Degree Papers for Masters in Statistics

Chen Xu

Advisor: Rina Foygel Barber

Approved _____ (Signatures of the advisors) _____

Date _____ (Date of the signatures) 5/5/20

May-5th, 2020

Efficient Predictive Inference with Jackknife+ under Ensemble Learning

Abstract

Ensemble learning is widely used in applications to make predictions in complex decision problems—for example, averaging models fitted to a sequence of samples bootstrapped from the available training data. While such methods offer more accurate, stable, and robust predictions and model estimates, much less is known about how to perform valid, assumption-lean inference on the output of these types of procedures. In this paper, we compare the previously proposed jackknife+-after-bootstrap (J+aB) with another variant called the jackknife+-with-bootstrap (J+wB), both of which are computationally efficient and theoretically valid methods for distribution-free predictive inference under ensemble learning. Both methods offer predictive coverage guarantee that holds with minimal assumptions. In particular, we do not assume the distribution of the data, the nature of the fitted model, or how the individual bootstrap estimators are aggregated—at worst, the failure rate of the predictive interval is non-asymptotically inflated by either a factor of 2 for J+aB or a factor of 2 plus an explicit analytic expression that can be made arbitrarily small for J+wB. Our numerical experiments verify the coverage and accuracy of the resulting predictive intervals on real data.

Keywords: Assumption-free inference; Bagging; Bootstrapping; Conformal inference; Ensemble learning; Exchangeability; Jackknife; Predictive inference; Stability

Contents

Acknowledgements

I want to thank **my parents**, Yongmei Li and Xiaobing Xu, and **my grandma**, Ping Fan, for their love, support, and dedicated upbringing. Without them, I would never have the opportunity to study abroad and get exposed to different cultures, let alone completing this work. Their unconditional love and teaching through examples constantly support me and move me forward through life. I want to thank all my other family members who have contributed to my growth with their love and support.

I want to thank **my fiancee**, Haoruo Zhao, for being one of the most supporting and helpful people in my life. I remember vividly countless times when she provides suggestions to my work, encourages me emotionally, and dedicates her time for me. She has taught me how to become more caring, thoughtful, and grateful in life. Besides my parents and grandma, she is undoubtedly my most important person in life.

I want to thank **Professor Rina Barber and Byol Kim**, both of whom are essential to the success of this work. Professor Rina Barber, through her teaching and encouragement, teaches me the rigor of doing research and transforms me into an aspiring young researcher. She is definitely the most important teacher during my education at the University of Chicago and had I not done research with her, I would not have applied PhD programs and being fortunate to go Georgia Tech ISyE in the future. My colleague Byol Kim teaches me the details of completing a research, advises me against the common pitfalls, and through contributing her knowledge, makes this work possible. Overall, I learned so much from both of them, even just witnessing the discussion between them during weekly meetings.

Lastly, I want to thank my friends, other faculty members, and anyone who has helped me throughout my learning and education. I wish you all the best in your life and first and foremost, stay safe during this pandemic.

1 Introduction

1.1 Distribution-Free Predictive Interval/Set

First, we introduce the problem of quantifying uncertainty in regression by providing a distribution-free predictive inference interval around the predicted value $\hat{\mu}(X_{n+1})$. Formally, we state the problem as follows:

Suppose we are given n independent and identically distributed (i.i.d.) observations

$$(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} \mathcal{P}$$

for some probability distribution $\mathcal{P}$ on $\mathbb{R}^p \times \mathbb{R}$. Given the available training data, we would now like to predict the value of the response Y_{n+1} for a new data point with features X_{n+1}, where we assume that (X_{n+1}, Y_{n+1}) is drawn from the same probability distribution $\mathcal{P}$. One common strategy is to fit a regression model $\hat{\mu} : \mathbb{R}^p \to \mathbb{R}$ by applying some regression algorithm to the training data $\{(X_i, Y_i)\}_{i=1}^n$, and then predicting $\hat{\mu}(X_{n+1})$ as our best estimate of the unseen test response Y_{n+1}. Yet, we are interested in quantifying the accuracy or error level of this prediction and in particular, we want to use the available information to build an interval around our estimate $\hat{\mu}(X_{n+1}) \pm$ (some margin of error) that we believe is likely to contain Y_{n+1}.

In general, we call $\hat{C} := \hat{C}(X_{n+1}) \subseteq \mathbb{R}$ a *predictive interval/set* if it maps X_{n+1} to a interval/set that is believed to contain Y_{n+1}.

Moreover, $\hat{C}$ satisfies *distribution-free predictive coverage* at level $1 - \alpha$ if

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}(X_{n+1})\right] \geq 1 - \alpha$$

for any distribution $\mathcal{P}$ of the data. This probability is with respect to the distribution of the $n + 1$ training and test data points, as well as any additional source of randomness used in obtaining $\hat{C}$ as $\hat{C}$ is implicitly a function over the training data as well. The bound must hold uniformly over all distributions $\mathcal{P}$.

1.2 The Jackknife and Jackknife+ Methods

In this subsection, we briefly go over common strategies to obtain $\hat{C}$ in distribution-free fashions and in particular, introduce the Jackknife and Jackknife+ methods that serve as motivations for our proposed methods.

Distribution-free prediction methods have garnered attention in recent years as wrapper methods for complex machine learning algorithms such as neural networks. The use of holdout or validation sets is a common, and computationally inexpensive, way to avoid overfitting and ensure distribution-free predictive coverage [Papadopoulos, 2008, Vovk, 2013, Lei et al., 2018], while methods such as cross-validation or leave-one-out cross-validation (also called the "jackknife") stabilize the results in practice but require some assumptions to analyze theoretically [Steinberger and Leeb, 2016, 2018, Barber et al., 2019]. Distribution-free guarantees are also obtained by the conformal prediction methodology of Vovk et al. [2005] [see also Lei et al., 2018].

Our method, as will be described in later sections, is inspired by the recent jackknife+ of Barber et al. [2019]. As suggested by the name, the jackknife+ is a simple modification of the jackknife approach to constructing predictive confidence intervals.

To briefly explain jackknife and jackknife+, we first define some notations.

1. $\mathcal{R}$ denotes any regression algorithm that maps a collection of training data set to a fitted regression function $\hat{\mu}$, which maps any new data X to a predicted value $\hat{Y} := \hat{\mu}(X)$.

   Compactly, if we have n training data points, $\hat{\mu} = \mathcal{R}(\{(X_i, Y_i)\}_{i=1}^n)$ and $\hat{\mu}_{\setminus i} = \mathcal{R}(\{(X_j, Y_j)\}_{j=1, j\neq i}^n)$ is the algorithm fitted on all but the i-th data point in the training data.

2. Given a collection of n numerical values indexed by i, $q^+_{\alpha,n}\{v_i\}$ and $q^-_{\alpha,n}\{v_i\}$ are the upper and lower α-quantiles, namely,

$$q^+_{\alpha,n}\{v_i\} = \text{ the } \lceil (1-\alpha)(n+1) \rceil \text{ th smallest value of } v_1, \ldots, v_n,$$

$$q^-_{\alpha,n}\{v_i\} = \text{ the } \lceil \alpha(n+1) \rceil \text{ th smallest value of } v_1, \ldots, v_n,$$

Under these notations, the usual jackknife prediction interval is given by

$$\hat{C}^{\mathrm{J}}_{\alpha,n}(x) = \hat{\mu}(x) \pm q^+_{\alpha,n}\{R_i\} = \left[q^-_{\alpha,n}\{\hat{\mu}(x) - R_i\}, q^+_{\alpha,n}\{\hat{\mu}(x) + R_i\}\right], \tag{1}$$

where $R_i = |Y_i - \hat{\mu}_{\setminus i}(X_i)|$ is the i-th leave-one-out residual. The intuitive reason why this construction should work is that the R_i's can well approximate the test residual $|Y_{n+1} - \hat{\mu}_{\setminus i}(X_{n+1})|$, because $\hat{\mu}_{\setminus i}$ is trained without using (X_i, Y_i). Surprisingly, however, fully assumption-free theoretical guarantees are impossible to achieve for the jackknife construction [see Barber et al., 2019, Theorem 2].

As an improvement, the jackknife+ replaces $\hat{\mu}$ in (1) with $\hat{\mu}_{\setminus i}$'s:

$$\hat{C}^{\mathrm{J+}}_{\alpha,n}(x) = \left[q^-_{\alpha,n}\{\hat{\mu}_{\setminus i}(x) - R_i\}, q^+_{\alpha,n}\{\hat{\mu}_{\setminus i}(x) + R_i\}\right].$$

It can be shown that $\hat{C}^{\mathrm{J+}}_{\alpha,n}(X_{n+1})$ satisfies

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}^{\mathrm{J+}}_{\alpha,n}(X_{n+1})\right] \geq 1 - 2\alpha$$

for any sample size n, irrespective of the data distribution and the choice of regression method, achieving an assumption-free, non-asymptotic coverage guarantee.

Intuitively, jackknife fails to achieve such a guarantee without additional assumptions because the test residual $|Y_{n+1} - \hat{\mu}(X_{n+1})|$ can be incomparable with the leave-one-out residuals $|Y_i - \hat{\mu}_{\setminus i}(X_i)|$. The former predictor $\hat{\mu}$ always uses one more observation to train the regression algorithm compared to the latter, so that for $\mathcal{R}$ that is sensitive to the perturbation of training data, the resulting predicted values by $\hat{\mu}$ can differ dramatically from those by $\hat{\mu}_{\setminus i}$.

1.3 Ensemble Learning Basics

After briefly introducing the problem and common strategies to solve it, we introduce our main interest in this paper, which is to apply jackknife+ for $\mathcal{R}_\varphi$ that is an ensemble regression algorithm with aggregation function φ. Specifically, ensemble predictions are obtained after applying a base regression method $\mathcal{R}$ to different training sets generated from the training data by a resampling procedure and then aggregate the predictions via φ.

Mathematically, we begin by creating multiple training data sets of size m,

$$S_1 = (i_{1,1}, \ldots, i_{1,m}), \ldots, S_B = (i_{B,1}, \ldots, i_{B,m}),$$

so that each S_b is a *multiset* of the original training data. For each b, we then compute a fitted function $\hat{\mu}_b$ on the b-th training data set S_b. These B fitted regression functions are finally aggregated using some aggregation function φ, which maps a collection of predictions $\hat{\mu}_1(x), \ldots, \hat{\mu}_B(x)$ to a single final prediction $\hat{\mu}_\varphi(x)$ for any feature vector $x \in \mathbb{R}^p$. This ensembled construction is formalized in Algorithm 1 below.

Algorithm 1 Ensembled learning

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$
Output: Ensembled regression function $\hat{\mu}_\varphi$
  for $b = 1, \ldots, B$ **do**
    Draw $S_b = (i_{b,1}, \ldots, i_{b,m})$ by sampling uniformly at random, with or without replacement, from $\{1, \ldots, n\}$.
    Compute $\hat{\mu}_b = \mathcal{R}((X_{i_{b,1}}, Y_{i_{b,1}}), \ldots, (X_{i_{b,m}}, Y_{i_{b,m}}))$.
  end for
  Define $\hat{\mu}_\varphi = \varphi(\hat{\mu}_1, \ldots, \hat{\mu}_B)$.

Here are some concrete examples:

- Two common ways to obtain the training data sets S_b are either *bootstrapping* or *subsampling*. The former subsamples m indices from $\{1, \ldots, n\}$ uniformly with replacement and the latter does so without replacement. It is typical to choose $m = n$ in bootstrapping and $m = 0.5n$ in subsampling.
- Common choices of the baseline regression algorithm $\mathcal{R}$ are linear or generalized linear regression, penalized linear regression such as Ridge or Lasso, a neural network, or a regression tree.
- The aggregation function φ is often chosen to be the median, mean, or trimmed mean, but other choices also exist, as we will see in Section 3 where we introduce the J+aB2 method.

For any base algorithm $\mathcal{R}$, when φ is chosen to be mean aggregation, the ensembled method run with bootstrapped S_b's is referred to as *bagging* [Breiman, 1996].

We remark that ensemble learning is a popular technique for enhancing the performance of machine learning algorithms. It is used to capture a complex model space with simple hypotheses which are often significantly easier to learn, or to increase the accuracy of an otherwise unstable procedure [see Hastie et al., 2009, Polikar, 2006, Rokach, 2010, and references therein].

1.4 Quantifying Uncertainty for Ensemble Learning

While ensembling is generally understood to provide a more robust and stable prediction as compared to the underlying base algorithm, there are substantial difficulties in developing inference procedures for ensemble methods with theoretical guarantees. For one thing, ensemble methods are frequently used with highly discontinuous and nonlinear base learners, and aggregating many of them leads to models that defy an easy analysis. For concreteness, consider regression trees such as CART that split nodes based on variance reduction. The problem is compounded by the fact that ensemble methods are typically employed in settings where good generative models of the data distribution are either unavailable or difficult to obtain. Thus, we would naturally want to use distribution-free predictive inference methods as introduced in Section 1.2 whose validity does not depend on knowing the distribution of the data or characterizing the behavior of the regression algorithm.

We might therefore consider applying a distribution-free method such as jackknife+ to the ensembled model. However, the computational cost of such a procedure is prohibitive, if it is applied naively. Under jackknife+, each ensembled model requires B many calls to the underlying base algorithm $\mathcal{R}$ to produce a single ensemble predictor $\hat{\mu}_{\varphi,\backslash i}$ that leaves out the ith training data, after which this procedure is repeated n many times to produce the full set of ensembled leave-one-out residuals $R_{\varphi,i}$—hence, Algorithm 1 is run once for each leave-one-out regression—for a total of Bn many calls to $\mathcal{R}$. Therefore, we should not naively apply the jackknife+ "wrapper" to an ensembled regression.

1.5 Related Work

Many ensemble methods can be cast as particular instances of bootstrap aggregating or bagging [Breiman, 1996]. Some of the earlier theoretical works were concerned with studying the impact of bagging on improving accuracy compared to the base method [Bühlmann and Yu, 2002, Buja and Stuetzle, 2006, Friedman and Hall, 2007]. These works are primarily focused on quantifying the improvement over the base (non-ensembled) algorithm.

The literature that deals with precise uncertainty quantification of ensembled estimators is substantially leaner. Meinshausen [2006], Athey et al. [2019], Lu and Hardin [2019] proposed methods for estimating conditional quantiles derived from the popular random forests [Ho, 1995, Breiman, 2001]. These methods can be used to construct valid prediction intervals, but their guarantees are necessarily approximate or asymptotic, and rely on additional conditions. By contrast, Sexton and Laake [2009], Wager et al. [2014], Mentch and Hooker [2016] studied methods for estimating the variance of the random forest estimator of the conditional mean by applying, in order, the jackknife-after-bootstrap (not jackknife+) [Efron, 1992] or the infinitesimal jackknife [Efron, 2014] or U-statistics theory. Roy and Larocque [2019] propose a heuristic for constructing prediction intervals with such variance estimates. For a comprehensive survey of statistical work related to random forests, we refer the reader to the literature review by Athey et al. [2019].

While our proposed methods are designed to be deployed in conjunction with bootstrap or ensemble methods, in flavor they are more closely linked to the growing literature on assumption-free predictive inference [see Vovk et al., 2005, Lei et al., 2018, and references there in]. Our paper is most closely related to the jackknife+ of Barber et al. [2019]. More recently,

Kuchibhotla and Ramdas [2019] looked at aggregating conformal inference after subsampling or bootstrapping. Their work proposes ensembling multiple runs of an inference procedure, while in contrast our present work seeks to provide inference for ensembled methods.

As mentioned in Section 1.4, applying jackknife+ (or, equivalently, jackknife) as a "wrapper" around an ensembled algorithm is computationally burdensome, with Bn many calls to the base learner (where B is the number of samples used for constructing an ensemble, and n is the sample size). To reduce this computational burden while ensuring distribution-free theoretical guarantees, we can instead consider using a holdout set to assess the predictive accuracy of an ensembled model, as studied by, e.g., Papadopoulos et al. [2002], Papadopoulos and Haralambous [2011]. However, when the sample size n is limited, we will achieve more accurate predictions with a cross-validation or jackknife type method, which avoids reducing the sample size in order to obtain a holdout set.

Finally, our work is most closely related to the idea of "out-of-bag" prediction intervals proposed for the jackknife, as a computationally efficient alternative to the naive idea of applying jackknife directly to an ensembled algorithm and thus requiring Bn calls to the base learner. Specifically, defining the function

$$\hat{\mu}_{\varphi\setminus i} = \varphi(\{\hat{\mu}_b : b = 1, \ldots, B, S_b \not\ni i\}),$$

which ensembles all models $\hat{\mu}_b$ whose subsample S_b does *not* train on the ith data point, Johansson et al. [2014] propose a prediction interval of the form

$$\hat{\mu}_{\varphi}(X_{n+1}) \pm q^{+}_{\alpha,n}(R_i) \quad \text{where} \quad R_i = |Y_i - \hat{\mu}_{\varphi\setminus i}(X_i)|. \tag{2}$$

Zhang et al. [2019] provide a theoretical analysis of this type of prediction interval, ensuring that predictive coverage holds asymptotically under additional assumptions. Devetyarov and Nouretdinov [2010], Löfström et al. [2013], Boström et al. [2017b,a], Linusson et al. [2019] study variants of this type of method, but distribution-free coverage is not guaranteed in any of these proposed methods.

The rest of the paper is organized as follows. Section 2 presents the method called the jackknife+-after-bootstrap (J+aB), which performs efficient predictive inference for ensemble learning. The J+aB algorithm, theoretical guarantees, and proofs are included. Section 3 considers and analyzes theoretically an alternative method to integrate J+ with ensembling, titled the jackknife+-with-bootstrap (J+wB). Section 4 empirically compares and contrasts the performance of J+aB and J+wB on real datasets and Section 5 discusses the empirical performance of these two methods in greater detail.

2 Jackknife+-after-bootstrap

In this section, we address the problem of inference for ensemble predictions, by proposing the jackknife+-after-bootstrap (J+aB) method[1], which only makes the necessary B many calls rather than Bn calls to the base regression algorithm, and is therefore highly efficient in

[1]The work in this section is a collaborative research with Professor Rina Barber and colleague Byol Kim. This preprint is titled Predictive Inference Is Free with the Jackknife+-after-Bootstrap [Kim, Xu, and Barber, 2020]

terms of the cost of model fitting. Specifically, J+aB obtains n many ensembled leave-one-out regressors with the *same* number of calls to the baseline regression method.

When run at a target predictive coverage level of $1-\alpha$, the method also provably provides at least $1-2\alpha$ coverage in the worst case. In terms of assumption, J+aB requires only independent and identically distributed data and symmetric base algorithm and aggregation function.

2.1 The Method and Algorithm

To obtain the i-th leave-one-out fitted predictor $\hat{\mu}_{\varphi\setminus i}$, we will simply aggregate the original models $\hat{\mu}_1, \ldots, \hat{\mu}_B$ with the caveat that we exclude any $\hat{\mu}_b$ whose training data set S_b includes data point i.

Algorithm 2 Jackknife+-after-bootstrap (J+aB)

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$
Output: Predictive interval $\hat{C}^{\mathrm{J+aB}}_{\alpha,n,B}$
  for $b = 1, \ldots, B$ **do**
    Draw $S_b = (i_{b,1}, \ldots, i_{b,m})$ by sampling uniformly at random (with or without replacement, as desired) from $\{1, \ldots, n\}$.
    Compute $\hat{\mu}_b = \mathcal{R}((X_{i_{b,1}}, Y_{i_{b,1}}), \ldots, (X_{i_{b,m}}, Y_{i_{b,m}}))$.
  end for
  for $i = 1, \ldots, n$ **do**
    Aggregate $\hat{\mu}_{\varphi\setminus i} = \varphi(\{\hat{\mu}_b : b = 1, \ldots, B,\ S_b \not\ni i\})$.
    Compute the residual, $R_i = |Y_i - \hat{\mu}_{\varphi\setminus i}(X_i)|$.
  end for
  Compute the jackknife+-after-bootstrap prediction interval: at each $x \in \mathbb{R}$,

$$\hat{C}^{\mathrm{J+aB}}_{\alpha,n,B}(x) = \left[q^-_{\alpha,n}\{\hat{\mu}_{\varphi\setminus i}(x) - R_i\}, q^+_{\alpha,n}\{\hat{\mu}_{\varphi\setminus i}(x) + R_i\}\right].$$

To run the jackknife+-after-bootstrap (J+aB) method given in Algorithm 2, we can see that all n leave-one-out models $\hat{\mu}_{\varphi\setminus i}$ are computed by aggregating subsets of the *same* underlying list of fitted models $\hat{\mu}_1, \ldots, \hat{\mu}_B$. Therefore, the cost of J+aB is essentially the same as that of ensembled learning (Algorithm 1) in any setting where the dominant computational cost comes from model training rather than model aggregation or function evaluation. For example, this will hold if φ is the mean or median while $\mathcal{R}$ is an expensive method such as a neural net.

Thus, for a prediction task, instead of a point estimate obtained by ensemble learning (Algorithm 1), we can provide a more informative prediction interval via jackknife+-after-bootstrap (Algorithm 2), essentially "for free."

2.2 Theory

To give guarantee for J+aB, we make two assumptions, one on the data distribution and the other on the base and aggregation algorithms.

Assumption 1 (i.i.d. data)**.** The training and test data are i.i.d.: $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \overset{\text{iid}}{\sim} \mathcal{P}$, where $\mathcal{P}$ is any distribution on $\mathbb{R}^p \times \mathbb{R}$.

Assumption 2 (symmetric algorithms)**.** For $k \geq 1$, any fixed k-tuple $((x_1, y_1), \ldots, (x_k, y_k)) \in \mathbb{R}^p \times \mathbb{R}$, and any permutation σ on $\{1, \ldots, k\}$, it holds that $\mathcal{R}((x_1, y_1), \ldots, (x_k, y_k)) = \mathcal{R}((x_{\sigma(1)}, y_{\sigma(1)}), \ldots, (x_{\sigma(k)}, y_{\sigma(k)}))$ and $\varphi(y_1, \ldots, y_k) = \varphi(y_{\sigma(1)}, \ldots, y_{\sigma(k)})$.

In other words, the base regression algorithm $\mathcal{R}$ and the aggregation φ are both invariant to the ordering of the input arguments.[2]

Assumption 1 is fairly standard in the distribution-free prediction literature [Vovk et al., 2005, Lei et al., 2018, Barber et al., 2019] (in fact, as in the conformal prediction literature, our results only require exchangeability of the $n+1$ data points—the i.i.d. assumption is a familiar special case). Assumption 2 is a natural condition in the setting where the data points are i.i.d. and therefore should logically be treated symmetrically.

Under these assumptions, we establish a coverage guarantee for the jackknife+-after-bootstrap prediction interval, with no assumptions on the data, the base algorithm, or the aggregation procedure. One interesting requirement is that the number of subsamples, B, is required to be *random* for this result to hold—we discuss this point later on.

Theorem 1. *Fix any integers $\tilde{B} \geq 1$ and $m \geq 1$, any base algorithm $\mathcal{R}$, and any aggregation function φ. Suppose jackknife+-after-bootstrap (Algorithm 2) is run with $B \sim \text{Binomial}(\tilde{B}, (1 - \frac{1}{n+1})^m)$ (in the case of sampling with replacement) or $B \sim \text{Binomial}(\tilde{B}, 1 - \frac{m}{n+1})$ (in the case of sampling without replacement). Then, under Assumptions 1 and 2, the jackknife+-after-bootstrap prediction interval satisfies*

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}^{\text{J+aB}}_{\alpha,n,B}(X_{n+1})\right] \geq 1 - 2\alpha,$$

where the probability holds with respect to the random draw of the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, the test data point (X_{n+1}, Y_{n+1}), and B.

2.3 Proof Sketch

2.3.1 Why Do We Need a Random B?

To see why B needs to be random in order to establish Theorem 1, it is instructive to go over the jackknife+ theory and understand how exchangeability is used in the proof to obtain a lower-bound on the coverage. The proof of the jackknife+ coverage guarantee [Barber et al., 2019, Theorem 1] is based on the observation that the event that the predictive interval fails to cover Y_{n+1}, implies the event that for at least $\lceil (1-\alpha)(n+1) \rceil$ leave-one-out fitted predictors $\hat{\mu}_{\setminus i}$, the residual for the new observation exceeds the residual for the i-th observation in

[2] If $\mathcal{R}$ and/or φ involve any randomization—for example if φ operates by sampling from the collection of predictions—then we can require that the outputs are equal in distribution under any permutation of the input arguments, rather than requiring that equality holds deterministically. In this case, the coverage guarantees in our theorems hold on average over the randomization in $\mathcal{R}$ and/or φ, in addition to the distribution of the data.

magnitude. In other words, if $Y_{n+1} \notin \hat{C}^{\mathrm{J+}}_{\alpha,n}(X_{n+1})$ then

$$\sum_{i=1}^{n} \mathbb{1}\left[\left|Y_{n+1} - \hat{\mu}_{\backslash i}(X_{n+1})\right| > \left|Y_i - \hat{\mu}_{\backslash i}(X_i)\right|\right] \geq (1-\alpha)(n+1).$$

Initially, it may seem that we cannot use exchangeability of the training and test data to study this event, since each training data point i appears in every $\hat{\mu}_{\backslash j}$ where $j \neq i$, while the test point $n+1$ is not used in any fitted model.

However, we can embed these n leave-one-out models $(\hat{\mu}_{\backslash i})_{i=1}^n$ into a larger collection in order to restore exchangeability. Consider the $(n+1) \times (n+1)$ array of leave-*two*-out fitted predictors $(\tilde{\mu}_{\backslash i,j})_{1 \leq i \neq j \leq n+1}$. Since the $n+1$ data points are assumed to be i.i.d., and this array constructs leave-two-out fitted models for each possible pair, the resulting $(n+1) \times (n+1)$ array is exchangeable, i.e., its distribution does not change if we permute the rows/columns. As $\tilde{\mu}_{\backslash n+1,i} = \hat{\mu}_{\backslash i}$ for $i = 1, \ldots, n$, any statement we make with $\tilde{\mu}_{\backslash n+1,i}$'s, which are embedded in the $(n+1) \times (n+1)$ exchangeable array, is related back to $\hat{\mu}_{\backslash i}$'s, and thus, to the jackknife+ interval. This construction underlies the theory for the jackknife+.

If we attempt to apply the jackknife+ proof for the J+aB algorithm, however, issues arise immediately. For example, define $\tilde{\mu}_{\varphi\backslash i,j} = \varphi(\{\hat{\mu}_b : S_b \not\ni i, j\})$, the aggregation of all fitted models $\hat{\mu}_b$ whose underlying subsampled or bootstrapped data set S_b does not include either i or j. For each $i = 1, \ldots, n$, we have $\tilde{\mu}_{\varphi\backslash n+1,i} = \hat{\mu}_{\varphi\backslash i}$, exactly as for the jackknife+ proof, and so it would seem that we can prove coverage of the jackknife+-after-bootstrap method by way of this larger $(n+1) \times (n+1)$ array of $\tilde{\mu}_{\varphi\backslash i,j}$'s.

Unfortunately, though, this larger array is *not* exchangeable—the jackknife+-after-bootstrap algorithm subtly violates symmetry even though $\mathcal{R}$ and φ are themselves symmetric. This is mostly easily seen by noting that there are always exactly B many subsampled or bootstrapped training data sets S_b that do not include the test observation $n+1$, whereas for any training observation $i = 1, \ldots, n$ the number of S_b's that do not contain i is usually smaller. It turns out that this issue can easily be addressed by simply drawing B from a Binomial distribution, as we will see next.

2.3.2 Proof of Theorem 1

We now prove the distribution-free guarantee of Theorem 1. Our proof follows the main idea of the jackknife+ guarantee [Barber et al., 2019, Theorem 1]—we lift the jackknife+-after-bootstrap method, which requires the construction of n many leave-one-out ensembled models $\hat{\mu}_{\varphi\backslash i}$, to an $(n+1) \times (n+1)$ array of leave-two-out models. Unlike the jackknife+ theory, here we must take care to ensure exchangeability within the collection of subsamples S_b. Here we sketch the argument; for completeness, full details are given in Appendix A.

Consider the "lifted" Algorithm 3. We can see that Algorithm 3 treats the $n+1$ data points symmetrically, and therefore the resulting array of residuals $(R_{ij} : i \neq j \in \{1, \ldots, n+1\})$ is exchangeable. Now, for each $i = 1, \ldots, n+1$, define $\tilde{\mathcal{E}}_i$ as the event that

$$\sum_{j \in \{1,\ldots,n+1\} \backslash \{i\}} \mathbb{1}\left[R_{ij} > R_{ji}\right] \geq (1-\alpha)(n+1).$$

By a simple counting argument and exchangeability, it can be shown that $\mathbb{P}[\tilde{\mathcal{E}}_{n+1}] \leq 2\alpha$, but we need to relate the event $\tilde{\mathcal{E}}_{n+1}$, defined based on the lifted jackknife+-after-bootstrap

Algorithm 3 Lifted jackknife+-after-bootstrap residuals

Input: Data $\{(X_i, Y_i)\}_{i=1}^{n+1}$
Output: Residuals $(R_{ij} : i \neq j \in \{1, \ldots, n+1\})$
  for $b = 1, \ldots, \tilde{B}$ **do**
    Draw $\tilde{S}_b = (i_{b,1}, \ldots, i_{b,m})$ uniformly at random, with or without replacement, from $\{1, \ldots, n+1\}$.
    Compute $\tilde{\mu}_b = \mathcal{R}((X_{i_{b,1}}, Y_{i_{b,1}}), \ldots, (X_{i_{b,m}}, Y_{i_{b,m}}))$.
  end for
  for pairs $i \neq j \in \{1, \ldots, n+1\}$ **do**
    Aggregate $\tilde{\mu}_{\varphi \setminus i,j} = \varphi(\{\tilde{\mu}_b : \tilde{S}_b \not\ni i, j\})$.
    Compute the residual, $R_{ij} = |Y_i - \tilde{\mu}_{\varphi \setminus i,j}(X_i)|$.
  end for

construction, back to the original jackknife+-after-bootstrap interval $\hat{C}^{\text{J+aB}}_{\alpha,n,B}(X_{n+1})$. Let $B = \sum_{b=1}^{\tilde{B}} \mathbb{1}[\tilde{S}_b \not\ni n+1]$, the number of $\tilde{S}_b$'s containing only training data, in the lifted construction, and let $1 \leq b_1 < \cdots < b_B \leq \tilde{B}$ be the corresponding indices. Note that the distribution of B is Binomial, as specified in the theorem. Now, for each $k = 1, \ldots, B$, define $S_k = \tilde{S}_{b_k}$. We can observe that each S_k is an independent uniform draw from $\{1, \ldots, n\}$ (with or without replacement). Therefore, we can equivalently consider running J+aB (Algorithm 2) with these particular subsamples or bootstrapped samples $S_1, \ldots, S_B$. Furthermore, for each $i = 1, \ldots, n$, this ensures that $\tilde{\mu}_{\varphi \setminus n+1,i} = \hat{\mu}_{\varphi \setminus i}$, that is, the leave-one-out models of the jackknife+-after-bootstrap methods coincide with the leave-two-out models of the lifted jackknife+-after-bootstrap. Thus, we have constructed a coupling of the jackknife+-after-bootstrap with its lifted version.

Now, define $\mathcal{E}_{n+1}$ as the event that

$$\sum_{i=1}^{n} \mathbb{1}\left[|Y_{n+1} - \hat{\mu}_{\varphi \setminus i}(X_{n+1})| > R_i\right] \geq (1-\alpha)(n+1),$$

where $R_i = |Y_i - \hat{\mu}_{\varphi \setminus i}(X_i)|$ as before. By the coupling we have just constructed, we can see that the event $\mathcal{E}_{n+1}$ is exactly equivalent to the lifted event $\tilde{\mathcal{E}}_{n+1}$, and thus, $\mathbb{P}[\mathcal{E}_{n+1}] = \mathbb{P}[\tilde{\mathcal{E}}_{n+1}] \leq 2\alpha$. It can be verified that if jackknife+-after-bootstrap fails to cover, i.e., if $Y_{n+1} \notin \hat{C}^{\text{J+aB}}_{\alpha,n,B}(X_{n+1})$, then the event $\mathcal{E}_{n+1}$ must occur, completing the proof.

3 Jackknife+-with-bootstrap

3.1 The Method and Algorithm

Note that in J+aB, we aggregate the individual predictors $\hat{\mu}^b$ in a leave-one-out fashion first *before* predicting on the new data X_{n+1} and X_i's to get point predictions and leave-one-out residuals. However, we can consider performing the aggregation and prediction in another order such that we first create the leave-one-residuals and point predictions on new data with the individual predictors and *then* aggregate the residuals not via taking mean or median but simply taking the quantiles. This motiviates the jackknife+-with-bootstrap variant that,

as we iterate through the n training data points, also produces predictive inference intervals by leaving out the i-th training data point.

In particular, the same B many bootstrapped predictors introduced in the ensemble learning algorithm (Algorithm 1) are used in Algorithm 4 below.

Algorithm 4 Jackknife+-with-bootstrap (J+wB)

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$
Output: Predictive interval $\hat{C}^{\text{J+wB}}_{\alpha,n,B}$ at x
  for $b = 1, \ldots, B$ **do**
    Draw $S_b = (i_{b,1}, \ldots, i_{b,m})$ by sampling uniformly at random (with or without replacement, as desired) from $\{1, \ldots, n\}$.
    Compute $\hat{\mu}_b = \mathcal{R}((X_{i_{b,1}}, Y_{i_{b,1}}), \ldots, (X_{i_{b,m}}, Y_{i_{b,m}}))$.
  end for
  $S^- = \{\}$,$S^+ = \{\}$
  for $b = 1, \ldots, B$ **do**
    for $i = 1, \ldots, n$ **do**
      if $i \notin S_b$ **then**
        $R_i^b = |Y_i - \hat{\mu}_b(X_i)|$
        $S^- = S^- \cup \{\hat{\mu}_b(x) - R_i^b\}$
        $S^+ = S^+ \cup \{\hat{\mu}_b(x) + R_i^b\}$
      else
        $S^- = S^-$, $S^+ = S^+$
      end if
    end for
  end for
  Define $s := |S^-| = |S^+|$
  Compute the jackknife+-after-bootstrap2 prediction interval: at each $x \in \mathbb{R}$,

$$\hat{C}^{\text{J+wB}}_{\alpha,n,B}(x) = \left[q^-_{\alpha,s}\{S^-\}, q^+_{\alpha,s}\{S^+\}\right].$$

By comparing Algorithm 2 and 4, we emphasize that when the dominant computational cost comes from model training rather than model aggregation or sorting a long list, these two methods have comparable computational costs. However, the theoretical guarantee relies on different assumptions—in particualr, it is non-asymptotic for Algorithm 4 but asymptitic for Algorithm 2, as we will see in the next section.

3.2 Theory

The same problem of violating exchangeability at a fixed B still persists in Algorithm 4 but is resolved when B is drawn from the same particular binomial distribution as in Theorem 1. In particular, we have the following theorem similar to Theorem 1:

Theorem 2. *Fix any base algorithm $\mathcal{R}$. Suppose jackknife+-with-bootstrap (Algorithm 4) is run with $B \sim \text{Binomial}(\tilde{B}, (1 - \frac{1}{n+1})^m)$ (in the case of sampling with replacement) or*

$B \sim \text{Binomial}(\tilde{B}, 1 - \frac{m}{n+1})$ *(in the case of sampling without replacement). Then, under Assumptions 1 and 2, the jackknife+-with-bootstrap prediction interval satisfies*

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}^{\text{J+wB}}_{\alpha,n,B}(X_{n+1})\right] \geq 1 - 2\alpha - \delta,$$

$$\delta \leq 2\alpha\sqrt{\frac{log(n(n+1)\tilde{B})}{\tilde{B}C}} + \frac{1}{\tilde{B}}, \text{ where}$$

$$C(n, m, \alpha) := \frac{(1 - \frac{2}{n+1})^m}{3(2 + \frac{1}{2\alpha})^2}$$

where the probability holds with respect to the random draw of the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$*, the test data point* (X_{n+1}, Y_{n+1})*, and* B*.*

We remark that at any finite sample size n, subsample size m, and tolerance error level α, the inflation term δ can be made arbitrary small as long as $\tilde{B} \to \infty$.

Both of the theorems above show that the distribution-free coverage guarantee of the jackknife+ extends immediately to the J+aB and J+wB with one intriguing twist: the number of bootstrapped or subsampled training sets, B, must be drawn at random rather than chosen in advance. In practice, this choice of B does not make any meaningful difference to the output of the algorithm, as long as $\tilde{B}$ is large. However, the theoretical arguments for these variants that allow us to obtain distribution-free coverage require this random B in an interesting way.

3.3 Proof Sketch

Here we sketch the argument that resembles the earlier proof of jackknife+-after-bootstrap as in Section 2.3.2; for completeness, full details are given in Appendix A. In particular, for each pair of indices $(i, j) \in [n+1] \times [n+1]$ such that $i \neq j$, the lifted jackknife+-with-bootstrap algorithm below computes the residuals $\tilde{R}^b_{ij}$ and $\tilde{R}^b_{ji}$ for $b \in \{1, ..., \tilde{B}\}$ via leave-two-out constructions:

Algorithm 5 Lifted jackknife+-with-bootstrap residuals

Input: Data $\{(X_i, Y_i)\}_{i=1}^{n+1}$
Output: Residuals $(\tilde{R}_{ij}^b : i \neq j \in \{1, \ldots, n+1\},\, b \in \{1, \ldots, \tilde{B}\})$

```
for b = 1, ..., B̃ do
    Draw S̃_b = (i_{b,1}, ..., i_{b,m}) uniformly at random (with or without replacement) from {1, ..., n+1}.
    Compute μ̃_b = R((X_{i_{b,1}}, Y_{i_{b,1}}), ..., (X_{i_{b,m}}, Y_{i_{b,m}})).
end for
for b = 1, ..., B̃ do
    for pairs (i, j) ∈ [n+1] × [n+1], i ≠ j do
        if (i, j) ∉ S̃_b then
            Compute the residuals R̃^b_{ij} = |Y_i − μ̃_b(X_i)|  R̃^b_{ji} = |Y_j − μ̃_b(X_j)|
        else
            Define the residuals R̃^b_{ij} = R̃^b_{ji} = ∞
        end if
    end for
end for
```

Given the lifted Algorithm 5, we can see that for each b, the resulting array of residuals $(\tilde{R}_{ij}^b, i \neq j \in \{1, ..., n+1\})$ is exchangeable. Therefore, by following the notation in section 2.3.2, for each $i = 1, ..., n+1$, we define $\tilde{\mathcal{E}}_i$ as the event that

$$\sum_{j \in \{1,\ldots,n+1\} \setminus \{i\}} \sum_{b:(i,j) \notin \tilde{S}_b} \mathbb{1}[\tilde{R}_{i,j}^b > \tilde{R}_{j,i}^b] \geq (1-\alpha)\Big(\sum_{j \in \{1,\ldots,n+1\} \setminus \{i\}} \sum_{b:(i,j) \notin \tilde{S}_b} \mathbb{1}\Big).$$

Via mimicking the argument in section 2.3.2, we also observe a natural coupling of the jackknife+-with-bootstrap with its lifted version, since $B := \sum_{b=1}^{\tilde{B}} \mathbb{1}[n+1 \notin \tilde{S}_b]$ is understood to follow a binomial distribution (under subsampling with or without replacement). Equivalently, if we execute Algorithm 4 with the binomially distributed B and define $\mathcal{E}_{n+1}$ as the event that

$$\sum_{j=1}^{n} \sum_{b:j \notin S_b} \mathbb{1}[|Y_{n+1} - \hat{\mu}^b(X_{n+1})| > R_j^b] \geq (1-\alpha)\Big(\sum_{j=1}^{n} \sum_{b:j \notin S_b} \mathbb{1}\Big)$$

where $R_j^b = |Y_j - \hat{\mu}^b(X_j)|$, we can see that the event $\mathcal{E}_{n+1}$ is exactly equivalent to the lifted event $\tilde{\mathcal{E}}_{n+1}$ under a fixed $\tilde{B}$ and therefore $\mathbb{P}[\mathcal{E}_{n+1}] = \mathbb{P}[\tilde{\mathcal{E}}_{n+1}]$. We complete the proof by verifying that if jackknife+-with-bootstrap fails to cover, i.e., if $Y_{n+1} \notin \hat{C}_{\alpha,n,B}^{\text{J+aB2}}(X_{n+1})$, then the event $\mathcal{E}_{n+1}$ must occur.

However, different from the argument in section 2.3.2 is the bound on $\mathbb{P}[\tilde{\mathcal{E}}_{n+1}]$, which is now less than or equal to $2\alpha(1+\epsilon) + n(n+1)\exp(-\tilde{B}p\frac{4\epsilon^2}{12(2+\epsilon)^2})$, $p := e^{-\frac{2}{k}}$, at any finite sample size n and user-specified ϵ. The details are provided in the appendix. Thus, $\mathbb{P}[Y_{n+1} \notin \hat{C}_{\alpha,n,B}^{\text{J+aB2}}(X_{n+1})] \leq 2\alpha(1+\epsilon) + n(n+1)\exp(-\tilde{B}p\frac{4\epsilon^2}{12(2+\epsilon)^2})$.

4 Experiments

Our experimental results aim to verify the coverage properties of jackknife+-after-bootstrap and jackknife+-with-bootstrap using different base algorithms, and to compare and contrast their performances.

4.1 Data

We use the same three real data sets as in [Barber et al., 2019], performing the same preprocessing steps on the data.

The Communities and Crime (COMMUNITIES) data set [Redmond and Baveja, 2002] contains information on 1994 communities with $d = 99$ covariates. The response Y is the per capita violent crime rate.

The BlogFeedback (BLOG) data set [Buza, 2014] contains 52397 blog posts with $d = 280$ covariates. The response is the number of comments left on the blog post in the following 24 hours, which we transform as $Y = \log(1 + \#\text{comments})$.

The Medical Expenditure Panel Survey 2016 (MEPS) data set (from the Agency for Healthcare Research and Quality) is described in [Ezzati-Rice et al., 2008]. The response is a composite score measuring use of medical services. There are 33005 data points with $d = 107$ covariates after the preprocessing steps following [Barber et al., 2019]. Since the distribution of the response is highly skewed, we use the transformation $Y = \log(1 + \text{utilization score})$.

4.2 Setup and Procedures

In all experiments, we fixed $\alpha = 0.1$ for a target coverage level of 90%. In each of 10 trials, a fixed $n = 200$ number of points were randomly sampled without replacement from the whole data set, and were used to train both the jackknife+-after-bootstrap and the jackknife+-with-bootstrap. We ran the jackknife+-after-bootstrap (Algorithm 2) with mean aggregation and jackknife+-with-bootstrap (Algorithm 4)using sampling *with* replacement. We varied m, the size of each bootstrap replicate, from $m = 0.1n$ to $m = n$ with an increment of $0.1n$. Fixing $\tilde{B} = 200$, a total of $B \sim \text{Binomial}(\tilde{B}, (1 - \frac{1}{n+1})^m)$ bootstrap replicates were drawn at each run.

For comparison purposes, we chose the same three base regression algorithms as used in [Barber et al., 2019] — namely, ridge regression (RIDGE), random forests (RF),[3] and neural networks (NN), with the same settings as used in [Barber et al., 2019]. We do not optimize these algorithms, as we are only interested in how the jackknife+-with-bootstrap performs with these models, and how it compares to the jackknife+-after-bootstrap.

4.3 Results

Table 1 below displays the resulting average coverage and average interval width for each data set and each base algorithm, where the size of each bootstrap replicate is $m = n$ for both methods.

[3]The RF base algorithm we use subsamples the features but not the observations.

Figures 1 and 2 below show average coverage and average width, respectively, for all pairs of base regression method (RIDGE or RF or NN) and data set (COMMUNITIES or MEPS or BLOG).

We remark that although the assumption-free lower-bound on the coverage is only $1-2\alpha$ for both the jackknife+-after-bootstrap and the jackknife+-with-bootstrap, Table 1 and Figure 1 and 2 make clear that both methods yield intervals with above or close to $1-\alpha$ coverage on average. The behavior is consistent for all data sets and base regression method combinations we consider, as well as across all m. In fact, the jackknife+-with-bootstrap significantly overcovers above $1-\alpha$ in some cases (i.e. when baselines are RF and NN), especially at low values of m. This reflects both the possible instability of individual bootstrap estimators trained with small numbers of training data and the ability of bagging to significantly improve the stability of the predictive inference method.

Also noteworthy is that, unless RIDGE is used as the baseline algorithm and m is close to n, the average interval widths under jackknife+-with-bootstrap are significantly wider. This phenomenon again shows the noisiness and instability of individual bootstrap estimators in comparison with the bagging estimator in the context of producing efficient predictive intervals. Moreover, the huge differences in interval widths between the two methods under RF and NN show that their performances can be greatly improved by bagging. This point is worth considering when choosing a baseline method. Overall, although the two predictive method have similar computational costs in terms of number of calls to the baseline algorithm, jackknife+-after-bootstrap as introduced in Algorithm 2 almost always produce more efficient and valid predictive inference intervals so it should be a preferable method to use in reality.

Table 1: The performances of jackknife+-after-bootstrap and jackknife+-with-bootstrap ($m = n$ and sampling with replacement) on all data sets for different base regression methods. (Results are averages over 10 independent training / test splits).

Data set/ Base algorithm	Method	Coverage (SE)	Width (SE)
Communities			
Ridge	J+wB	0.919 (0.005)	0.507 (0.012)
	J+aB	0.904 (0.005)	0.478 (0.010)
RF	J+wB	0.983 (0.001)	0.782 (0.015)
	J+aB	0.904 (0.007)	0.500 (0.015)
NN	J+wB	0.970 (0.003)	0.781 (0.021)
	J+aB	0.910 (0.012)	0.575 (0.025)
MEPS			
Ridge	J+wB	0.920 (0.008)	4.418 (0.055)
	J+aB	0.890 (0.010)	4.195 (0.062)
RF	J+wB	0.992 (0.001)	6.383 (0.091)
	J+aB	0.912 (0.005)	4.108 (0.058)
NN	J+wB	0.941 (0.007)	4.798 (0.086)
	J+aB	0.910 (0.005)	4.473 (0.070)
Blog			
Ridge	J+wB	0.904 (0.007)	3.118 (0.129)
	J+aB	0.895 (0.008)	2.991 (0.148)
RF	J+wB	0.963 (0.004)	3.522 (0.105)
	J+aB	0.902 (0.004)	2.620 (0.063)
NN	J+wB	0.933 (0.005)	3.470 (0.116)
	J+aB	0.896 (0.007)	3.046 (0.102)

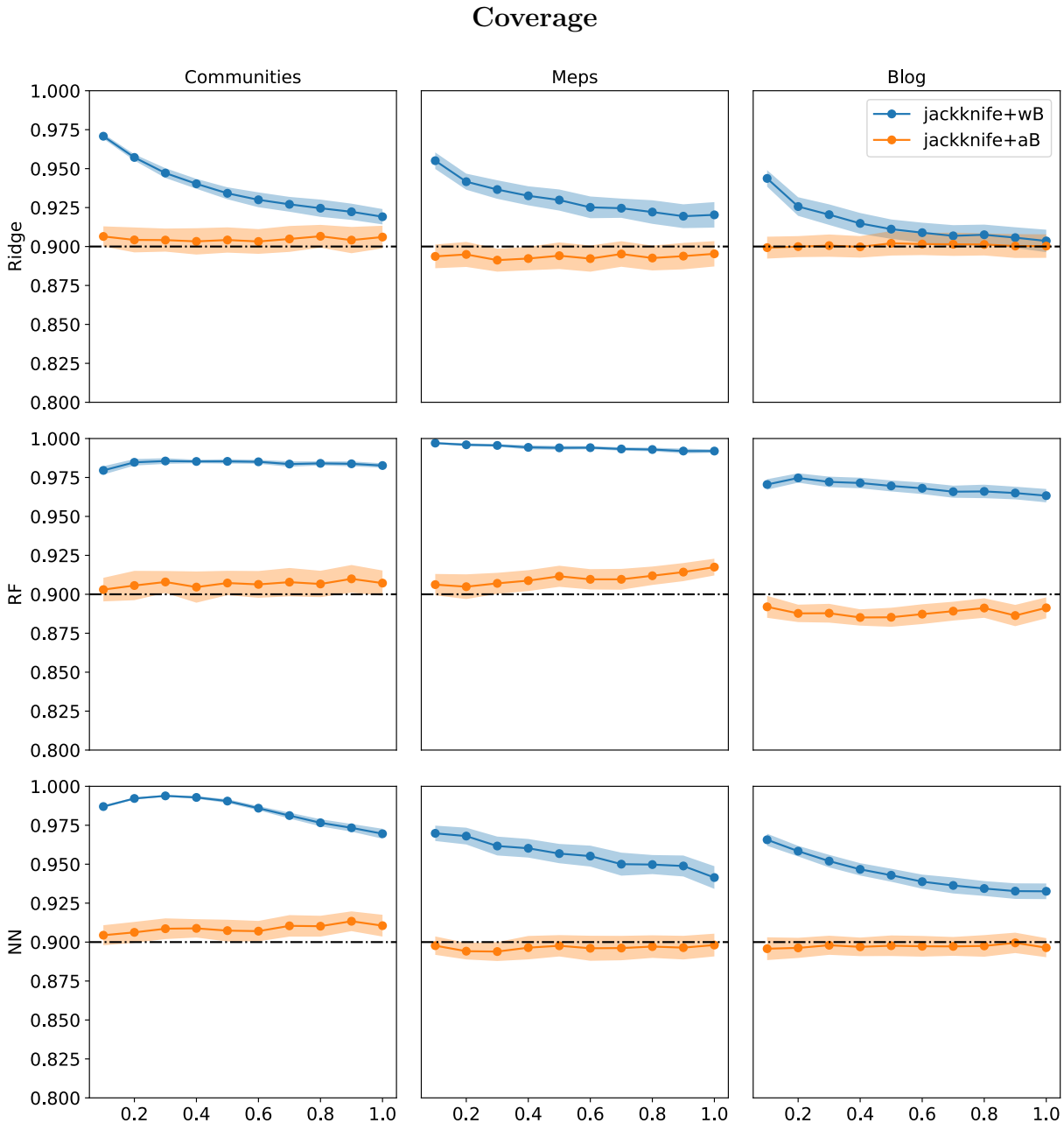

Figure 1: Average coverage results on all three data sets with three different base regression methods as m/n changes. The lines show the average, and the shaded areas show +/- one standard error, over 10 trials. The black dash-dotted line is the $1 - \alpha$ target coverage.

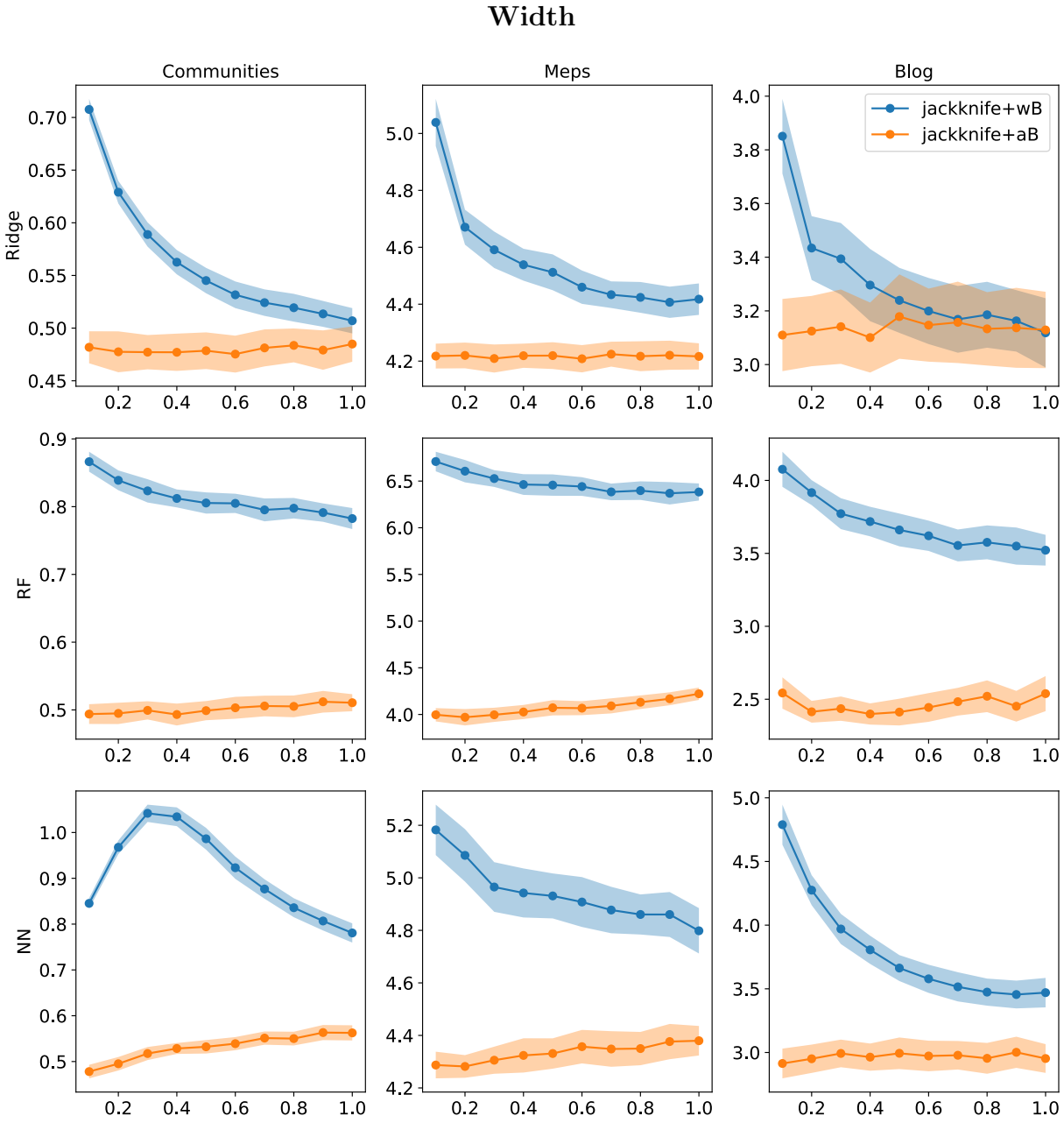

Figure 2: Average width results on all three data sets with three different base regression methods as m/n changes. The lines show the average, and the shaded areas show +/- one standard error, over 10 trials.

5 Discussion

In this section, we briefly discuss around the previous experimental results, that J+aB almost always produces more efficient predictive inference intervals under all baselines than J+wB, while the latter method also significantly over covers so that despite its intervals being valid, they are not informative.

Although our main focus in this paper has been quantifying uncertainties in prediction via producing valid predictive intervals, we can better understand why these two method behave so differently by first focusing on the greater predictive accuracy of bagged predictors over individuals ones. In particular, [Breiman, 1996] has emphasized in his well-known paper that theoretically and experimentally, bagging can improve predictive accuracy especially when the baseline algorithm is unstable. In our case, when the baseline algorithm is either neural network or random forest whose trees only subsample features, the collection of baseline predictors $\hat{\mu}^b$ can be unstable, especially when each of them is trained with a small subset of the available training data. In comparison, $\hat{\mu}_\varphi$ as the bagged predictor over these baseline predictors is much less sensitive to the perturbation of training set. We point out that such heuristic analysis has a more formal formulation under algorithmic stability in learning theory. For more details, one can refer to the work by [Bousquet and Elisseeff, 2002] and in this context of analyzing predictive inference for jackknife+, by [Barber et al., 2019]

The above observation can have two potential effects on the performance of J+aB and J+wB:

First, the bagged predictor can have smaller generalization error under $l2$ norm. That is, applied to X_{n+1}, $\hat{\mu}_\varphi(X_{n+1})$ can be closer to the underlying true value Y_{n+1} than $\hat{\mu}^b(X_{n+1})$. In our case, this means the center for the predictive interval for a new training data is more accurate under the bagged predictor.

Second, the leave-one-out residuals are less prone to having extreme values under the bagged predictor. In other words, under J+aB, the residuals $R_i := |Y_i - \hat{\mu}_{\varphi\setminus i}(X_i)|, i = 1, \ldots, n$ are less likely to be unusually large. This situation happens because for any particular i, $\hat{\mu}_{\varphi\setminus i}(X_i)$ is also closer to Y_i than the prediction $\hat{\mu}^b(X_i)$ by any baseline predictor $\hat{\mu}^b$ where $i \notin S_b$. As a consequence, the widths of predictive intervals under the bagged predictor are also smaller, without sacrificing valid coverage.

Combining the first and the second part, we can intuitively understand how the greater predictive accuracy of the bagged predictor over baseline predictors implies the empirical observation that J+aB produces more efficient intervals than J+wB does, without sacrificing validity when the latter can more easily become too conservative and over covers.

6 Conclusion

We introduce and compare the jackknife+-after-bootstrap and the jackknife+-with-bootstrap, both of which are computationally efficient method for constructing predictive intervals with assumption-free coverage guarantee. While the former method requires aggregation, empirical results have shown its ability to produce more efficient and valid methods than the latter method without aggregation. Both methods provide mechanisms for quantifying uncertainty in predictions which are both straightforward to implement and easy to interpret

and can therefore be easily integrated into existing regression models. In the future, it will be interesting to investigate various stability assumptions with which the theoretical coverage guarantee can be closer to the $1-\alpha$ target level.

A Additional Proofs

A.1 Proof of Theorem 1

For completeness, we give the full details of the proof of Theorem 1; a sketch of the proof is presented in Section 2.3.2.

Denote Algorithm 3 by $\tilde{\mathcal{A}}$. We view $\tilde{\mathcal{A}}$ as mapping a given input $\{(X_i, Y_i)\}_{i=1}^{n+1}$ and a collection of subsamples or bootstrapped samples $\tilde{S}_1, \ldots, \tilde{S}_B$ to a matrix of residuals $R \in \mathbb{R}^{(n+1)\times(n+1)}$, where

$$R_{ij} = \begin{cases} \left|Y_i - \tilde{\mu}_{\varphi\setminus i,j}(X_i)\right| & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases}$$

For any permutation σ on $\{1, \ldots, n+1\}$, let Π_σ stand for its matrix representation—that is, $\Pi_\sigma \in \{0,1\}^{(n+1)\times(n+1)}$ has entries $(\Pi_\sigma)_{\sigma(i),i} = 1$ for each i, and zeros elsewhere. Furthermore, for each subsample or bootstrapped sample $\tilde{S}_b = \{i_{b,1}, \ldots, i_{b,m}\}$, write $\sigma(\tilde{S}_b) = \{\sigma(i_{b,1}), \ldots, \sigma(i_{b,m})\}$.

We now claim that

$$R \stackrel{d}{=} \Pi_\sigma R \Pi_\sigma^\top, \tag{3}$$

for any fixed permutation σ on $\{1, \ldots, n+1\}$. Here R is the residual matrix obtained by a run of Algorithm 3, namely,

$$R = \tilde{\mathcal{A}}\left((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}); \tilde{S}_1, \ldots, \tilde{S}_B\right).$$

To see why (3) holds, observe that deterministically, we have

$$\Pi_\sigma R \Pi_\sigma^\top = \tilde{\mathcal{A}}\left((X_{\sigma(1)}, Y_{\sigma(1)}), \ldots, (X_{\sigma(n+1)}, Y_{\sigma(n+1)}); \sigma(\tilde{S}_1), \ldots, \sigma(\tilde{S}_B)\right).$$

Furthermore, we have

$$\left((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})\right) \stackrel{d}{=} \left((X_{\sigma(1)}, Y_{\sigma(1)}), \ldots, (X_{\sigma(n+1)}, Y_{\sigma(n+1)})\right)$$

by Assumption 1, and

$$\left(\tilde{S}_1, \ldots, \tilde{S}_B\right) \stackrel{d}{=} \left(\sigma(\tilde{S}_1), \ldots, \sigma(\tilde{S}_B)\right)$$

since subsampling or resampling treats all the indices the same. Finally, the subsamples or bootstrapped samples (i.e., the $\tilde{S}_b$'s) are drawn independently of the data points (i.e., the (X_i, Y_i)'s). Combining these calculations yields (3).

Next, given R, define a "tournament matrix" $A = A(R)$ as

$$A_{ij} = \begin{cases} \mathbb{1}\left[R_{ij} > R_{ji}\right] & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases}$$

It is easily checked that $A(\Pi_\sigma R \Pi_\sigma^\top) = \Pi_\sigma A(R) \Pi_\sigma^\top$, and hence (3) implies that

$$A \stackrel{d}{=} \Pi_\sigma A \Pi_\sigma^\top. \tag{4}$$

Let $S_\alpha(A)$ be the set of row indices with row sums greater than or equal to $(1-\alpha)(n+1)$, i.e.,

$$S_\alpha(A) = \left\{ i = 1, \ldots, n+1 : \sum_{j=1}^{n+1} A_{ij} \geq (1-\alpha)(n+1) \right\}.$$

The argument of Step 3 in the proof of Barber et al. [2019, Theorem 1] applies to the lifted J+aB "tournament matrix" A, and it holds deterministically that

$$|S_\alpha(A)| \leq 2\alpha(n+1). \tag{5}$$

On the other hand, if j is any index, and σ is any permutation that swaps indices $n+1$ and j, then

$$\mathbb{P}\big[n+1 \in S_\alpha(A)\big] = \mathbb{P}\big[j \in S_\alpha(\Pi_\sigma A \Pi_\sigma^\top)\big] = \mathbb{P}\big[j \in S_\alpha(A)\big].$$

The first two events are the same, and the second equality uses (4). Thus,

$$\mathbb{P}\big[n+1 \in S_\alpha(A)\big] = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbb{P}\big[j \in S_\alpha(A)\big] = \frac{1}{n+1} \mathbb{E}\left[\sum_{j=1}^{n+1} \mathbb{1}\left[j \in S_\alpha(A)\right] \right] = \frac{\mathbb{E}|S_\alpha(A)|}{n+1} \leq 2\alpha. \tag{6}$$

Note that the event $\big[n+1 \in S_\alpha(A)\big]$ is exactly the event $\tilde{\mathcal{E}}_{n+1}$, defined in Section 2.3.2. As described in the proof sketch in Section 2.3.2 of the main paper, we can couple this lifted event to the event $\mathcal{E}_{n+1}$, also defined in Section 2.3.2 in terms of the actual jackknife+-after-bootstrap, as follows. Let $B = \sum_{b=1}^{\tilde{B}} \mathbb{1}\left[\tilde{S}_b \not\ni n+1\right]$, the number of $\tilde{S}_b$'s containing only training data, and let $1 \leq b_1 < \cdots < b_B \leq \tilde{B}$ be the corresponding indices. Note that the distribution of B is Binomial, as specified in the theorem. Now, for each $k = 1, \ldots, B$, define $S_k = \tilde{S}_{b_k}$. We can observe that each S_k is an independent uniform draw from $\{1, \ldots, n\}$ (with or without replacement). Therefore, we can equivalently consider running J+aB (Algorithm 2) with these particular subsamples or bootstrapped samples $S_1, \ldots, S_B$, in which case it holds deterministically that $\tilde{\mu}_{\varphi \backslash n+1,i} = \hat{\mu}_{\varphi \backslash i}$ for each $i = 1, \ldots, n$. This ensures that $|Y_{n+1} - \tilde{\mu}_{\varphi \backslash n+1,i}(X_{n+1})| = |Y_{n+1} - \hat{\mu}_{\varphi \backslash i}(X_{n+1})|$ and $|Y_i - \tilde{\mu}_{\varphi \backslash i,n+1}(X_i)| = |Y_i - \hat{\mu}_{\varphi \backslash i}(X_i)|$, and thus,

$$\mathbb{P}[\mathcal{E}_{n+1}] = \mathbb{P}[\tilde{\mathcal{E}}_{n+1}] \leq 2\alpha.$$

Finally, as in Step 1 in the proof of Barber et al. [2019, Theorem 1], it easily follows from the definition of $\hat{C}^{\text{J+aB}}_{\alpha,n,B}$ that if $Y_{n+1} \notin \hat{C}^{\text{J+aB}}_{\alpha,n,B}(X_{n+1})$ then the event $\mathcal{E}_{n+1}$ must occur. Indeed, if $Y_{n+1} \notin \hat{C}^{\text{J+aB}}_{\alpha,n,B}(X_{n+1})$, then either Y_{n+1} falls below the lower bound, i.e.,

$$\sum_{i=1}^{n} \mathbb{1}\left[Y_{n+1} - \hat{\mu}_{\varphi \backslash i}(X_{n+1}) < \left| Y_i - \hat{\mu}_{\varphi \backslash i}(X_i) \right| \right] \geq (1-\alpha)(n+1),$$

or Y_{n+1} exceeds the upper bound, i.e.,

$$\sum_{i=1}^{n} \mathbb{1}\left[Y_{n+1} - \hat{\mu}_{\varphi \backslash i}(X_{n+1}) > \left| Y_i - \hat{\mu}_{\varphi \backslash i}(X_i) \right| \right] \geq (1-\alpha)(n+1),$$

and the above two expressions imply

$$\sum_{i=1}^{n} \mathbb{1}\left[\left|Y_{n+1} - \hat{\mu}_{\varphi\setminus i}(X_{n+1})\right| > \left|Y_i - \hat{\mu}_{\varphi\setminus i}(X_i)\right| \right] \geq (1-\alpha)(n+1).$$

Therefore, we conclude that

$$\mathbb{P}\left[Y_{n+1} \notin \hat{C}^{\text{J+aB}}_{\alpha,n,B}(X_{n+1})\right] \leq 2\alpha,$$

thus proving the theorem.

A.2 Proof of Theorem 2

There are two essential differences between this proof and the one above. The first is the definition of the tournament matrix A and the second is the bound on $\mathbb{E}[|S_\alpha(A)|]$. Details are given below.

Given the residual matrices $\tilde{R}^b, b \in \{1, \ldots, \tilde{B}\}$ with entries $\tilde{R}^b_{ij}$ from Algorithm 5, define a tournament matrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$ as

$$A_{ij} := \begin{cases} \sum_{b:i,j\notin\tilde{S}_b} \mathbb{1}[\tilde{R}^b_{ij} > \tilde{R}^b_{ji}] & i \neq j \\ 0 & i = j \end{cases}$$

Since for all b and all permutation matrices Π_σ defined under a permutation σ on $\{1, ..., n+1\}$

$$\tilde{R}^b \overset{d}{=} \Pi_\sigma \tilde{R}^b \Pi_\sigma^\top,$$

Our tournament matrix A still enjoys the property that

$$A \overset{d}{=} \Pi_\sigma A \Pi_\sigma^\top.$$

Now, by defining

$$S_\alpha(A) = \{i = 1, ..., n+1 : \sum_{j=1, j\neq i}^{n+1} A_{ij} \geq (1-\alpha)s_i\},$$

where $s_i := \sum_{j=1,j\neq i}^{n+1} \sum_{b:(i,j)\notin\tilde{S}_b} \mathbb{1}$, we see the event $[n+1 \in S_\alpha(A)]$ is exactly the event $\tilde{\mathcal{E}}_{n+1}$ defined in section 3.3. Intuitively, s_i is the total number of comparisons between residual i and all other residuals $j, j \neq i$.

On the other hand, if j is any index, and σ is any permutation that swaps indices $n+1$ and j, then

$$\mathbb{P}\big[n+1 \in S_\alpha(A)\big] = \mathbb{P}\big[j \in S_\alpha(\Pi_\sigma A \Pi_\sigma^\top)\big] = \mathbb{P}\big[j \in S_\alpha(A)\big],$$

continues to be true as before. As a consequence,

$$\mathbb{P}\big[n+1 \in S_\alpha(A)\big] = \frac{\mathbb{E}[S_\alpha(A)]}{n+1}.$$

Finally, to bound $\mathbb{E}[S_\alpha(A)]$, we mimic Step 3 of proof of Theorem 1 in Barber et al., 2019. For each $i \in S_\alpha(A)$:

$$
\begin{aligned}
(1-\alpha)s_i &\leq \sum_{j=1, j\neq i}^{n+1} \sum_{b:(i,j)\notin \tilde{S}_b} \mathbb{1}\{\tilde{R}_{ij}^b > \tilde{R}_{ji}^b\} && \text{(by definition)} \\
&= \sum_{j\in S_\alpha(A), j\neq i} \sum_{b:(i,j)\notin \tilde{S}_b} \mathbb{1}\{\tilde{R}_{ij}^b > \tilde{R}_{ji}^b\} + \sum_{j\notin S_\alpha(A)} \sum_{b:(i,j)\notin \tilde{S}_b} \mathbb{1}\{\tilde{R}_{ij}^b > \tilde{R}_{ji}^b\} \\
&\leq \sum_{j\in S_\alpha(A), j\neq i} \sum_{b:(i,j)\notin \tilde{S}_b} \mathbb{1}\{\tilde{R}_{ij}^b > \tilde{R}_{ji}^b\} + s_i - \sum_{j\in S_\alpha(A), j\neq i} \sum_{b:(i,j)\notin \tilde{S}_b} \mathbb{1}
\end{aligned}
$$

Where the last inequality holds if we fix $\mathbb{1}\{\tilde{R}_{ij}^b > \tilde{R}_{ji}^b\}$ as 1 for $j \notin S_\alpha(A)$ and apply the definition of $s_i := \sum_{j=1, j\neq i}^{n+1} \sum_{b:(i,j)\notin \tilde{S}_b} \mathbb{1}$.

For notation simplicity, we abbreviate $S_\alpha(A)$ as S and let $s_{ij} := \sum_{b:(i,j)\notin \tilde{S}_b} \mathbb{1}$ to be the number of comparisons between residuals i and j. Continue, we see that:

$$
\sum_{j\in S, j\neq i} s_{ij} \leq \alpha s_i + \sum_{j\in S, j\neq i} \sum_{b:(i,j)\notin \tilde{S}_b} \mathbb{1}\{\tilde{R}_{ij}^b > \tilde{R}_{ji}^b\}
$$

Summing over $i \in S$, we get

$$
\begin{aligned}
\sum_{i\in S} \sum_{j\in S, j\neq i} s_{ij} &\leq \alpha \sum_{i\in S} s_i + \sum_{i\in S} \sum_{j\in S, j\neq i} \sum_{b:(i,j)\notin \tilde{S}_b} \mathbb{1}\{\tilde{R}_{ij}^b > \tilde{R}_{ji}^b\} \\
&\leq \alpha \sum_{i\in S} s_i + 0.5 \sum_{i\in S} \sum_{j\in S} s_{ij}
\end{aligned}
$$

Rearrange, we see

$$
\sum_{i\in S} \sum_{j\in S, j\neq i} s_{ij} \leq 2\alpha \sum_{i\in S} \sum_{j=1, j\neq i}^{n+1} s_{ij}
$$

Now, define $s_{min} := \min_{(i,j)\in[n+1]\times[n+1], i\neq j} s_{ij}$ and $s_{max} := \max_{(i,j)\in[n+1]\times[n+1], i\neq j} s_{ij}$ to be the minimum and maximum number of comparisons between any two distinct pair of indices. Then, we can give lower bound for the left hand side and upper bound for the right hand side in the inequality above in terms of $|S|$ as follows:

$$
\sum_{i\in S} \sum_{j\in S, j\neq i} s_{ij} \geq s_{min} |S|(|S|-1)
$$

$$
2\alpha \sum_{i\in S} \sum_{j=1, j\neq i}^{n+1} s_{ij} \leq 2\alpha s_{max} |S|[(n+1)-1]
$$

Rearrange, we see that

$$
|S| \leq 2\alpha n (\frac{s_{max}}{s_{min}}) \tag{7}
$$

Now, to ensure $\frac{s_{max}}{s_{min}}$ on the right hand side of inequality (7) is close to 1, we first notice that for each pair of distinct indices (i,j), $s_{ij} := \sum_b 1\{(i,j) \notin \tilde{S}_b\} \sim \text{Binomial}(\tilde{B}, (1-\frac{2}{n+1})^m)$. Therefore, s_{max} and s_{min} are the maximum and minimum over $\frac{(n+1)n}{2}$ many possibly dependent $\text{Binomial}(\tilde{B}, (1-\frac{2}{n+1})^m)$ random variables.

Next, given a small $\epsilon > 0$, we first want to show that $\frac{s_{max}}{s_{min}} \leq 1+\epsilon \iff \frac{s_{max}-s_{min}}{s_{min}} \leq \epsilon$ with high probability.

Define $p := (1-\frac{2}{n+1})^m$, we have each $s_{ij} \sim \text{Binomial}(\tilde{B}, p)$ with mean $\mu = \tilde{B}p$.

As a result, by defining $\epsilon' := \frac{\epsilon}{1+\frac{\epsilon}{2}}$, we get

$$\begin{aligned}
\mathbb{P}\big[s_{max} \geq (1+\frac{\epsilon'}{2})\mu\big] &= \mathbb{P}\big[\exists (i,j) \text{ pair s.t } s_{ij} \geq (1+\frac{\epsilon'}{2})\mu\big] \\
&\leq \sum_{i,j} \mathbb{P}\big[s_{ij} \geq (1+\frac{\epsilon'}{2})\mu\big] \\
&\leq \frac{n(n+1)}{2}\exp(-\tilde{B}p\frac{\epsilon'^2}{12})
\end{aligned}$$

where the first inequality follows by union bound and the last inequality follows by applying Chernoff upper tail bound on a binomial random variable with mean $\mu = \tilde{B}p$.

Similarly, $\mathbb{P}\big[s_{min} \leq (1-\frac{\epsilon'}{2})\mu\big] \leq \frac{n(n+1)}{2}\exp(-\tilde{B}p\frac{\epsilon'^2}{12})$ by union bound and Chernoff lower tail bound on the same binomial random variable.

Combine the two probability bounds, we observe that

$$\mathbb{P}\big[s_{min} \geq (1-\frac{\epsilon'}{2})\mu, s_{max} \leq (1+\frac{\epsilon'}{2})\mu\big] \geq 1 - n(n+1)\exp(-\tilde{B}p\frac{\epsilon'^2}{12})$$

which, in combination with the earlier definition of ϵ', implies that

$$\begin{aligned}
\mathbb{P}\big[\frac{s_{max}-s_{min}}{s_{min}} > \frac{\epsilon'\mu}{(1-\frac{\epsilon'}{2})\mu} = \epsilon\big] &= \mathbb{P}\big[\frac{s_{max}}{s_{min}} > 1+\epsilon\big] \\
&< n(n+1)\exp(-\tilde{B}p\frac{4\epsilon^2}{12(2+\epsilon)^2}).
\end{aligned}$$

Therefore,

$$\mathbb{P}\big[|S| \leq 2\alpha(1+\epsilon)n\big] \geq 1 - n(n+1)\exp(-\tilde{B}p\frac{4\epsilon^2}{12(2+\epsilon)^2}) \tag{8}$$

As a consequence, we obtain that

$$\begin{aligned}
\mathbb{E}[|S|] &= \mathbb{E}[|S|\,\mathbb{1}[|S| \leq 2\alpha(1+\epsilon)n] + |S|\,\mathbb{1}[|S| > 2\alpha(1+\epsilon)n]] \\
&\leq (2\alpha(1+\epsilon)n)\mathbb{P}\big[|S| \leq 2\alpha(1+\epsilon)n\big] + (n+1)\mathbb{P}\big[|S| > 2\alpha(1+\epsilon)n\big] \\
&\leq (2\alpha(1+\epsilon)n) + n(n+1)^2\exp(-\tilde{B}p\frac{4\epsilon^2}{12(2+\epsilon)^2})
\end{aligned}$$

where the first inequality holds by linearity of expectation and the fact that $|S| \leq n+1$ deterministicly and the second inequalty holds by the probability bound (8)

Combine the results, we see that

$$\begin{aligned}
\mathbb{P}\big[Y_{n+1} \notin \hat{C}^{\text{J+wB}}_{\alpha,n,B}(X_{n+1})\big] &= \mathbb{P}\big[n+1 \in S\big] \\
&= \frac{E[S_\alpha(A)]}{n+1} \\
&\leq 2\alpha(1+\epsilon) + n(n+1)\exp(-\tilde{B}p\frac{\epsilon^2}{3(2+\epsilon)^2}) \\
&= 2\alpha + \delta
\end{aligned}$$

with $\delta := 2\alpha\epsilon + n(n+1)\exp(-\tilde{B}p\frac{\epsilon^2}{3(2+\epsilon)^2})$.

Now, we can assume $\epsilon \leq \frac{1}{2\alpha}$, as otherwise $\delta > 1$, causing $\mathbb{P}\big[Y_{n+1} \notin \hat{C}^{\text{J+wB}}_{\alpha,n,B}(X_{n+1})\big]$ to be trivially bounded.

Thus, $n(n+1)\exp(-\frac{\tilde{B}p}{3}\frac{\epsilon^2}{(2+\epsilon)^2}) \leq n(n+1)\exp(-C\tilde{B}\epsilon^2)$, where $C = C(n,m,\alpha) := \frac{p}{3(2+\frac{1}{2\alpha})^2}$ is a constant.

Let $\epsilon := \sqrt{\frac{\log(n(n+1)\tilde{B})}{\tilde{B}C}}$, we see that $\epsilon \to \infty$ as $\tilde{B} \to \infty$ for any fixed n.

As a consequence, as $\tilde{B} \to \infty$,

$$2\alpha\epsilon = 2\alpha\sqrt{\frac{\log(n(n+1)\tilde{B})}{\tilde{B}C}} \to 0$$

$$n(n+1)\exp(-C\tilde{B}\epsilon^2) = \frac{1}{\tilde{B}} \to 0$$

Thus,

$$\delta \leq 2\alpha\sqrt{\frac{\log(n(n+1)\tilde{B})}{\tilde{B}C}} + \frac{1}{\tilde{B}} \to 0$$

.

Hence, $\mathbb{P}\big[Y_{n+1} \notin \hat{C}^{\text{J+wB}}_{\alpha,n,B}(X_{n+1})\big] \leq 2\alpha$ as $\tilde{B} \to \infty$

References

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Ann. Statist.*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709. URL https://projecteuclid.org:443/euclid.aos/1547197251.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+, 2019. arXiv preprint.

H. Boström, L. Asker, R. Gurung, I. Karlsson, T. Lindgren, and P. Papapetrou. Conformal prediction using random survival forests. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 812–817, 2017a. ISBN null. doi: 10.1109/ICMLA.2017.00-57.

Henrik Boström, Henrik Linusson, Tuve Löfström, and Ulf Johansson. Accelerating difficulty estimation for conformal regression forests. *Annals of Mathematics and Artificial Intelligence*, 81(1):125–144, 2017b. doi: 10.1007/s10472-017-9539-9. URL https://doi.org/10.1007/s10472-017-9539-9.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996. doi: 10.1007/BF00058655. URL https://doi.org/10.1007/BF00058655.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Peter Bühlmann and Bin Yu. Analyzing bagging. *Ann. Statist.*, 30(4):927–961, Aug 2002. doi: 10.1214/aos/1031689014. URL https://doi.org/10.1214/aos/1031689014.

Andreas Buja and Werner Stuetzle. Observations on bagging. *Statist. Sinica*, 16(2):323–351, 2006. ISSN 1017-0405.

Krisztian Buza. Feedback prediction for blogs. In Myra Spiliopoulou, Lars Schmidt-Thieme, and Ruth Janning, editors, *Data Analysis, Machine Learning and Knowledge Discovery*, pages 145–152. Springer International Publishing, 2014. ISBN 978-3-319-01595-8.

Dmitry Devetyarov and Ilia Nouretdinov. Prediction with confidence based on a random forest classifier. In Harris Papadopoulos, Andreas S. Andreou, and Max Bramer, editors, *Artificial Intelligence Applications and Innovations*, pages 37–44, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-16239-8.

Bradley Efron. Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):83–127, 1992. ISSN 00359246. URL http://www.jstor.org/stable/2345949.

Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014. doi: 10.1080/01621459.2013.823775. URL https://doi.org/10.1080/01621459.2013.823775. PMID: 25346558.

Trena M Ezzati-Rice, Frederick Rohde, and Janet Greenblatt. Sample design of the medical expenditure panel survey household component, 1998–2007. Methodology Report 22, Agency for Healthcare Research and Quality, Rockville, MD, Mar 2008. URL http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr22/mr22.pdf.

Jerome H. Friedman and Peter Hall. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683, 2007. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2006.06.002. URL http://www.sciencedirect.com/science/article/pii/S0378375806001339. Special Issue on Nonparametric Statistics and Related Topics: In honor of M.L. Puri.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Ensemble Learning*, pages 605–624. Springer New York, 2009. ISBN 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7_16. URL https://doi.org/10.1007/978-0-387-84858-7_16.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282. IEEE, Aug 1995. doi: 10.1109/ICDAR.1995.598994.

Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine Learning*, 97(1):155–176, 2014. doi: 10.1007/s10994-014-5453-0. URL https://doi.org/10.1007/s10994-014-5453-0.

Byol Kim, Chen Xu, and Rina Foygel Barber. Predictive inference is free with the jackknife+-after-bootstrap. *arXiv preprint arXiv:2002.09025*, 2020.

Arun K. Kuchibhotla and Aaditya K. Ramdas. Nested conformal prediction and the generalized jackknife+, 2019. arXiv preprint.

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. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

H. Linusson, U. Johansson, and H. Boström. Efficient conformal predictor ensembles. *Neurocomputing*, 2019. doi: https://doi.org/10.1016/j.neucom.2019.07.113. URL http://www.sciencedirect.com/science/article/pii/S0925231219316108.

T. Löfström, U. Johansson, and H. Boström. Effective utilization of data in inductive conformal prediction using ensembles of neural networks. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013. ISBN 2161-4393. doi: 10.1109/IJCNN.2013.6706817.

Benjamin Lu and Johanna Hardin. A unified framework for random forest prediction error estimation, 2019. arXiv preprint.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-33745174860&partnerID=40&md5=93784fdaa267f35c561a576c66aad9ff.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41, 2016. URL http://jmlr.org/papers/v17/14-168.html.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In Paula Fritzsche, editor, *Tools in Artificial Intelligence*, pages 325–330. InTech, 2008. ISBN 978-953-7619-03-9. URL http://www.intechopen.com/books/tools_in_artificial_intelligence/inductive_conformal_prediction__theory_and_application_to_neural_networks.

Harris Papadopoulos and Haris Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011. doi: https://doi.org/10.1016/j.neunet.2011.05.008. URL http://www.sciencedirect.com/science/article/pii/S089360801100150X.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36755-0.

Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, Mar 2006. ISSN 1558-0830. doi: 10.1109/MCAS.2006.1688199.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(Mar):660–678, 2002. ISSN 0377-2217. doi: 10.1016/S0377-2217(01)00264-8. URL http://www.sciencedirect.com/science/article/pii/S0377221701002648.

Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39, Feb 2010. ISSN 1573-7462. doi: 10.1007/s10462-009-9124-7. URL https://doi.org/10.1007/s10462-009-9124-7.

Marie-Hélène Roy and Denis Larocque. Prediction intervals with random forests. *Statistical Methods in Medical Research*, 29(1):205–229, 2020/03/01 2019. doi: 10.1177/0962280219829885. URL https://doi.org/10.1177/0962280219829885.

Joseph Sexton and Petter Laake. Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis*, 53(3):801–811, 2009. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2008.08.007. URL http://www.sciencedirect.com/science/article/pii/S0167947308003988. Computational Statistics within Clinical Research.

Lukas Steinberger and Hannes Leeb. Leave-one-out prediction intervals in linear regression models with many variables, 2016. arXiv preprint.

Lukas Steinberger and Hannes Leeb. Conditional predictive inference for high-dimensional stable algorithms, 2018. arXiv preprint.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2-3):349–376, 2013. ISSN 08856125. doi: 10.1007/s10994-013-5355-6. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84880107869&doi=10.1007%2fs10994-013-5355-6&partnerID=40&md5=18a26112b5a5b1b33e6b108388ea3854.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, 2005. ISBN 978-0-387-25061-8. doi: 10.1007/b106715. URL https://doi.org/10.1007/b106715.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15: 1625–1651, 2014. URL http://jmlr.org/papers/v15/wager14a.html.

Haozhe Zhang, Joshua Zimmerman, Dan Nettleton, and Daniel J. Nordman. Random forest prediction intervals. *The American Statistician*, pages 1–15, 04 2019. doi: 10.1080/00031305.2019.1585288. URL https://doi.org/10.1080/00031305.2019.1585288.