\alpha'}\{a_i\} + \varepsilon$ with $\alpha' = \alpha - \delta_\alpha$, then there are at least $(n+1)\delta_\alpha$ pairs of (a_i, b_i) such that $b_i > a_i + \varepsilon$.

Proof. This result was directly obtained in [3] "by definition of quantiles," while a more complete proof might be helpful. Without loss of generality, we take $\varepsilon = 0$. (Otherwise, we consider $\{a_i', b_i\}$ instead, where $a_i' = a_i + \varepsilon$.) We first define three sets as follows.
$$S_A \equiv \{i : a_i \leqslant a_{(j)}\},$$
$$S_B \equiv \{i : b_i \geqslant b_{(k)}\},$$
$$S_C \equiv \{i : a_i < b_i\}.$$
It is obvious that for any $i \in S_A \cap S_B$, we have $a_i \leqslant a_{(j)} < b_{(k)} \leqslant b_i$. Hence $i \in S_C$. That is,
$$S_A \cap S_B \subseteq S_C.$$
As a consequence, $|S_A \cap S_B| \leqslant |S_C|$, where $|\cdot|$ denotes the cardinality of a set. Then, note that
$$S_A = (S_A \cap S_B) \cup (S_A \cap S_B^c),$$
where S_B^c is the complement of S_B, and
$$|S_A| = |S_A \cap S_B| + |S_A \cap S_B^c|.$$
Therefore, we have
$$|S_C| \geqslant |S_A \cap S_B| = |S_A| - |S_A \cap S_B^c| \geqslant |S_A| - |S_B^c| = j - (k-1) = j-k+1.$$
Now consider $j = \lceil(n+1)(1-\alpha')\rceil$ and $k = \lceil(n+1)(1-\alpha)\rceil$ with $\alpha' = \alpha - \delta_\alpha$. Note for any $\beta \in (0,1)$ that $n+1 = \lceil(n+1)(1-\beta)\rceil + \lfloor(n+1)\beta\rfloor$. Then we have
$$\begin{aligned}
j-k+1 &= \lceil(n+1)(1-\alpha')\rceil - \lceil(n+1)(1-\alpha)\rceil + 1 \\
&= \lfloor(n+1)\alpha\rfloor - \lfloor(n+1)\alpha'\rfloor + 1 \\
&= \lfloor(n+1)(\alpha'+\delta_\alpha)\rfloor - \lfloor(n+1)\alpha'\rfloor + 1 \\
&\geqslant \lfloor(n+1)\alpha'\rfloor + \lfloor(n+1)\delta_\alpha\rfloor - \lfloor(n+1)\alpha'\rfloor + 1 \\
&= \lfloor(n+1)\delta_\alpha\rfloor + 1 \\
&\geqslant (n+1)\delta_\alpha,
\end{aligned}$$
completing the proof. $\qquad\square$

We are now ready to prove Theorem 9. Basically, we use the stability property once and the concentration of measure twice to establish probabilistic deviation bounds for $|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)|$, $|\mathsf{rf}(X) - \mathsf{RF}_B(X)|$, and $|\mathsf{rf}^{\backslash i}(X) - \mathsf{RF}_B^{\backslash i}(X)|$, respectively.

22

D.1 Using the stability property to control $|\mathsf{rf}(X) - \mathsf{rf}^{\setminus i}(X)|$

Following the same idea as in [3], we consider the event

$$A_l = \text{``} q_{n,\alpha}\{\tilde{r}_i\} > q_{n,\alpha_2}\{r_i\} + \varepsilon_2 \text{''},$$

where $\alpha_2 = \alpha - \sqrt{\nu_2}$. We then have that

$$1 - \alpha \leqslant \mathbb{P}\left(|Y_{n+1} - \mathsf{rf}(X_{n+1})| \leqslant q_{n,\alpha}\{\tilde{r}_i\} \text{ and } A_l\right) + \mathbb{P}\left(|Y_{n+1} - \mathsf{rf}(X_{n+1})| \leqslant q_{n,\alpha}\{\tilde{r}_i\} \text{ and } A_l^c\right)$$
$$\leqslant \mathbb{P}(A_l) + \mathbb{P}\left(\tilde{r}_{n+1} \leqslant q_{n,\alpha_2}\{r_i\} + \varepsilon_2\right),$$

where A_l^c is the complement of A_l. Now, we calculate $\mathbb{P}(A_l)$. By Lemma 5, event A_l implies that there are at least $(n+1)\sqrt{\nu_2}$ pairs of $(\tilde{r}_i, r_i)$ such that $\tilde{r}_i > r_i + \varepsilon_2$, and thus

$$\mathbb{P}(A_l) \leqslant \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}\{\tilde{r}_i > r_i + \varepsilon_2\} \geqslant (n+1)\sqrt{\nu_2}\right)$$

$$\leqslant \frac{\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{I}\{\tilde{r}_i > r_i + \varepsilon_2\}\right]}{(n+1)\sqrt{\nu_2}} \text{ (by Markov's inequality)}$$

$$= \frac{n\mathbb{E}\left[\mathbb{I}\{\tilde{r}_i > r_i + \varepsilon_2\}\right]}{(n+1)\sqrt{\nu_2}} \text{ (by iid data)}$$

$$= \frac{n\mathbb{P}\left(\tilde{r}_i > r_i + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}}$$

$$= \frac{n\mathbb{P}\left(|Y_i - \widetilde{\mathsf{rf}}^{\setminus i}(X_i)| > |Y_i - \mathsf{rf}^{\setminus i}(X_i)| + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}}$$

$$= \frac{n\mathbb{P}\left(|Y_i - \widetilde{\mathsf{rf}}^{\setminus i}(X_i)| > |Y_i - \widetilde{\mathsf{rf}}^{\setminus (n+1,i)}(X_i)| + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \text{ (by definitions of } \widetilde{\mathsf{rf}}^{\setminus i}, \mathsf{rf}, \mathsf{rf}^{\setminus i})$$

$$= \frac{n\mathbb{P}\left(|Y_{n+1} - \widetilde{\mathsf{rf}}^{\setminus (n+1)}(X_{n+1})| > |Y_{n+1} - \widetilde{\mathsf{rf}}^{\setminus (i,n+1)}(X_{n+1})| + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \text{ (by iid data)}$$

$$= \frac{n\mathbb{P}\left(|Y_{n+1} - \mathsf{rf}(X_{n+1})| > |Y_{n+1} - \mathsf{rf}^{\setminus i}(X_{n+1})| + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \text{ (by definitions of } \widetilde{\mathsf{rf}}^{\setminus i}, \mathsf{rf}, \mathsf{rf}^{\setminus i})$$

$$\leqslant \frac{n\mathbb{P}\left(|Y_{n+1} - \mathsf{rf}(X_{n+1}) - Y_{n+1} + \mathsf{rf}^{\setminus i}(X_{n+1})| > \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \text{ (because } |a| - |b| \leqslant |a - b|)$$

$$= \frac{n\mathbb{P}\left(|\mathsf{rf}(X_{n+1}) - \mathsf{rf}^{\setminus i}(X_{n+1})| > \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}}$$

$$\leqslant \frac{n\nu_2}{(n+1)\sqrt{\nu_2}} \text{ (by stability of rf)}$$

$$\leqslant \sqrt{\nu_2}.$$

As a consequence, we have

$$\mathbb{P}\left(|Y_{n+1} - \mathsf{rf}(X_{n+1})| \leqslant q_{n,\alpha_2}\{r_i\} + \varepsilon_2\right) \geqslant 1 - \alpha - \mathbb{P}(A_l) \geqslant 1 - \alpha - \sqrt{\nu_2} = 1 - \alpha_2 - 2\sqrt{\nu_2}.$$

That is to say, we have a reduced lower bound of coverage using an ε_2-inflated interval constructed from $\{r_i\}$. However, $\{r_i\}$ is unknown, and we want to further approximate r_i by R_i.

D.2 Using the concentration of measure to control $|\mathsf{rf}^{\setminus i}(X) - \mathsf{RF}_B^{\setminus i}(X)|$

To this end, we similarly define another event:

$$A_l' = \text{``} q_{n,\alpha_2}\{r_i\} > q_{n,\alpha_3}\{R_i\} + \varepsilon_3 \text{''},$$

23

where $\alpha_3 = \alpha_2 - \sqrt{\nu_3}$, and we have

$$1 - \alpha_2 - 2\sqrt{\nu_2}$$
$$\leqslant \mathbb{P}\left(|Y_{n+1} - \mathsf{rf}(X_{n+1})| \leqslant q_{n,\alpha_2}\{r_i\} + \varepsilon_2 \text{ and } A_l'\right) + \mathbb{P}\left(|Y_{n+1} - \mathsf{rf}(X_{n+1})| \leqslant q_{n,\alpha_2}\{r_i\} + \varepsilon_2 \text{ and } A_l'^c\right)$$
$$\leqslant \mathbb{P}(A_l') + \mathbb{P}\left(|Y_{n+1} - \mathsf{rf}(X_{n+1})| \leqslant q_{n,\alpha_3}\{R_i\} + \varepsilon_2 + \varepsilon_3\right),$$

where $A_l'^c$ is the complement of A_l'. Applying Lemma 5 again yields

$$\mathbb{P}(A_l') \leqslant \mathbb{P}\left(\sum_{i=1}^n \mathbb{I}\{r_i > R_i + \varepsilon_3\} \geqslant (n+1)\sqrt{\nu_3}\right)$$
$$\leqslant \frac{\mathbb{E}\left[\sum_{i=1}^n \mathbb{I}\{r_i > R_i + \varepsilon_3\}\right]}{(n+1)\sqrt{\nu_3}} \text{ (by Markov's inequality)}$$
$$\leqslant \frac{n\mathbb{P}\left(r_i > R_i + \varepsilon_3\right)}{(n+1)\sqrt{\nu_3}} \text{ (by iid data)}$$
$$= \frac{n\mathbb{P}\left(|Y_i - \mathsf{rf}^{\backslash i}(X_i)| > |Y_i - \mathsf{RF}_B^{\backslash i}(X_i)| + \varepsilon_3\right)}{(n+1)\sqrt{\nu_3}}$$
$$\leqslant \frac{n\mathbb{P}\left(|Y_i - \mathsf{rf}^{\backslash i}(X_i) - Y_i + \mathsf{RF}_B^{\backslash i}(X_i)| > \varepsilon_3\right)}{(n+1)\sqrt{\nu_3}} \text{ (because } |a| - |b| \leqslant |a - b|)$$
$$= \frac{n\mathbb{P}\left(|\mathsf{rf}^{\backslash i}(X_i) - \mathsf{RF}_B^{\backslash i}(X_i)| > \varepsilon_3\right)}{(n+1)\sqrt{\nu_3}}$$
$$= \frac{n\mathbb{P}\left(|\mathsf{rf}^{\backslash i}(X) - \mathsf{RF}_B^{\backslash i}(X)| > \varepsilon_3\right)}{(n+1)\sqrt{\nu_3}} \text{ (because } X_i \text{ and } X \text{ are iid, and } \mathsf{rf}^{\backslash i} \text{ and } \mathsf{RF}_B^{\backslash i} \text{ do not depend on } X_i)$$
$$\leqslant \frac{n\nu_3}{(n+1)\sqrt{\nu_3}} \text{ (by concentration of measure)}$$
$$\leqslant \sqrt{\nu_3}.$$

So we have

$$\mathbb{P}\left(|Y_{n+1} - \mathsf{rf}(X_{n+1})| \leqslant q_{n,\alpha_3}\{R_i\} + \varepsilon_2 + \varepsilon_3\right) \geqslant 1 - \alpha_2 - 2\sqrt{\nu_2} - \mathbb{P}(A_l') \geqslant 1 - \alpha_3 - 2\sqrt{\nu_2} - 2\sqrt{\nu_3}.$$

D.3 Using the concentration of measure to control $|\mathsf{rf}(X) - \mathsf{RF}_B(X)|$

Finally, we need to replace $\mathsf{rf}(X_{n+1})$ by $\mathsf{RF}_B(X_{n+1})$. Note that for $t > 0$,

$$\mathbb{P}\left(|Y - \mathsf{RF}_B(X)| > t + \varepsilon_1\right) = \mathbb{P}\left(|Y - \mathsf{rf}(X) + \mathsf{rf}(X) - \mathsf{RF}_B(X)| > t + \varepsilon_1\right)$$
$$\leqslant \mathbb{P}\left(|Y - \mathsf{rf}(X)| > t\right) + \mathbb{P}\left(|\mathsf{rf}(X) - \mathsf{RF}_B(X)| > \varepsilon_1\right)$$
$$\leqslant \mathbb{P}\left(|Y - \mathsf{rf}(X)| > t\right) + \nu_1,$$

which implies

$$\mathbb{P}\left(|Y - \mathsf{RF}_B(X)| \leqslant t + \varepsilon_1\right) \geqslant \mathbb{P}\left(|Y - \mathsf{rf}(X)| \leqslant t\right) - \nu_1.$$

Let $t = q_{n,\alpha_3}\{R_i\} + \varepsilon_2 + \varepsilon_3$, and we arrive at

$$\mathbb{P}\left(|Y - \mathsf{RF}_B(X)| \leqslant q_{n,\alpha_3}\{R_i\} + \varepsilon_2 + \varepsilon_3 + \varepsilon_1\right) \geqslant \mathbb{P}\left(|Y - \mathsf{rf}(X)| \leqslant q_{n,\alpha_3}\{R_i\} + \varepsilon_2 + \varepsilon_3\right) - \nu_1$$
$$\geqslant 1 - \alpha_3 - 2\sqrt{\nu_2} - 2\sqrt{\nu_3} - \nu_1,$$

which completes the proof of Theorem 9. Moreover, for $\nu_1 \in (0, 1)$, we have $\sqrt{\nu_1} > \nu_1$. So

$$2\sqrt{\nu_2} + 2\sqrt{\nu_3} + \nu_1 \leqslant 2(\sqrt{\nu_2} + \sqrt{\nu_3} + \sqrt{\nu_1}) = 6 \times \frac{\sqrt{\nu_2} + \sqrt{\nu_3} + \sqrt{\nu_1}}{3} \leqslant 6 \times \sqrt{(\nu_1 + \nu_2 + \nu_3)/3} = 2\sqrt{3}\sqrt{\nu_{n,B}}.$$

Hence we eventually have

$$\mathbb{P}\left(|Y - \mathsf{RF}_B(X)| \leqslant q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}\right) \leqslant 1 - \alpha - 2\sqrt{3}\sqrt{\nu_{n,B}},$$

which leads to the informal version of the theorem.

24

E Proof of Theorem 10

E.1 Using the stability property to control $|\mathsf{rf}(X) - \mathsf{rf}^{\backslash i}(X)|$

Since we have assumed $\{\widetilde{r}_i\}, i \in [n+1]$ have no ties, by the second part of Lemma 4, we have for the test point $(X, Y) = (X_{n+1}, Y_{n+1})$ that

$$\mathbb{P}\left(|Y_{n+1} - \widetilde{\mathsf{rf}}^{\backslash(n+1)}(X_{n+1})| \leqslant q_{n,\alpha}\{\widetilde{r}_i\}\right) = \mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha}\{\widetilde{r}_i\}\right) \leqslant 1 - \alpha + \frac{1}{n+1}.$$

Now consider the event that

$$A_u = \text{``} q_{n,\alpha_2}\{r_i\} - \varepsilon_2 \leqslant q_{n,\alpha}\{\widetilde{r}_i\} \text{''},$$

where $\alpha_2 = \alpha + \sqrt{\nu_2}$. We denote A_u^c as the complement of A_u. Note that in this case the ancillary quantity α_2 is greater than α, while in the proof of Theorem 9, α_2 is less than α. We then have

$$
\begin{aligned}
\mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha_2}\{r_i\} - \varepsilon_2\right) &= \mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha_2}\{r_i\} - \varepsilon_2 \text{ and } A_u\right) + \mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha_2}\{r_i\} - \varepsilon_2 \text{ and } A_u^c\right) \\
&\leqslant \mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha_2}\{r_i\} - \varepsilon_2 \text{ and } A_u\right) + \mathbb{P}(A_u^c) \\
&\leqslant \mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha}\{\widetilde{r}_i\}\right) + \mathbb{P}(A_u^c) \\
&\leqslant 1 - \alpha + \frac{1}{n+1} + \mathbb{P}(A_u^c).
\end{aligned}
$$

Next, we bound $\mathbb{P}(A_u^c)$. By Lemma 5, event A_u^c implies there are at least $(n+1)\sqrt{\nu_2}$ pairs of $(r_i, \widetilde{r}_i)$ such that $r_i > \widetilde{r}_i + \varepsilon_2$, and thus

$$
\begin{aligned}
\mathbb{P}(A_u^c) &\leqslant \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}\{r_i > \widetilde{r}_i + \varepsilon_2\} \geqslant (n+1)\sqrt{\nu_2}\right) \\
&\leqslant \frac{\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{I}\{r_i > \widetilde{r}_i + \varepsilon_2\}\right]}{(n+1)\sqrt{\nu_2}} \text{ (by Markov's inequality)} \\
&= \frac{n\mathbb{E}\left[\mathbb{I}\{r_i > \widetilde{r}_i + \varepsilon_2\}\right]}{(n+1)\sqrt{\nu_2}} \text{ (by iid data)} \\
&= \frac{n\mathbb{P}\left(r_i > \widetilde{r}_i + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \\
&= \frac{n\mathbb{P}\left(|Y_i - \mathsf{rf}^{\backslash i}(X_i)| > |Y_i - \widetilde{\mathsf{rf}}^{\backslash i}(X_i)| + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \\
&= \frac{n\mathbb{P}\left(|Y_i - \widetilde{\mathsf{rf}}^{\backslash(n+1,i)}(X_i)| > |Y_i - \widetilde{\mathsf{rf}}^{\backslash i}(X_i)| + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \text{ (by definitions of } \widetilde{\mathsf{rf}}^{\backslash i}, \mathsf{rf}, \mathsf{rf}^{\backslash i}) \\
&= \frac{n\mathbb{P}\left(|Y_{n+1} - \widetilde{\mathsf{rf}}^{\backslash(i,n+1)}(X_{n+1})| > |Y_{n+1} - \widetilde{\mathsf{rf}}^{\backslash(n+1)}(X_{n+1})| + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \text{ (by iid data)} \\
&= \frac{n\mathbb{P}\left(|Y_{n+1} - \mathsf{rf}^{\backslash i}(X_{n+1})| > |Y_{n+1} - \mathsf{rf}(X_{n+1})| + \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \text{ (by definitions of } \widetilde{\mathsf{rf}}^{\backslash i}, \mathsf{rf}, \mathsf{rf}^{\backslash i}) \\
&\leqslant \frac{n\mathbb{P}\left(|Y_{n+1} - \mathsf{rf}^{\backslash i}(X_{n+1}) - Y_{n+1} + \mathsf{rf}(X_{n+1})| > \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \text{ (because } |a| - |b| \leqslant |a - b|) \\
&= \frac{n\mathbb{P}\left(|\mathsf{rf}(X_{n+1}) - \mathsf{rf}^{\backslash i}(X_{n+1})| > \varepsilon_2\right)}{(n+1)\sqrt{\nu_2}} \\
&\leqslant \frac{n\nu_2}{(n+1)\sqrt{\nu_2}} \text{ (by stability of rf)} \\
&\leqslant \sqrt{\nu_2}.
\end{aligned}
$$

25

Hence we have

$$\mathbb{P}\left(\tilde{r}_{n+1} \leqslant q_{n,\alpha_2}\{r_i\} - \varepsilon_2\right) \leqslant 1 - \alpha + \frac{1}{n+1} + \sqrt{\nu_2} = 1 - \alpha_2 + \frac{1}{n+1} + 2\sqrt{\nu_2}.$$

Note that this proof also works for a general stable algorithm, as stated in Corollary 3.

E.2 Using the concentration of measure to control $|\mathsf{rf}^{\backslash i}(X) - \mathsf{RF}_B^{\backslash i}(X)|$

We further define an event

$$A'_u = \text{``}q_{n,\alpha_3}\{R_i\} - \varepsilon_3 \leqslant q_{n,\alpha_2}\{r_i\}\text{''},$$

where $\alpha_3 = \alpha_2 + \sqrt{\nu_3}$. We denote A'^c_u as the complement of A'_u. Again, we increase rather than decrease α_3 in this case, as opposed to in the proof of the lower bound. We then have

$$\mathbb{P}(\tilde{r}_{n+1} \leqslant q_{n,\alpha_3}\{R_i\} - \varepsilon_2 - \varepsilon_3)$$

$$= \mathbb{P}\left(\tilde{r}_{n+1} \leqslant q_{n,\alpha_3}\{R_i\} - \varepsilon_2 - \varepsilon_3 \text{ and } A'_u\right) + \mathbb{P}\left(\tilde{r}_{n+1} \leqslant q_{n,\alpha_3}\{R_i\} - \varepsilon_2 - \varepsilon_3 \text{ and } A'^c_u\right)$$

$$\leqslant \mathbb{P}\left(\tilde{r}_{n+1} \leqslant q_{n,\alpha_3}\{R_i\} - \varepsilon_2 - \varepsilon_3 \text{ and } A'_u\right) + \mathbb{P}(A'^c_u)$$

$$\leqslant \mathbb{P}\left(\tilde{r}_{n+1} \leqslant q_{n,\alpha_2}\{r_i\} - \varepsilon_2\right) + \mathbb{P}(A'^c_u)$$

$$\leqslant 1 - \alpha_2 + \frac{1}{n+1} + 2\sqrt{\nu_2} + \mathbb{P}(A'^c_u).$$

By Lemma 5, A'^c_u implies there exist at least $(n+1)\sqrt{\nu_3}$ pairs of (R_i, r_i) such that $R_i > r_i + \varepsilon_3$, and thus

$$\mathbb{P}(A'^c_u) \leqslant \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}\{R_i > r_i + \varepsilon_3\} \geqslant (n+1)\sqrt{\nu_3}\right)$$

$$\leqslant \frac{\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{I}\{R_i > r_i + \varepsilon_3\}\right]}{(n+1)\sqrt{\nu_3}} \text{ (by Markov's inequality)}$$

$$= \frac{n\mathbb{P}(R_i > r_i + \varepsilon_3)}{(n+1)\sqrt{\nu_3}} \text{ (by iid data)}$$

$$= \frac{n\mathbb{P}\left(|Y_i - \mathsf{RF}_B^{\backslash i}(X_i)| > |Y_i - \mathsf{rf}^{\backslash i}(X_i)| + \varepsilon_3\right)}{(n+1)\sqrt{\nu_3}}$$

$$\leqslant \frac{n\mathbb{P}\left(|Y_i - \mathsf{RF}_B^{\backslash i}(X_i) - Y_i + \mathsf{rf}^{\backslash i}(X_i)| > \varepsilon_3\right)}{(n+1)\sqrt{\nu_3}} \text{ (because } |a| - |b| \leqslant |a - b|)$$

$$= \frac{n\mathbb{P}\left(|\mathsf{rf}^{\backslash i}(X_i) - \mathsf{RF}_B^{\backslash i}(X_i)| > \varepsilon_3\right)}{(n+1)\sqrt{\nu_3}}$$

$$= \frac{n\mathbb{P}\left(|\mathsf{rf}^{\backslash i}(X) - \mathsf{RF}_B^{\backslash i}(X)| > \varepsilon_3\right)}{(n+1)\sqrt{\nu_3}} \text{ (because } X_i \text{ and } X \text{ are iid, and } \mathsf{rf}^{\backslash i} \text{ and } \mathsf{RF}_B^{\backslash i} \text{ do not depend on } X_i)$$

$$\leqslant \frac{n\nu_3}{(n+1)\sqrt{\nu_3}} \text{ (by concentration of measure)}$$

$$\leqslant \sqrt{\nu_3}.$$

We then have

$$\mathbb{P}(\tilde{r}_{n+1} \leqslant q_{n,\alpha_3}\{R_i\} - \varepsilon_2 - \varepsilon_3) \leqslant 1 - \alpha_3 + \frac{1}{n+1} + 2\sqrt{\nu_2} + 2\sqrt{\nu_3}.$$

26

E.3 Using the concentration of measure to control $|\mathsf{rf}(X) - \mathsf{RF}_B(X)|$

We write $R_{n+1} = |Y_{n+1} - \mathsf{RF}_B(X_{n+1})|$. Then

$$
\mathbb{P}(R_{n+1} \leqslant q_{n,\alpha_3}\{R_i\} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3)
$$
$$
= \mathbb{P}\left(R_{n+1} - \widetilde{r}_{n+1} + \widetilde{r}_{n+1} \leqslant q_{n,\alpha_3}\{R_i\} - \varepsilon_1 - \varepsilon_2 - \varepsilon_3\right)
$$
$$
\leqslant \mathbb{P}\left(R_{n+1} - \widetilde{r}_{n+1} < -\varepsilon_1\right) + \mathbb{P}\left(\widetilde{r}_{n+1} \leqslant q_{n,\alpha_3}\{R_i\} - \varepsilon_2 - \varepsilon_3\right)
$$
$$
\leqslant \mathbb{P}\left(|R_{n+1} - \widetilde{r}_{n+1}| > \varepsilon_1\right) + 1 - \alpha_3 + \frac{1}{n+1} + 2\sqrt{\nu_2} + 2\sqrt{\nu_3}
$$
$$
= \mathbb{P}\left(||Y_{n+1} - \mathsf{RF}_B(X_{n+1})| - |Y_{n+1} - \mathsf{rf}(X_{n+1})|| > \varepsilon_1\right) + 1 - \alpha_3 + \frac{1}{n+1} + 2\sqrt{\nu_2} + 2\sqrt{\nu_3}
$$
$$
\leqslant \mathbb{P}\left(|\mathsf{rf}(X) - \mathsf{RF}_B(X)| > \varepsilon_1\right) + 1 - \alpha_3 + \frac{1}{n+1} + 2\sqrt{\nu_2} + 2\sqrt{\nu_3} \text{ (because } ||a| - |b|| \leqslant |a - b|)
$$
$$
\leqslant 1 - \alpha_3 + \frac{1}{n+1} + \nu_1 + 2\sqrt{\nu_2} + 2\sqrt{\nu_3}, \text{ (by concentration of measure)}
$$

which completes the proof. Again, for $\nu_1 \in (0, 1)$, the upper bound is $1 - \alpha_3 + \frac{1}{n+1} + O(\sqrt{\nu_{n,B}})$.

The proof of Theorem 9 bears some resemblance to that of Theorem 10 because of the symmetry consideration. For example, "$\mathsf{rf}(X)$ is close to $\mathsf{rf}^{\backslash i}(X)$" also means "$\mathsf{rf}^{\backslash i}(X)$ is close to $\mathsf{rf}(X)$." Hence the probabilistic deviation bounds established above apply in two directions, and the upper bound can be proved. Still, we use the stability property once and the concentration of measure twice.

F Proof of Theorem 11

We start from $\mathbb{P}(|Y - \mathsf{RF}_B(X)| \leqslant q_{n,\alpha}\{R_i\} + \varepsilon_{n,B})$. Denote $R = |Y - \mathsf{RF}_B(X)|$, and let its CDF be F_n, where we explicitly use n to denote the size of the training set. Then we can rewrite $\mathbb{P}(|Y - \mathsf{RF}_B(X)| \leqslant q_{n,\alpha}\{R_i\} + \varepsilon_{n,B})$ as

$$
\mathbb{P}(|Y - \mathsf{RF}_B(X)| \leqslant q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) = \mathbb{E}[F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B})],
$$

where the expectation is with respect to $q_{n,\alpha}\{R_i\}$, which is random. By the assumption, when $n \geqslant n_0$, the family $\{F_n(t)\}_{n \geqslant n_0}$ is uniformly equicontinuous, which means that for any $\nu > 0$, there exists some $\delta > 0$ that is independent of t and n, such that $|F_n(t') - F_n(t)| \leqslant \nu$ for all $|t' - t| \leqslant \delta$ and $n \geqslant n_0$. Then we have for $n \geqslant n_0$ and $\nu > 0$ that

$$
F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\})
$$
$$
= F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\} + \delta) + F_n(q_{n,\alpha}\{R_i\} + \delta) - F_n(q_{n,\alpha}\{R_i\})
$$
$$
\leqslant \sup_t \left[F_n(t + \varepsilon_{n,B}) - F_n(t + \delta)\right] + \sup_t \left[F_n(t + \delta) - F_n(t)\right]
$$
$$
\leqslant \sup_t \left[F_n(t + \varepsilon_{n,B}) - F_n(t + \delta)\right] + \nu.
$$

However, $F_n(t)$ as a CDF is monotonically increasing, and for any t, $F_n(t + \varepsilon_{n,B}) - F_n(t + \delta) \leqslant 0$ as long as $\varepsilon_{n,B} \leqslant \delta$. By Corollary 2, $\lim_{n \to \infty} \varepsilon_{n,B} = 0$. So there exists some n_1 such that when $n \geqslant n_1$, we have $\varepsilon_{n,B} \leqslant \delta$. Hence we conclude that for any $\nu > 0$, for all $n \geqslant \max\{n_0, n_1\}$

$$
0 \leqslant F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\}) \leqslant \nu \text{ a.s.},
$$

implying that

$$
\lim_{n \to \infty}\{F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\})\} = 0 \text{ a.s.}
$$

As F_n is bounded, then by the bounded dominance theorem, we have

$$
\lim_{n \to \infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\})] = \mathbb{E}[\lim_{n \to \infty}\{F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\})\}] = 0.
$$

Under the conditions of Corollary 2, both $\varepsilon_{n,B}$ and $\nu_{n,B}$ go to 0 as $n \to \infty$, and the lower bound (11) implies that

$$
1 - \alpha \leqslant \liminf_{n \to \infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B})]
$$
$$
= \liminf_{n \to \infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\}) + F_n(q_{n,\alpha}\{R_i\})]
$$
$$
= \lim_{n \to \infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\})] + \liminf_{n \to \infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\})]
$$
$$
= \liminf_{n \to \infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\})].
$$

27

Similarly, the upper bound (12) implies that

$$1 - \alpha \geqslant \limsup_{n\to\infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B})]$$

$$= \limsup_{n\to\infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\}) + F_n(q_{n,\alpha}\{R_i\})]$$

$$= \lim_{n\to\infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\} + \varepsilon_{n,B}) - F_n(q_{n,\alpha}\{R_i\})] + \limsup_{n\to\infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\})]$$

$$= \limsup_{n\to\infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\})].$$

Combining these results, we have

$$1 - \alpha \leqslant \liminf_{n\to\infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\})] \leqslant \limsup_{n\to\infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\})] \leqslant 1 - \alpha.$$

Thus,

$$\lim_{n\to\infty} \mathbb{P}(R \leqslant q_{n,\alpha}\{R_i\}) = \lim_{n\to\infty} \mathbb{E}[F_n(q_{n,\alpha}\{R_i\})] = 1 - \alpha,$$

completing the proof.

28