Predictive inference with the jackknife+

Rina Foygel Barber*, Emmanuel J. Candès†,
Aaditya Ramdas‡, Ryan J. Tibshirani‡§

June 2, 2020

Abstract

This paper introduces the *jackknife+*, which is a novel method for constructing predictive confidence intervals. Whereas the jackknife outputs an interval centered at the predicted response of a test point, with the width of the interval determined by the quantiles of leave-one-out residuals, the jackknife+ also uses the leave-one-out predictions at the test point to account for the variability in the fitted regression function. Assuming exchangeable training samples, we prove that this crucial modification permits rigorous coverage guarantees regardless of the distribution of the data points, for any algorithm that treats the training points symmetrically. Such guarantees are not possible for the original jackknife and we demonstrate examples where the coverage rate may actually vanish. Our theoretical and empirical analysis reveals that the jackknife and the jackknife+ intervals achieve nearly exact coverage and have similar lengths whenever the fitting algorithm obeys some form of stability. Further, we extend the jackknife+ to K-fold cross validation and similarly establish rigorous coverage properties. Our methods are related to *cross-conformal prediction* proposed by Vovk [2015] and we discuss connections.

1 Introduction

Suppose that we have i.i.d. training data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$, and a new test point (X_{n+1}, Y_{n+1}) drawn independently from the same distribution. We would like to fit a regression model to the training data, i.e., a function $\widehat{\mu} : \mathbb{R}^d \to \mathbb{R}$ where $\widehat{\mu}(x)$ predicts Y_{n+1} given a new feature vector $X_{n+1} = x$, and then provide a

*Department of Statistics, University of Chicago
†Departments of Statistics and Mathematics, Stanford University
‡Department of Statistics and Data Science, Carnegie Mellon University
§Machine Learning Department, Carnegie Mellon University

prediction interval for the test point—an interval around $\widehat{\mu}(X_{n+1})$ that is likely to contain the true test response value Y_{n+1}. Specifically, given some target coverage level $1-\alpha$, we would like to construct a prediction interval $\widehat{C}_{n,\alpha}$ as a function of the n training data points, such that

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1})\right\} \geq 1-\alpha,$$

where the probability is taken with respect to a new test point (X_{n+1}, Y_{n+1}) as well as with respect to the training data.

A naive solution might be to use the residuals on the training data, $|Y_i - \widehat{\mu}(X_i)|$, to estimate the typical prediction error on the new test point—for instance, we might consider the prediction interval

$$\widehat{\mu}(X_{n+1}) \pm \Big(\text{the } (1-\alpha) \text{ quantile of } |Y_1 - \widehat{\mu}(X_1)|, \ldots, |Y_n - \widehat{\mu}(X_n)|\Big). \tag{1}$$

However, in practice, this interval would typically undercover (meaning that the probability that Y_{n+1} lies in this interval would be lower than the target level $1-\alpha$), since due to overfitting, the residuals on the training data points $i = 1, \ldots, n$ are typically smaller than the residual on the previously unseen test point, i.e., $|Y_{n+1} - \widehat{\mu}(X_{n+1})|$.

In order to avoid the overfitting problem, the jackknife prediction method computes a margin of error with a leave-one-out construction:

- For each $i = 1, \ldots, n$, fit the regression function $\widehat{\mu}_{-i}$ to the training data with the ith point removed, and compute the corresponding leave-one-out residual, $|Y_i - \widehat{\mu}_{-i}(X_i)|$.
- Fit the regression function $\widehat{\mu}$ to the full training data, and output the prediction interval

$$\widehat{\mu}(X_{n+1}) \pm \Big(\text{the } (1-\alpha) \text{ quantile of } |Y_1 - \widehat{\mu}_{-1}(X_1)|, \ldots, |Y_n - \widehat{\mu}_{-n}(X_n)|\Big). \tag{2}$$

Intuitively, this method should have the right coverage properties on average since it avoids overfitting—the leave-one-out residuals $|Y_i - \widehat{\mu}_{-i}(X_i)|$ reflect the typical magnitude of the error in predicting a new data point after fitting to a sample size n (or, almost equivalently, $n-1$), unlike the naive method where the residuals on the training data are likely to be too small due to overfitting.

However, the jackknife procedure does not have any universal theoretical guarantees. Although many results are known under asymptotic settings or under assumptions of stability of the regression algorithm $\widehat{\mu}$ (we will give an overview below), it is nonetheless the case that, for settings where $\widehat{\mu}$ is unstable, the jackknife method may lose predictive coverage—for example, we will see in our simulations in Section 7 that the jackknife can have extremely poor coverage using least squares regression when the sample size n is close to the dimension d.

In this paper, we introduce a new method, the *jackknife+*, that provides non-asymptotic coverage guarantees under no assumptions beyond the training and test data being exchangeable. We will see that the jackknife+ offers, in the worst case, a $1-2\alpha$ coverage rate (where $1-\alpha$ is the target), while the original jackknife may even have zero coverage in degenerate examples. On the other hand, empirically we often observe that the two methods yield nearly identical intervals and both achieve $1-\alpha$ coverage. Theoretically, we will see that under a suitable notion of stability, the jackknife+ and jackknife both provably yield close to $1-\alpha$ coverage.

1.1 Background

The idea of resampling or subsampling from the available data, in order to assess the accuracy of our parameter estimates or predictions, has a rich history in the statistics literature. Early works developing the jackknife and bootstrap methods include Quenouille [1949, 1956], Tukey [1958], Miller [1974], Efron [1979], Stine [1985]. Several papers from this period include leave-one-out methods for assessing or calibrating predictive accuracy, similar to the predictive interval constructed in (2) above, e.g., Stone [1974], Geisser [1975], Butler and Rothman [1980], generally using the term "cross-validation" to refer to this approach. (In this work, we will instead use the term "jackknife" to refer to the leave-one-out style of prediction methods, as is common in the modern literature.) Efron and Gong [1983] provides an overview of the early literature on these types of methods.

While this rich literature demonstrated extensive evidence of the reliable performance of the jackknife in practice, relatively little has been known about the theoretical properties of this type of method until recently. Steinberger and Leeb [2016, 2018] have developed results proving valid predictive coverage of the jackknife under assumptions of algorithmic stability, meaning that the fitted model $\widehat{\mu}$ and its leave-one-out version $\widehat{\mu}_{-i}$ are required to give similar predictions at the test point. This work builds on earlier results by Bousquet and Elisseeff [2002], which study generalization bounds for risk minimization through the framework of stability conditions; an earlier work in this line is that of Devroye and Wagner [1979], which give analogous results for classification risk.

In contrast to cross-validation methods, which perform well but are difficult to analyze theoretically, we can instead consider a simple *validation* or *holdout* method. We first partition the training data as $\{1,\ldots,n\} = S_{\text{train}} \cup S_{\text{holdout}}$, then fit $\widehat{\mu}_{\text{train}}$ on the subset S_{train} of the training data and construct a predictive interval

$$\widehat{\mu}_{\text{train}}(X_{n+1}) \pm \Big(\text{the } (1-\alpha) \text{ quantile of } |Y_i - \widehat{\mu}_{\text{train}}(X_i)|,\, i \in S_{\text{holdout}}\Big). \tag{3}$$

Papadopoulos [2008], Vovk [2012], Lei et al. [2018] study this type of method, under the name *split conformal prediction* or *inductive conformal prediction*, through the framework of exchangeability, and prove that $1-\alpha$ predictive coverage holds with

no assumptions on the algorithm $\mathcal{A}$ or on the distribution of the data (with a small correction to the definition of the quantile). This method is also computationally very cheap, as we only need to fit a single regression function $\widehat{\mu}_{\text{train}}$—in contrast, jackknife and cross-validation methods require running the regression many times. However, these benefits come at a statistical cost. If the training size $|S_{\text{train}}|$ is much smaller than n, then the fitted model $\widehat{\mu}_{\text{train}}$ may be a poor fit, leading to wide prediction intervals; if instead we decide to take $|S_{\text{train}}| \approx n$ then instead $|S_{\text{holdout}}|$ is very small, leading to high variability.

Finally, Vovk [2015], Vovk et al. [2018] proposed the *cross-conformal prediction* method, which is closely related to the jackknife+. We describe the cross-conformal method, and the previously known theoretical guarantees, in detail later on. Their work is based on the conformal prediction method (see Vovk et al. [2005], Lei et al. [2018] for background), which provably achieves distribution-free predictive coverage at the target level $1-\alpha$ but at an extremely high computational cost.

1.2 Notation

Before proceeding, we first define some notation. First, for any values v_i indexed by $i = 1, \ldots, n$, define[1]

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

the $1-\alpha$ quantile of the empirical distribution of these values. Similarly, we will let $\widehat{q}^{-}_{n,\alpha}\{v_i\}$ denote the α quantile of the distribution,

$$\widehat{q}^{-}_{n,\alpha}\{v_i\} = \text{ the } \lfloor \alpha(n+1) \rfloor \text{-th smallest value of } v_1, \ldots, v_n \; = -\widehat{q}^{+}_{n,\alpha}\{-v_i\}.$$

With this notation, the "naive" prediction interval in (1) can be defined as

$$\widehat{C}^{\text{naive}}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}^{+}_{n,\alpha}\big\{|Y_i - \widehat{\mu}(X_i)|\big\}. \tag{4}$$

Second, we will write $\mathcal{A}$ to denote the algorithm mapping a training data set of any size, to the fitted regression function. Formally, $\mathcal{A}$ is a map from $\cup_{m \geq 1} \big(\mathbb{R}^d \times \mathbb{R}\big)^m$ (i.e., the collection of all training sets of any size $m \geq 1$), to the space of functions $\mathbb{R}^d \to \mathbb{R}$. For example, when $\widehat{\mu}$ is the regression function fitted on the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we can write

$$\widehat{\mu} = \mathcal{A}\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big). \tag{5}$$

[1] In defining the quantiles $\widehat{q}^{+}_{n,\alpha}$ of the residuals, we use $(1-\alpha)(n+1)$ rather than $(1-\alpha)n$ to correct for the finite sample size—we will see later on why this correction is natural. For the jackknife, it is perhaps more common to see n in place of $n+1$, i.e., the residual quantile is defined slightly differently, but for large n the difference is negligible. Formally, if $\alpha < \frac{1}{n+1}$ and so $(1-\alpha)(n+1) > n$, then we set $\widehat{q}^{+}_{n,\alpha}\{v_i\} = \infty$.

Similarly, to compute the leave-one-out residuals for the jackknife, we let

$$\widehat{\mu}_{-i} = \mathcal{A}\Big((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)\Big), \tag{6}$$

and then the jackknife prediction interval (2) can be written as

$$\widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}^{+}_{n,\alpha}\big\{R^{\text{LOO}}_i\big\}, \tag{7}$$

where $R^{\text{LOO}}_i = |Y_i - \widehat{\mu}_{-i}(X_i)|$ denotes the ith leave-one-out residual.

From this point on, we will assume without comment that $\mathcal{A}$ satisfies a symmetry condition, namely, $\mathcal{A}$ must be invariant to reordering the data, i.e.,

$$\mathcal{A}\Big((X_{\pi(1)}, Y_{\pi(1)}), \ldots, (X_{\pi(m)}, Y_{\pi(m)})\Big) = \mathcal{A}\Big((X_1, Y_1), \ldots, (X_m, Y_m)\Big) \tag{8}$$

for any sample size $m \geq 1$, any points $(X_1, Y_1), \ldots, (X_m, Y_m)$, and any permutation π of the indices $\{1, \ldots, m\}$.

2 The jackknife+

Our jackknife+ method is a modification of the jackknife (7). Defining $\widehat{\mu}_{-i}$ as in (6), the jackknife+ prediction interval is given by:

$$\widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1}) = \left[\widehat{q}^{-}_{n,\alpha}\big\{\widehat{\mu}_{-i}(X_{n+1}) - R^{\text{LOO}}_i\big\},\ \widehat{q}^{+}_{n,\alpha}\big\{\widehat{\mu}_{-i}(X_{n+1}) + R^{\text{LOO}}_i\big\}\right]. \tag{9}$$

To compare this to the usual jackknife, we observe that $\widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})$ can equivalently be defined as

$$\widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1}) = \left[\widehat{q}^{-}_{n,\alpha}\big\{\widehat{\mu}(X_{n+1}) - R^{\text{LOO}}_i\big\},\ \widehat{q}^{+}_{n,\alpha}\big\{\widehat{\mu}(X_{n+1}) + R^{\text{LOO}}_i\big\}\right].$$

The constructions of the usual jackknife and the new jackknife+ are compared in Figure 1. While both versions of jackknife use the leave-one-out residuals, the difference is that for jackknife, we center our interval on the predicted value $\widehat{\mu}(X_{n+1})$ fitted on the full training data, while for jackknife+ we use the leave-one-out predictions $\widehat{\mu}_{-i}(X_{n+1})$ for the test point.

Figure 1 illustrates that, if the leave-one-out fitted functions $\widehat{\mu}_{-i}$ are all quite similar to $\widehat{\mu}$, which was fitted on the full training data, then the two methods should return nearly identical prediction intervals. On the other hand, in settings where the regression algorithm is extremely sensitive to the training data, such that removing one data point can substantially change the predicted value at X_{n+1}, the output may be quite different. In Section 5, we will examine the role of this type of instability in $\widehat{\mu}$ more closely.

To give one further interpretation of the difference between the two methods, while the jackknife interval $\widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})$ is defined as a symmetric interval around

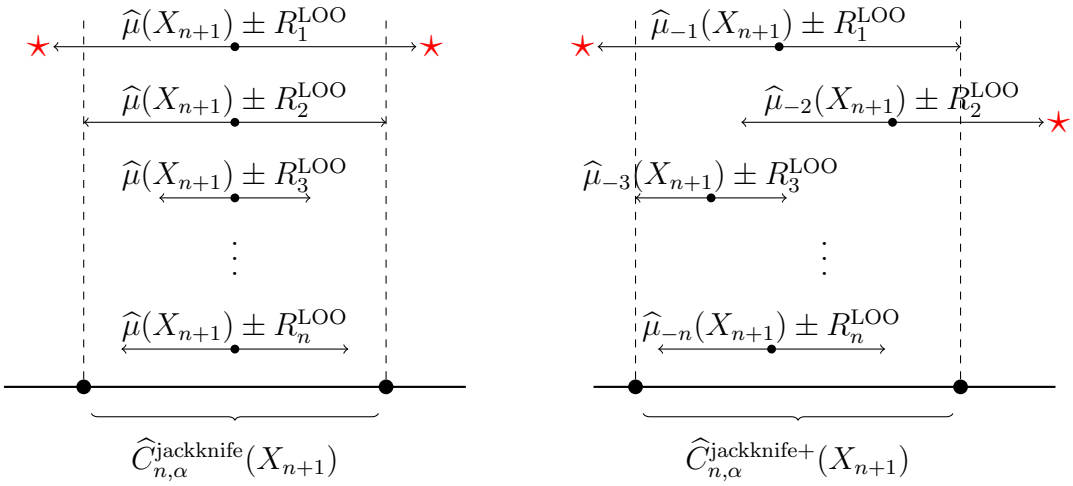


Figure 1: Illustration of the usual jackknife and the new jackknife+. The resulting prediction intervals are chosen so that, on either side, the boundary is exceeded by a sufficiently small proportion of the two sided arrows—above, these are marked with a star.

the prediction $\widehat{\mu}(X_{n+1})$ for the test point (7), the jackknife+ interval can be interpreted as an interval around the *median* prediction,

$$\mathrm{Median}\Big(\widehat{\mu}_{-1}(X_{n+1}), \ldots, \widehat{\mu}_{-n}(X_{n+1})\Big),$$

which is guaranteed to lie inside $\widehat{C}^{\mathrm{jackknife+}}_{n,\alpha}(X_{n+1})$ for $\alpha \leq \frac{1}{2}$ (in general, however, the jackknife+ interval will not be symmetric around this median prediction).

As detailed in Section 7, the jackknife and jackknife+ often perform nearly identically in practice (and generally achieve an empirical coverage level very close to the target $1-\alpha$), but in some more challenging examples where the regression algorithm is less stable, the original jackknife may lose coverage while jackknife+ still achieves the target coverage level.

Finally, we remark that in settings where the distribution of $Y|X$ is highly skewed, it may be more natural to consider an asymmetric version of this method; we consider this extension in Appendix A.

2.1 Assumption-free guarantees

Remarkably, although the jackknife+ method appears to only be a slight modification of jackknife, our main result proves that the jackknife+ is guaranteed to achieve predictive coverage at the level $1-2\alpha$, without making any assumptions on the distribution of the data (X, Y) or the nature of the regression method $\mathcal{A}$.

For this theorem, and all results that follow, all probabilities are stated with respect to the distribution of the training data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the

test data point (X_{n+1}, Y_{n+1}) drawn i.i.d. from an arbitrary distribution P, and we assume implicitly that the regression method $\mathcal{A}$ is invariant to the ordering of the data (8). We will treat the sample size $n \geq 2$ and the target coverage level $\alpha \in [0, 1]$ as fixed throughout.

Theorem 1. *The jackknife+ prediction interval satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - 2\alpha.$$

This result is proved in Section 6 using the exchangeability of the $n+1$ data points $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$—we remark that this theorem actually holds more generally under the assumption that the $n+1$ data points are exchangeable, with the i.i.d. assumption as a special case.

In practice, we generally expect to achieve the target level $1 - \alpha$ with either version of the jackknife. A natural question is whether the factor of 2 appearing in the coverage guarantee for jackknife+ is real, or is merely an artifact of the proof. We would also want to know whether analogous results may be possible for the original jackknife.

In fact, our next result constructs explicit pathological examples to see that, without making assumptions, we cannot improve our theoretical guarantee for the jackknife+ (i.e., we cannot remove the factor of 2 appearing in Theorem 1), and no guarantee at all is possible for the jackknife. For completeness, we also construct an example to show zero coverage for the naive method, although for that method we expect to see undercoverage in practice, not just in pathological examples.

Theorem 2. *For any sample size $n \geq 2$, any $\alpha \in [\frac{1}{n+1}, 1]$, and any dimension $d \geq 1$, there exists a distribution on $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ and a regression algorithm $\mathcal{A}$, such that the predictive coverage of the naive prediction interval (4) and the jackknife prediction interval (7) satisfy*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{naive}}_{n,\alpha}(X_{n+1})\right\} = \mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})\right\} = 0.$$

Furthermore, if $\alpha \leq \frac{1}{2}$, there exists a distribution on $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ and a regression algorithm $\mathcal{A}$, such that the predictive coverage of jackknife+ satisfies

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\right\} \leq 1 - 2\alpha + 6\sqrt{\frac{\log n}{n}}.$$

The proof of this theorem, and the proofs for all our theoretical results presented below, are deferred to Appendix B. The example for the original jackknife is simple—we choose the regression algorithm $\mathcal{A}$ so that models fitted at sample size n are always less accurate than models fitted at sample size $n-1$. The construction for jackknife+ is substantially more technical, and is similar in spirit to the example

sketched in Vovk [2015, Appendix A] for cross-conformal predictors in the setting of exchangeable data. (The constant 6 on the vanishing term in the bound for jackknife+ is simply an artifact of the proof, and can certainly be improved with a more careful construction.)

2.2 The jackknife-minmax method

We have seen that the best possible coverage guarantee for jackknife+, in the assumption-free setting, is $1-2\alpha$ rather than the target level $1-\alpha$. To address this gap, we can consider a more conservative alternative to the jackknife+, which will remove the factor of 2 from the theoretical bound. We define the jackknife-minmax method as follows:

$$\widehat{C}^{\text{jack-mm}}_{n,\alpha}(X_{n+1}) = \left[\min_{i=1,\ldots,n} \widehat{\mu}_{-i}(X_{n+1}) - \widehat{q}^{+}_{n,\alpha}\big\{R^{\text{LOO}}_i\big\},\ \max_{i=1,\ldots,n} \widehat{\mu}_{-i}(X_{n+1}) + \widehat{q}^{+}_{n,\alpha}\big\{R^{\text{LOO}}_i\big\}\right]. \quad (10)$$

It is simple to verify that this interval is strictly more conservative than jackknife+, meaning that for any data set, we have

$$\widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1}) \subseteq \widehat{C}^{\text{jack-mm}}_{n,\alpha}(X_{n+1}).$$

The advantage that jackknife-minmax provides is that, without any assumptions on the algorithm or distribution of the data, it always achieves the target coverage rate.

Theorem 3. *The jackknife-minmax prediction interval satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jack-mm}}_{n,\alpha}(X_{n+1})\right\} \geq 1-\alpha.$$

In practice, however, we will see that the jackknife-minmax prediction interval is generally too conservative.

3 CV+ for K-fold cross-validation

Suppose that we split the training sample into K disjoint subsets $S_1, \ldots, S_K$ each of size $m = n/K$ (assumed to be an integer). Let

$$\widehat{\mu}_{-S_k} = \mathcal{A}\Big((X_i, Y_i) : i \in \{1,\ldots,n\}\backslash S_k\Big)$$

be the regression function fitted onto the training data with the kth subset removed. To assess the quality of our regression algorithm using cross-validation (CV), we would consider the residuals from this K-fold process, namely,

$$R^{\text{CV}}_i = \big|Y_i - \widehat{\mu}_{-S_{k(i)}}(X_i)\big|,\ i = 1,\ldots,n,$$

where $k(i) \in \{1, \ldots, K\}$ identifies the subset that contains i, i.e., $i \in S_{k(i)}$. Using these residuals, we can define the CV+ prediction interval as

$$\widehat{C}^{\mathrm{CV+}}_{n,K,\alpha}(X_{n+1}) = \left[\widehat{q}^{-}_{n,\alpha}\left\{\widehat{\mu}_{-S_{k(i)}}(X_{n+1}) - R^{\mathrm{CV}}_i\right\},\ \widehat{q}^{+}_{n,\alpha}\left\{\widehat{\mu}_{-S_{k(i)}}(X_{n+1}) + R^{\mathrm{CV}}_i\right\}\right]. \quad (11)$$

Of course, jackknife+ can be viewed as a special case of CV+, by setting $K = n$. The advantage of the CV+ method, when we choose a smaller K, is that we only need to compute K rather than n models—however, this will likely come at the cost of slightly wider intervals, because the models $\widehat{\mu}_{-S_k}$ are fitted using a lower sample size (i.e., $n(1-1/K)$) and will lead to slightly larger residuals.

3.1 Assumption-free guarantee for CV+

Our next result verifies that the CV+ prediction interval enjoys essentially the same worst-case coverage guarantee as jackknife+.

Theorem 4. *The K-fold CV+ prediction interval satisfies the following coverage guarantees:*

(a) (Adapted from Vovk and Wang [2012], Vovk et al. [2018].)

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\mathrm{CV+}}_{n,K,\alpha}(X_{n+1})\right\} \geq 1 - 2\alpha - \frac{2(1-1/K)}{n/K+1}.$$

(b)

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\mathrm{CV+}}_{n,K,\alpha}(X_{n+1})\right\} \geq 1 - 2\alpha - \frac{1-K/n}{K+1}.$$

Combining the two bounds, it follows that for all K,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\mathrm{CV+}}_{n,K,\alpha}(X_{n+1})\right\} \geq 1 - 2\alpha - \min\left\{\frac{2(1-1/K)}{n/K+1}, \frac{1-K/n}{K+1}\right\} \geq 1 - 2\alpha - \sqrt{2/n}.$$

The first part of this result, part (a), is derived from the work Vovk and Wang [2012], Vovk et al. [2018]—we give more details on this in Section 3.2 below. This known result proves that the worst-case coverage is essentially $1-2\alpha$ when K is sufficiently small, i.e., $K \ll n$. Our new work, proving part (b), completes the picture by giving a meaningful bound for the case where K is large (at the extreme, $K = n$ for leave-one-out methods). By combining the two bounds, we see that coverage is essentially $1-2\alpha$ at *any* K, since the excess noncoverage is at most $\sqrt{2/n}$ uniformly over any choice of K.

We can also compare our result to the holdout or split conformal method (3), which is equivalent to fitting a model $\widehat{\mu}_{-S_1}$ and constructing the prediction interval

using the quantile of the residuals R_i^{CV} for $i \in S_1$, but using only a single subset S_1 (without repeating K times for each fold in the partition $S_1, \ldots, S_K$). As discussed earlier, this method offers an assumption-free guarantee of $1-\alpha$ coverage, but this comes at the cost of higher variance due to the single split—in contrast, CV+ reduces variance by averaging over all K splits, but at the cost of a weaker theoretical guarantee.

3.2 Related method: cross-conformal predictors

Our proposed CV+ prediction interval is related to the *cross-conformal prediction* method of Vovk [2015], Vovk et al. [2018], which (in its symmetric version) returns the predictive set

$$\widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1}) = \Bigg\{ y \in \mathbb{R} : \frac{\tau + \sum_{i=1}^{n} \mathbb{1}\left\{ \left| y - \widehat{\mu}_{-S_{k(i)}}(X_{n+1}) \right| < R_i^{\mathrm{CV}} \right\} + \tau \mathbb{1}\left\{ \left| y - \widehat{\mu}_{-S_{k(i)}}(X_{n+1}) \right| = R_i^{\mathrm{CV}} \right\}}{n+1} > \alpha \Bigg\}. \tag{12}$$

Here $\tau \sim \mathrm{Unif}[0,1]$ introduces randomization into the method. By comparing to CV+, we can verify that

$$\widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1}) \subseteq \widehat{C}^{\mathrm{CV+}}_{n,K,\alpha}(X_{n+1}) \tag{13}$$

deterministically (we demonstrate this in Appendix B.2.1). The two methods will sometimes produce the same output, but not always—in particular, $\widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1})$ may in principle return a predictive set that is a disjoint union of multiple intervals, while CV+ always returns an interval.

We next compare our theoretical findings with the known results for cross-conformal. Vovk et al. [2018] show that the K-fold cross-conformal method has coverage at least[2]

$$1 - 2\alpha - 2(1-\alpha)\frac{1-1/K}{n/K+1}. \tag{14}$$

When K is small, this additional term is negligible, and so we essentially have $1-2\alpha$ coverage for cross-conformal. However for large K, such as $K = n$ for the leave-one-out method, their earlier result does not yield a meaningful guarantee—the guaranteed coverage level is zero. In contrast, our new assumption-free result

[2] Vovk et al. [2018] do not state this coverage result directly, but instead prove $1-2\alpha$ coverage for a modification of the cross-conformal method; however, these two formulations can be shown to be equivalent. We give details in Appendix B.2.1.

in Theorem 1 proves $1-2\alpha$ coverage for the jackknife+ method (i.e., with $K=n$ folds), and Theorem 4 ensures $1-2\alpha-\sqrt{2/n}$ coverage for K-fold CV+ at any choice of K.

Remark 1. *By examining the proofs of Theorems 1 and 4, we can see that the arguments apply directly to the K-fold (or n-fold) cross-conformal method; that is, our proofs for these theorems also establish that*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\mathrm{CV+}}_{n,K,\alpha}(X_{n+1})\right\} \geq 1-2\alpha-\min\left\{\frac{2(1-1/K)}{n/K+1}, \frac{1-K/n}{K+1}\right\} \geq 1-2\alpha-\sqrt{2/n}$$

for K-fold cross-conformal with any K. In the special case that $K=n$ we are guaranteed coverage $\geq 1-2\alpha$. The first term in the minimum was established by Vovk et al. [2018] as presented in (14) above, but the second term (which allows for meaningful coverage for large values of K, e.g., $K=n$) is a new result.

3.3 An alternative method: conformal prediction

The final related method we present is *conformal prediction* [Vovk et al., 2005]. (We will sometimes refer to this method as "full" conformal prediction in order to distinguish it from the split conformal or cross-conformal methods described above.) Given the base algorithm $\mathcal{A}$, the full conformal prediction method outputs a prediction set (which consists of a union of one or more intervals) constructed as follows:

$$\widehat{C}^{\mathrm{conf}}_{n,\alpha}(X_{n+1}) = \left\{y \in \mathbb{R} : \left|y-\widehat{\mu}^{y}(X_{n+1})\right| \leq \widehat{q}^{+}_{n,\alpha}\left\{\left|Y_i-\widehat{\mu}^{y}(X_i)\right|\right\}\right\}, \tag{15}$$

where

$$\widehat{\mu}^{y} = \mathcal{A}\Big((X_1,Y_1),\ldots,(X_n,Y_n),(X_{n+1},y)\Big)$$

denotes the output of the algorithm run on the training data augmented with the hypothesized test point (X_{n+1},y). In other words, to determine whether to include a value y in the prediction set at a new point X_{n+1}, we need to train the algorithm on the training+test data (as though $Y_{n+1}=y$ were the true response value), and then see whether the residual of the test point "conforms" with the residuals on the remaining n points. The exchangeability of the test and training data ensures that

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\mathrm{conf}}_{n,\alpha}(X_{n+1})\right\} \geq 1-\alpha,$$

i.e., coverage at the target level. However, this desirable theoretical property comes at a high cost—we can see by construction of the prediction interval $\widehat{C}^{\mathrm{conf}}_{n,\alpha}(x)$ that the training algorithm $\mathcal{A}$ needs to be rerun for every test point feature vector x

we might consider, and for every possible response value $y \in \mathbb{R}$ (or in practice, for each y in a fine grid over $\mathbb{R}$). In certain special cases there are computational tricks allowing for efficient calculation of the prediction set—for example linear regression or ridge regression [Burnaev and Vovk, 2014], and the Lasso [Lei, 2017]. Outside of these special cases, full conformal is prohibitively expensive in practice, even on moderately sized data sets; while it provides an extremely elegant and theoretically rigorous framework for distribution-free inference, it is not practical in many applied settings.

4 Summary of coverage guarantees and computational costs

In light of these theoretical results, which method should a statistician choose in practice? The following table summarizes the theoretical results behind each of the methods under consideration, and the typical empirical performance that we have observed.

Method	Assumption-free theory	Typical empirical coverage
Naive (4)	No guarantee	$< 1-\alpha$
Split conf. (holdout) (3)	$\geq 1-\alpha$ coverage	$\approx 1-\alpha$
Jackknife (7)	No guarantee	$\approx 1-\alpha$, or $< 1-\alpha$ if $\widehat{\mu}$ unstable
Jackknife+ (9)	$\geq 1-2\alpha$ coverage	$\approx 1-\alpha$
Jackknife-minmax (10)	$\geq 1-\alpha$ coverage	$> 1-\alpha$
Full conformal (15)	$\geq 1-\alpha$ coverage	$\approx 1-\alpha$, or $> 1-\alpha$ if $\widehat{\mu}$ overfits
K-fold CV+ (11)	$\geq 1-2\alpha$ coverage	$\gtrsim 1-\alpha$
K-fold cross-conf. (12)	$\geq 1-2\alpha$ coverage	$\gtrsim 1-\alpha$

Given the theoretical and empirical properties of the various options, we therefore recommend the jackknife+ as a practical alternative to the usual jackknife predictive intervals. On the one hand, the empirical performance of the jackknife+ is nearly identical to that of the original jackknife (assuming we avoid pathological examples), with both methods giving intervals of nearly the same width and achieving close to the target $1-\alpha$ coverage level. However, while the jackknife offers no theoretical guarantees in the absence of stability assumptions, the jackknife+ achieves at least $1-2\alpha$ coverage in the worst possible case. On the other hand, the methods achieving $1-\alpha$ (rather than $1-2\alpha$) coverage guarantees are either less statistically efficient in the sense of producing wider intervals (split conformal uses models fitted on a smaller portion of the data while jackknife-minmax is generally too conservative), or suffer from computational infeasibility (full conformal is computationally prohibitive aside from perhaps a few special cases).

We now turn to a direct comparison of the computational costs of these eight methods. The split conformal and naive methods require only one run of the regression algorithm $\mathcal{A}$ (to fit $\widehat{\mu}$ on the full training data), while each of the jackknife methods requires n runs (to fit $\widehat{\mu}_{-i}$ for each $i = 1, \ldots, n$—and one additional run to fit $\widehat{\mu}$, in the case of the original jackknife). If the training sample size n is so large that fitting n regression functions is not feasible, we may instead prefer to use K-fold cross-validation. In contrast, the full conformal method must train $\mathcal{A}$ many more times—once for each possible combination of a test point feature vector x and a possible response value y—except for special cases such as linear regression or ridge regression. These observations are summarized below:

Method	Model training cost	Model evaluation cost
Naive (4)	1	$n + n_{\text{test}}$
Split conf. (holdout) (3)	1	"
Jackknife (7)	n	"
Jackknife+ (9)	n	$n_{\text{test}} \cdot n$
Jackknife-minmax (10)	n	"
K-fold CV+ (11)	K	$n + n_{\text{test}} \cdot K$
K-fold cross-conf. (12)	K	"
Full conformal (15)	$n_{\text{test}} \cdot n_{\text{grid}}$	$n_{\text{test}} \cdot n_{\text{grid}} \cdot n$

This table compares the computational cost (ignoring constants) of each method when run on a training sample of size n, for producing prediction sets on n_{test} many test points. The middle column ("Model training cost") counts the number of times that the model fitting algorithm $\mathcal{A}$ is run on a training data set[3] of size (up to) n. The value n_{grid} denotes the number of grid points of possible y values (a fine grid over $\mathbb{R}$), used in the construction of the full conformal prediction method (15). The last column ("Model evaluation cost") counts the number of times we evaluate a fitted model $\widehat{\mu}$ on a single new data point. In most settings, the model training cost is dominant—for example, training a neural network is far more costly than evaluating the prediction of a trained network on a new example. There are important exceptions, however, such as K-nearest neighbors, where computing a prediction incurs the cost of identifying the K neighbors of the test point.

[3]It is worth mentioning that for several common regression algorithms, the n leave-one-out residuals can be obtained without refitting n times, but by simply reweighting the in-sample training residuals. Examples include linear smoothing methods like ordinary least squares, kernel ridge regression, kernel smoothing, thin plate splines and smoothing splines. Another interesting example is random forests, where the i-th leave-one-out fit can be obtained by simply ignoring all trees containing the i-th point.

5 Guarantees under stability assumptions

Next, we consider how adding stability assumptions—conditions that ensure that the fitted regression function $\widehat{\mu}$ is not too sensitive to perturbations of the training data set—can improve the theoretical guarantees of the jackknife and its variants. (For simplicity, we only consider leave-one-out methods, and do not examine K-fold cross-validation here.)

5.1 In-sample and out-of-sample stability

Fix any $\epsilon \geq 0$, $\nu \in [0,1]$, any sample size $n \geq 2$, and any distribution P on (X, Y). We say that a regression algorithm $\mathcal{A}$ satisfies (ϵ, ν) out-of-sample stability with respect to the distribution P and sample size n if, for all $i \in \{1, \ldots, n\}$,

$$\mathbb{P}\left\{|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})| \leq \epsilon\right\} \geq 1 - \nu, \tag{16}$$

for $\widehat{\mu}$ and $\widehat{\mu}_{-i}$ defined as before in (5) and (6). The probability above is taken with respect to the distribution of the data points $(X_1, Y_1), \ldots, (X_n, Y_n)$, (X_{n+1}, Y_{n+1}) drawn i.i.d. from P. Similarly, $\mathcal{A}$ satisfies (ϵ, ν) in-sample stability with respect to the distribution P and sample size n if, for all $i \in \{1, \ldots, n\}$,

$$\mathbb{P}\left\{|\widehat{\mu}(X_i) - \widehat{\mu}_{-i}(X_i)| \leq \epsilon\right\} \geq 1 - \nu. \tag{17}$$

Naturally, since the data points are exchangeable, if (16) or (17) holds for any single $i \in \{1, \ldots, n\}$ then it holds for all $i \in \{1, \ldots, n\}$. These types of conditions appear elsewhere in the literature—for example Bousquet and Elisseeff [2002] define similar conditions, termed "hypothesis stability" and "pointwise hypothesis stability".

While the out-of-sample and in-sample stability properties may at first appear similar, they are extremely different in practice. Out-of-sample stability requires that, for a test point that is *independent of the training data*, the predicted value does not change much if we remove one point from the training data. In contrast, in-sample stability requires that, *for a point in the training data set*, the predicted value does not change much if we remove this point itself from the training data set. In a scenario where the model fitting algorithm suffers from strong overfitting, we would expect in-sample stability to be very poor, while out-of-sample stability may still hold—for example, we will see in Section 5.5 that this is the case for K-nearest-neighbor methods. On the other hand, strongly convex regularization, such as ridge regression, induces both in- and out-of-sample stability [Bousquet and Elisseeff, 2002, Example 3]. This is not the case, however, for sparse regression methods (e.g., ℓ_1 regularization), which are proved by Xu et al. [2012] to be incompatible with in-sample stability.

5.2 Summary of stability results

Before giving the details of our theoretical results, we summarize our findings on the various methods' predictive coverage guarantees, with and without stability assumptions:

Method	Assumption-free theory	Out-of-sample stability	In-sample and out-of-sample stability
Naive (4)	No guarantee	No guarantee	$\approx 1-\alpha$
Jackknife (7)	No guarantee	$\approx 1-\alpha$	$\approx 1-\alpha$
Jackknife+ (9)	$1-2\alpha$	$\approx 1-\alpha$	$\approx 1-\alpha$
Jackknife-minmax (10)	$1-\alpha$	$1-\alpha$	$1-\alpha$

The assumption free results are the same as those discussed in Section 4, while the results under stability assumptions are presented next in Theorems 5 and 6.

5.3 Out-of-sample stability and the jackknife

We will next prove that out-of-sample stability is sufficient for the jackknife and jackknife+ methods to achieve the target coverage rate, with a slight modification. Define

$$\widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \Big(\widehat{q}^{+}_{n,\alpha}\big\{R^{\text{LOO}}_i\big\} + \epsilon\Big),$$

and similarly, define

$$\widehat{C}^{\text{jackknife+},\epsilon}_{n,\alpha}(X_{n+1}) = \Big[\widehat{q}^{-}_{n,\alpha}\big\{\widehat{\mu}_{-i}(X_{n+1}) - R^{\text{LOO}}_i\big\} - \epsilon,\ \widehat{q}^{+}_{n,\alpha}\big\{\widehat{\mu}_{-i}(X_{n+1}) + R^{\text{LOO}}_i\big\} + \epsilon\Big],$$

which we refer to as the ϵ-*inflated* versions of the jackknife and jackknife+ predictive intervals.

Theorem 5. *Suppose that the regression algorithm $\mathcal{A}$ satisfies the (ϵ, ν) out-of-sample stability property (16) with respect to the data distribution P and the sample size n. Then the ϵ-inflated jackknife prediction interval satisfies*

$$\mathbb{P}\Big\{Y_{n+1} \in \widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha}(X_{n+1})\Big\} \geq 1-\alpha-2\sqrt{\nu}.$$

Similarly, the 2ϵ-inflated jackknife+ prediction interval satisfies

$$\mathbb{P}\Big\{Y_{n+1} \in \widehat{C}^{\text{jackknife+},2\epsilon}_{n,\alpha}(X_{n+1})\Big\} \geq 1-\alpha-4\sqrt{\nu}.$$

(The different amounts of inflation, ϵ for jackknife versus 2ϵ for jackknife+, are simply an artifact of the particular definition of out-of-sample stability that we use, and should not be interpreted as a meaningful difference between these two methods.)

We remark that if we additionally assume that, in the data distribution, $Y|X$ has a bounded conditional density (for example, $Y = \mu(X) + \mathcal{N}(0, \sigma^2)$ for some unknown true mean function $\mu(\cdot)$), then the result of Theorem 5 is sufficient to ensure that the (non-inflated) jackknife and jackknife+ intervals achieve close to target coverage. The reason is this: if the conditional density of $Y|X$ is bounded by some constant $c < \infty$, then very little probability can be captured by inflating the interval. Specifically,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha}(X_{n+1}) \backslash \widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})\right\} \leq 2\epsilon c.$$

Combined with the result of Theorem 5, this proves that

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha - 2\sqrt{\nu} - 2\epsilon c.$$

Similarly, for jackknife+, we have

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha - 4\sqrt{\nu} - 4\epsilon c.$$

5.4 In-sample stability and overfitting

To contrast the scenarios of in-sample and out-of-sample stability, we will next demonstrate that adding the in-sample stability assumption would in fact be sufficient for the "naive" prediction interval, defined earlier in (4), to have coverage at roughly the target level. Its ϵ-inflated version is defined as

$$\widehat{C}^{\text{naive},\epsilon}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \left(\widehat{q}^{+}_{n,\alpha}\{|Y_i - \widehat{\mu}(X_i)|\} + \epsilon\right). \tag{18}$$

Recall from Section 1 that we would typically expect $\widehat{C}^{\text{naive}}_{n,\alpha}(X_{n+1})$ to undercover severely due to the overfitting problem (thus inspiring the use of the jackknife to avoid this issue), and similarly $\widehat{C}^{\text{naive},\epsilon}_{n,\alpha}(X_{n+1})$ will also undercover whenever ϵ is too small to correct for overfitting. This is often the case even when out-of-sample stability is satisfied. With in-sample stability, however, this is no longer the case—in other words, the in-sample stability property is essentially assuming that inflation by ϵ is sufficient to correct for overfitting.

Theorem 6. *Suppose that the regression algorithm $\mathcal{A}$ satisfies both the (ϵ, ν) in-sample stability property (17) and the (ϵ, ν) out-of-sample stability property (16) with respect to the data distribution P and the sample size n. Then the naive prediction interval satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{naive},2\epsilon}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha - 4\sqrt{\nu}.$$

5.5 Example: K-nearest-neighbors

To give an illustrative example, consider a K-nearest-neighbor (K-NN) method. This style of example is also considered in Steinberger and Leeb [2018, Example 4.1], and was studied earlier by Devroye and Wagner [1979] in the context of estimating the error of a classifier, and by Bousquet and Elisseeff [2002] in the context of error in regression. Given a training data set $(X_1, Y_1), \ldots, (X_n, Y_n)$ and a new test point x, our prediction is

$$\widehat{\mu}(x) = \frac{1}{K} \sum_{i \in N(x)} Y_i,$$

where $N(x) \subset \{1, \ldots, n\}$ is the set of the K nearest neighbors to the test point x, i.e., the K indices i giving the smallest values of $\|X_i - x\|_2$ (of course, we can replace the ℓ_2 norm with any other metric). We will assume for simplicity that there are no ties between these distances (for instance, the X_i points might be continuously distributed on $\mathbb{R}^d$, or we apply a random tie-breaking rule). Now consider out-of-sample stability. Let $N(X_{n+1})$ and $N_{-i}(X_{n+1})$ be the K-nearest-neighbor sets for the test point X_{n+1} given the full training data, or the training data with data point i removed, respectively. Then we can easily verify that

$$i \notin N(X_{n+1}) \;\Leftrightarrow\; N(X_{n+1}) = N_{-i}(X_{n+1}) \;\Rightarrow\; \widehat{\mu}(X_{n+1}) = \widehat{\mu}_{-i}(X_{n+1}).$$

Therefore,

$$\mathbb{P}\left\{|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})| = 0\right\} \geq \mathbb{P}\left\{i \notin N(X_{n+1})\right\} = 1 - \frac{K}{n},$$

where the last step holds by exchangeability of the n training points. This proves that the K-NN method satisfies (ϵ, ν)-out-of-sample stability with $\epsilon = 0$ and $\nu = K/n$. (In contrast, we cannot hope for a similar argument to guarantee in-sample stability, since we will always have $i \in N(X_i)$; that is, X_i is one of its own nearest neighbors—and so in general we will have $\widehat{\mu}(X_i) \neq \widehat{\mu}_{-i}(X_i)$.)

Applying the conclusion of Theorem 5 to this setting, then, we see that K-NN leads to a coverage rate at least

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha - 2\sqrt{K/n}$$

for the jackknife, and

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha - 4\sqrt{K/n}$$

for the jackknife+. These results hold with no assumptions whatsoever on the distribution of the data—and in particular, we do not need to assume that the K-NN prediction is accurate or consistent on the given data.

5.6 Comparison to existing results

As mentioned above, Bousquet and Elisseeff [2002] study stability in the context of generalization bounds for regression, with the aim of bounding risk rather than predictive inference. The predictive accuracy of the jackknife under assumptions of algorithm stability was explored by Steinberger and Leeb [2016] for the linear regression setting, and in a more general setting by Steinberger and Leeb [2018]. Their stability assumption (see, e.g., Steinberger and Leeb [2018, Definition 1]) is essentially equivalent to our out-of-sample stability condition (16). However, the theory obtained in their work is asymptotic, and relies also on distributional assumptions (see Steinberger and Leeb [2018, (C1)]), namely, that $Y_i = \mathbb{E}[Y_i \mid X_i] + \nu_i$ where the noise ν_i is continuously distributed and is independent of X_i (for example, this does not allow for heteroskedasticity). In contrast, our guarantee, in Theorem 5, offers a simple finite-sample coverage guarantee with no distributional assumptions, requiring only algorithm stability.

6 Proof of Theorem 1

Suppose for a moment that we have access to the test point (X_{n+1}, Y_{n+1}) as well as the training data. For any indices $i, j \in \{1, \ldots, n+1\}$ with $i \neq j$, let $\widetilde{\mu}_{-(i,j)}$ define the regression function fitted on the training plus test data, with points i and j removed. (We use $\widetilde{\mu}$ rather than $\widehat{\mu}$ to remind ourselves that the model is fitted on a subset of the combined training and test data $i = 1, \ldots, n+1$, rather than a subset of only the training data.) Note that $\widetilde{\mu}_{-(i,j)} = \widetilde{\mu}_{-(j,i)}$ for any $i \neq j$, and that $\widetilde{\mu}_{-(i,n+1)} = \widehat{\mu}_{-i}$ for any $i = 1, \ldots, n$.

Next, we define a matrix of residuals, $R \in \mathbb{R}^{(n+1)\times(n+1)}$, with entries

$$R_{ij} = \begin{cases} +\infty, & i = j, \\ \left|Y_i - \widetilde{\mu}_{-(i,j)}(X_i)\right|, & i \neq j, \end{cases}$$

i.e., the off-diagonal entries represent the residual for the ith point when both i and j are left out of the regression. We also define a comparison matrix, $A \in \{0,1\}^{(n+1)\times(n+1)}$, with entries

$$A_{ij} = \mathbb{1}\{R_{ij} > R_{ji}\}.$$

In other words, A_{ij} is the indicator for the event that, when excluding data points i and j from the regression, data point i has higher residual than data point j. Naturally we see that $A_{ij} = 1$ implies $A_{ji} = 0$, for any i, j. We note that this comparison matrix construction is also examined by Vovk [2015, Appendix A], where it is used to establish that leave-one-out conformal methods fail to achieve $1-\alpha$ coverage.

Next, we are interested in finding data points i with unusually large residuals—the ones that are hardest to predict. We will define a set $\mathcal{S}(A) \subseteq \{1, \ldots, n+1\}$ of "strange" points,[4]

$$\mathcal{S}(A) = \{i \in \{1, \ldots, n+1\} : A_{i\bullet} \geq (1-\alpha)(n+1)\},$$

where $A_{i\bullet} = \sum_{j=1}^{n+1} A_{ij}$ is the ith row sum of A. In other words, the ith point is "strange" (i.e., $i \in \mathcal{S}(A)$) if it holds that, when we compare the residual R_{ij} of the ith point against residual R_{ji} for the jth point (for each $j \neq i$), the residual R_{ij} for the ith point is the larger one, for a sufficiently high fraction of these comparisons.

From this point on, the proof will proceed as follows:

- Step 1: we will establish deterministically that $|\mathcal{S}(A)| \leq 2\alpha(n+1)$, that is, for any comparison matrix A it is impossible to have more than $2\alpha(n+1)$ many strange points.
- Step 2: using the fact that the data points are i.i.d. (or more generally exchangeable), we will show that the probability that the test point $n+1$ is strange (i.e., $n+1 \in \mathcal{S}(A)$) is therefore bounded by 2α.
- Step 3: finally, we will verify that the jackknife+ interval can only fail to cover the test response value Y_{n+1} if $n+1$ is a strange point.

Step 1: bounding the number of strange points This bound is essentially a consequence of Landau's theorem for tournaments [Landau, 1953]. For data points i and j, we say that data point i "wins" its game against data point j, if $A_{ij} = 1$; that is, point i has a higher residual than point j, under the corresponding regression $\widetilde{\mu}_{-(i,j)}$. Note that each strange point $i \in \mathcal{S}(A)$ can lose against at most $\alpha(n+1)-1$ other strange points—this is because point i must win against at least $(1-\alpha)(n+1)$ points in total since it is strange, and as we have defined it, point i cannot win against itself.

Let $s = |\mathcal{S}(A)|$ denote the number of strange points. The key realization is now that, if we think about grouping each *pair* of strange points by the losing point, then we see that there are at most

$$s \cdot \big(\alpha(n+1) - 1\big)$$

pairs of strange points. This is because there are at most s unique possibilities for the loser, and for each such loser, it can only lose against at most $\alpha(n+1)-1$ other strange points, as argued above. In other words, we have established

$$\frac{s(s-1)}{2} \leq s \cdot \big(\alpha(n+1) - 1\big),$$

and rearranging gives $s \leq 2\alpha(n+1) - 1 < 2\alpha(n+1)$, as desired.

[4]The authors thank an anonymous reviewer for suggesting this presentation of the proof.

Step 2: exchangeability of the data points We next leverage the exchangeability of the data points to show that, since there are at most $2\alpha(n+1)$ strange points among a total of $n+1$ points, it follows that the test point has at most 2α probability of being strange—this reasoning uses the exchangeability of the data in exactly the same way as the conformal prediction literature [Vovk et al., 2005].

To establish this formally, since the data points $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable and the regression fitting algorithm $\mathcal{A}$ is invariant to the ordering of the data points (the symmetry condition (8)), it follows that $A \stackrel{\mathrm{d}}{=} \Pi A \Pi^\top$ for any $(n+1)\times(n+1)$ permutation matrix Π, where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution. In particular, for any index $j \in \{1, \ldots, n+1\}$, suppose we take Π to be any permutation matrix with $\Pi_{j,n+1} = 1$ (i.e., corresponding to a permutation mapping $n+1$ to j). Then, deterministically, we have

$$n+1 \in \mathcal{S}(A) \;\Leftrightarrow\; j \in \mathcal{S}\big(\Pi A \Pi^\top\big),$$

and therefore,

$$\mathbb{P}\left\{n+1 \in \mathcal{S}(A)\right\} = \mathbb{P}\left\{j \in \mathcal{S}\big(\Pi A \Pi^\top\big)\right\} = \mathbb{P}\left\{j \in \mathcal{S}(A)\right\}$$

for all $j = 1, \ldots, n+1$. In other words, if we compare an arbitrary training point j versus the test point $n+1$, these two points are equally likely to be strange, by exchangeability of the data. We can then calculate

$$\mathbb{P}\left\{n+1 \in \mathcal{S}(A)\right\} = \frac{1}{n+1}\sum_{j=1}^{n+1} \mathbb{P}\left\{j \in \mathcal{S}(A)\right\} = \frac{\mathbb{E}\left[|\mathcal{S}(A)|\right]}{n+1} \leq 2\alpha,$$

where the last step applies the result of Step 1.

Step 3: connecting to jackknife+ Finally, we need to relate the question of coverage of the jackknife+ interval, to our notion of strange points. Suppose that $Y_{n+1} \notin \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})$. This means that either

$$Y_{n+1} > \widehat{q}^{+}_{n,\alpha}\big\{\widehat{\mu}_{-i}(X_{n+1}) + R^{\text{LOO}}_i\big\},$$

which implies that $Y_{n+1} > \widehat{\mu}_{-j}(X_{n+1}) + R^{\text{LOO}}_j$ for at least $(1-\alpha)(n+1)$ many indices $j \in \{1, \ldots, n\}$, or otherwise

$$Y_{n+1} < \widehat{q}^{-}_{n,\alpha}\big\{\widehat{\mu}_{-i}(X_{n+1}) - R^{\text{LOO}}_i\big\},$$

which implies that $Y_{n+1} < \widehat{\mu}_{-j}(X_{n+1}) - R_j^{\mathrm{LOO}}$ for at least $(1-\alpha)(n+1)$ many indices $j \in \{1, \ldots, n\}$. In either case, then, we have

$$\begin{aligned}(1-\alpha)(n+1) &\leq \sum_{j=1}^{n} \mathbb{1}\left\{Y_{n+1} \notin \widehat{\mu}_{-j}(X_{n+1}) \pm R_j^{\mathrm{LOO}}\right\} \\ &= \sum_{j=1}^{n} \mathbb{1}\left\{\left|Y_j - \widehat{\mu}_{-j}(X_j)\right| < \left|Y_{n+1} - \widehat{\mu}_{-j}(X_{n+1})\right|\right\} \\ &= \sum_{j=1}^{n+1} \mathbb{1}\left\{R_{j,n+1} < R_{n+1,j}\right\} = \sum_{j=1}^{n+1} A_{n+1,j},\end{aligned}$$

and therefore $n+1 \in \mathcal{S}(A)$, that is, point $n+1$ is strange. Combining this with the result of Step 2, we have

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_{n,\alpha}^{\text{jackknife+}}(X_{n+1})\right\} \leq \mathbb{P}\left\{n+1 \in \mathcal{S}(A)\right\} \leq 2\alpha.$$

7 Empirical results

In this section, we compare seven methods—naive (4), jackknife (7), jackknife+ (9), jackknife-minmax (10), CV+ (11), split conformal (3), and full conformal (15)—on simulated and real data. Code for reproducing all results and figures is available online.[5]

7.1 Simulations

We first examine the performance of the various prediction intervals on a simulated example, using least squares as our regression method. We will see that when the training sample size n is equal or approximately equal to the dimension d, the instability of the least squares method (due to poor conditioning of the $n \times d$ design matrix) leads to a wide disparity in performance between the various methods. This simulation is thus designed to demonstrate the role of stability in the performance of these various methods.

7.1.1 Data and methods

Our target coverage level is $1 - \alpha = 0.9$. We use training sample size $n = 100$, and repeat the experiment at each dimension $d = 5, 10, \ldots, 200$, with i.i.d. data points (X_i, Y_i) generated as

$$X_i \sim \mathcal{N}(0, I_d) \text{ and } Y_i \mid X_i \sim \mathcal{N}(X_i^\top \beta, 1).$$

[5] http://www.stat.uchicago.edu/~rina/jackknife.html

The true coefficient vector β is drawn as $\beta = \sqrt{10} \cdot u$ for a uniform random unit vector $u \in \mathbb{R}^d$. The regression method $\mathcal{A}$ is simply least squares, with the convention that if the linear system is underdetermined then we take the solution with the lowest ℓ_2 norm (the limit of ridge regression as the regularization tends to zero). Specifically, for training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we return the regression function $\widehat{\mu}(x) = x^\top \widehat{\beta}$,

$$\widehat{\beta} = X_{\text{mat}}^\dagger Y_{\text{vec}},$$

where X_{mat} denotes the $n \times d$ matrix of covariates, Y_{vec} the vector of responses, and $\dagger$ the Moore–Penrose pseudoinverse.

We then generate 100 test data points from the same distribution, and calculate the empirical probability of coverage (i.e., the proportion of test points for which the prediction interval computed at the X value contains the Y value) and the average width of the prediction interval.

7.1.2 Results

Figure 2 displays the results of the simulation, averaged over 50 trials (where each trial has an independent draw of the training sample of $n = 100$ and the test sample of size 100).

When $d < n$, the jackknife and jackknife+ show very similar performance, with approximately the right coverage level $1 - \alpha = 0.9$ and with nearly identical interval width. For $d \approx n$ (the regime where least squares is quite unstable), the jackknife has substantial undercoverage—at $d = n$ the jackknife shows coverage rate around 0.5, and continues to show substantial undercoverage when d is slightly larger than n. In this regime, the jackknife+ continues to show the right coverage level, at the cost of a prediction interval that is only slightly wider than the jackknife. For large d, the jackknife and jackknife+ again show very similar performance. In fact, this connects to recent work on interpolation methods (methods that achieve zero training error). Specifically, Hastie et al. [2019] study "ridgeless" regression (i.e., the least squares solution with the lowest ℓ_2 norm, as in our simulation), and demonstrate that this provides a stable solution with good test error as long as d is *either* sufficiently small or sufficiently large relative to n. We see a similar phenomenon in the predictive coverage performance of the jackknife.

As expected, the jackknife-minmax is over-conservative, with typical coverage higher than $1 - \alpha = 0.9$ across all dimensions d, while the naive method drastically undercovers due to increasing overfitting as d grows (and in fact, at $d \geq n$, the training error is exactly zero, so the prediction intervals have width zero and coverage zero.)

When $d > n$, we note that full conformal prediction will always have infinite length intervals since for every potential y in the (X_{n+1}, y) pair, all $n + 1$ residuals will equal zero. Naturally, in such a situation, full conformal will have coverage

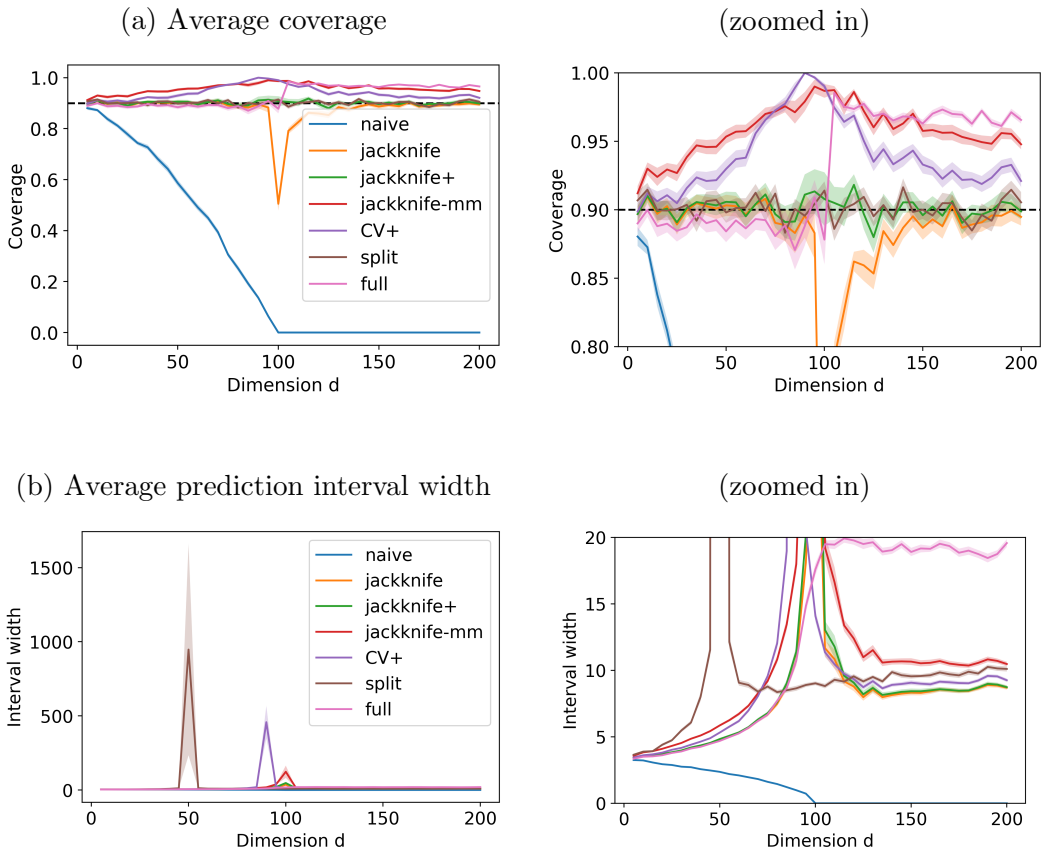


Figure 2: Simulation results, showing the coverage and width of the predictive interval for all methods. The solid lines show the mean over 50 independent trials, with shading to show $\pm$ one standard error. We observe that the jackknife undercovers around $d = 100$ due to instability (since $n = 100$). Jackknife+ and split conformal are the only two methods that maintain the correct coverage level throughout without under- or over-covering, but we can observe that jackknife+ often produces shorter intervals than split conformal. (See text for more details.)

equal to one deterministically. In practice, it is common to modify the conformal prediction method by truncating to a finite range, e.g., to the observed range of Y values in the training data (which has minimal effect on the coverage guarantee [Chen et al., 2018]); this is why we see finite length intervals for full conformal in our simulation results.

Split conformal is the only method other than jackknife+ to maintain coverage at 0.9 throughout. (Note that since split conformal trains on half of the data, its length spikes near $d = 50$, rather than $d = 100$ as for the other methods; this is simply due to the change in sample size $n/2 = 50$ used in training. This is a result of instability of OLS when $n \approx d$ and is not reflective of comparisons between holdout and jackknife+.)

7.2 Real data

We next compare the various methods on three real data sets. We will try three regression algorithms: ridge regression, random forests, and neural networks (details given below). Our aim in these experiments is to demonstrate the typical performance of the various prediction interval methods in a real data setting; we do not seek to optimize the base methods used as our regression algorithms, but are only interested in how the various prediction interval methods behave in comparison to each other. Due to the high computational cost of the full conformal method, we do not include it in the comparison.

7.2.1 Data

The Communities and Crime data set[6] [Redmond and Baveja, 2002] contains information on 1994 communities, with covariates such as median income, distribution of ages, family size, etc., and the goal of predicting a response variable defined as the per capita violent crime rate. After removing categorical variables and variables with missing data, $d = 99$ covariates remain.

The BlogFeedback data set[7] [Buza, 2014] contains 52397 data points, each corresponding to a single blog post. The goal is to predict the response variable of the number of comments left on the blog post in the following 24 hours, using $d = 280$ covariates such as the length of the post, the number of comments on previous posts, etc. Since the distribution of the response is extremely skewed, we transform it as $Y = \log(1 + \#\text{ comments})$.

The Medical Expenditure Panel Survey 2016 data set,[8] provided by the Agency for Healthcare Research and Quality, contains data on individuals' utilization of

[6] http://archive.ics.uci.edu/ml/datasets/communities+and+crime

[7] https://archive.ics.uci.edu/ml/datasets/BlogFeedback

[8] https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192

medical services such as visits to the doctor, hospital stays, etc. Details on the data collection for older versions of this data set are described in Ezzati-Rice et al. [2008]. We select a subset of relevant features, such as age, race/ethnicity, family income, occupation type, etc. After splitting categorical features into dummy variables to encode each category separately, the resulting dimension is $d = 107$. The goal is to predict the health care system utilization of each individual, which is a composite score reflecting the number of visits to a doctor's office, hospital visits, days in nursing home care, etc. With missing data removed, this data set contains 33005 data points. Since the distribution of the response is highly skewed, we transform it as $Y = \log(1 + (\text{utilization score}))$.

7.2.2 Methods

Our procedure is the same for each of the three data sets. We randomly sample $n = 200$ data points from the full data set, to use as the training data. The remaining points form the test set.

We run our experiment using three different regression algorithms $\mathcal{A}$—namely, ridge regression, random forests, and neural networks. The details of these algorithms is as follows:

- For ridge regression, we define $\widehat{\mu}(x) = \widehat{\beta}_0 + x^\top \widehat{\beta}$ for

$$\widehat{\beta}_0, \widehat{\beta} = \arg\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^d} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - \beta_0 - X_i^\top \beta)^2 + \lambda \|\beta\|_2^2 \right\},$$

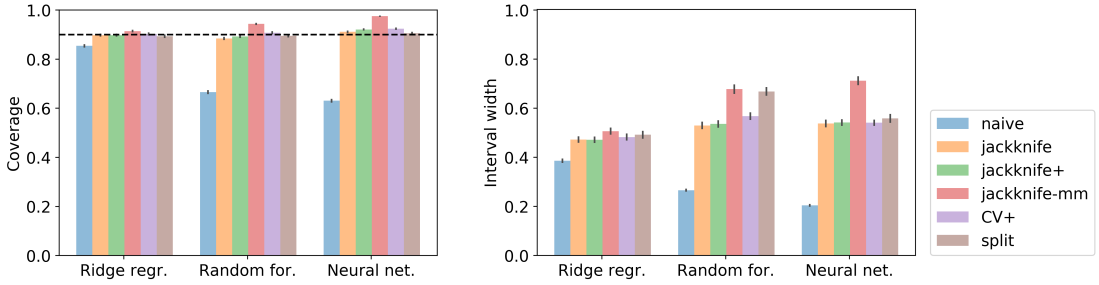
  where the penalty parameter is chosen as $\lambda = 0.001 \|X_{\text{mat}}\|^2$, where $X_{\text{mat}} \in \mathbb{R}^{n \times d}$ is the covariate matrix of the training data, and $\|X_{\text{mat}}\|$ is its spectral norm. This choice is to accommodate situations in which the matrix X_{mat} does not have full column rank as in the case where $d > n$. In such cases, the solution above is nearly the least-squares solution with minimum ℓ_2 norm.

- For random forests, we use the `RandomForestRegressor` method from the `scikit-learn` package [Pedregosa et al., 2011] in Python, with 20 trees grown for each random forest using the mean absolute error criterion, and with default settings otherwise.

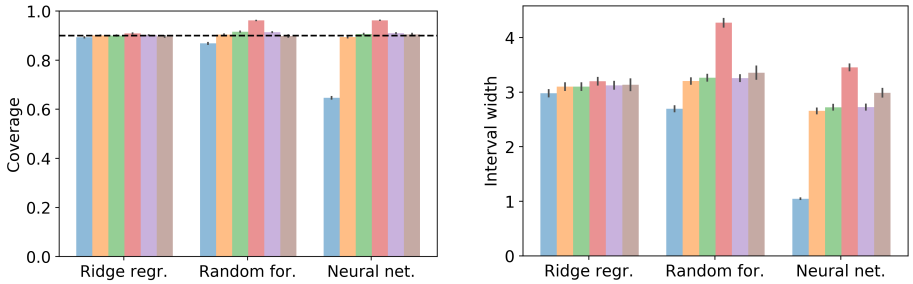
- For neural networks, we use the `MLPRegressor` method also from `scikit-learn`, run with the L-BFGS solver and the logistic activation function, and with default settings otherwise.

For each choice of $\mathcal{A}$, we construct six prediction intervals (naive, jackknife, jackknife+, jackknife-minmax, CV+, split conformal), and calculate their empirical coverage rate and their average width on the test set. We then repeat this procedure 20 times, with the train/test split formed randomly each time, and report the mean and standard error over these 20 trials.

(a) Communities and crime data set



(b) BlogFeedback data set



(c) Medical Expenditure Panel Survey data set

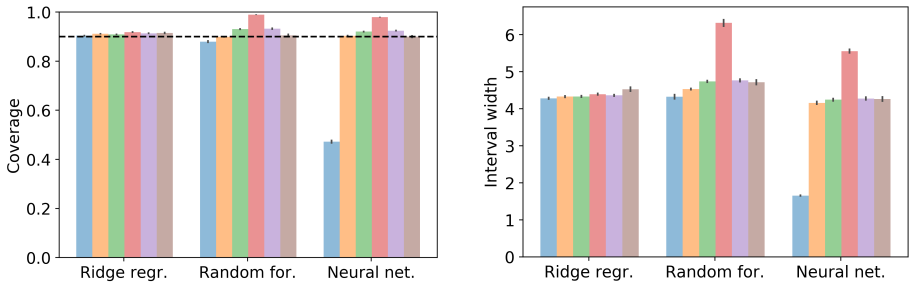


Figure 3: Results on three real data sets, using either ridge regression, random forests, or neural networks as the regression algorithm. The bar plots show the coverage and the width of the predictive interval for all methods. The figures display the mean over 20 independent trials (i.e., splits into training and test data), with error bars to show $\pm$ one standard error. In general, the naive method undercovers while jackknife-minmax overcovers, and the remaining methods have well calibrated coverage. In terms of their interval lengths, we typically (but not necessarily) get the expected order: jackknife $<$ jackknife+ $<$ 10-fold CV+ $<$ split conformal.

7.2.3 Results

Figure 3 displays the results of the real data experiments. For each data set, each regression algorithm, and each one of the six prediction interval methods, the figure plots the average coverage and average width, together with their standard errors across the 20 independent trials.

We see that the jackknife and jackknife+ methods both yield empirical coverage extremely close to the target level of 90%, and have very similar predictive interval widths. However, in some settings, the jackknife+ shows slightly higher coverage than jackknife, and slightly wider prediction intervals. These settings correspond to regression methods with greater instability. As expected, the naive method undercovers in some settings and the jackknife-minmax is generally overly conservative. Split conformal performs reasonably well: its length and coverage is sometimes comparable to the jackknife+, but is also significantly wider in some instances. Intuitively, if the best regression function in the considered function class is simple and the dataset is large, split conformal should perform fine even though it uses $n/2$ points for training and $n/2$ for calibration; however, in settings where the dataset is small relative to the complexity of the best regressor, then we should observe significant gains in using $n-1$ points for training and n points for calibration. One phenomenon that is not visible in the empirical results is that split conformal is a randomized method, with output varying slightly depending on the random split, while jackknife+ is a deterministic method on any fixed training data set.

8 Summary

The jackknife+ differs from the jackknife in that it uses the quantiles of

$$\widehat{\mu}_{-i}(X_{n+1}) \pm R_i^{\mathrm{LOO}} = \widehat{\mu}(X_{n+1}) + \Big(\widehat{\mu}_{-i}(X_{n+1}) - \widehat{\mu}(X_{n+1})\Big) \pm R_i^{\mathrm{LOO}},$$

instead of those of $\widehat{\mu}(X_{n+1}) \pm R_i^{\mathrm{LOO}}$, to build predictive intervals. By applying the shifts $\widehat{\mu}_{-i}(X_{n+1}) - \widehat{\mu}(X_{n+1})$, the jackknife+ effectively accounts for the (possible) algorithm instability, yielding rigorous coverage guarantees under no assumptions other than exchangeable samples. This, together with its empirical performance on real data, makes it a better choice than the jackknife in practice. In cases where the jackknife+ is computationally prohibitive, K-fold CV+ offers an attractive alternative. Here, it would be interesting to see if the coverage guarantees for the latter method can be somewhat sharpened.

Acknowledgements

The authors are grateful to the American Institute of Mathematics for supporting and hosting our collaboration. R.F.B. was partially supported by the National

Science Foundation via grant DMS–1654076 and by an Alfred P. Sloan fellowship. E.J.C. was partially supported by the Office of Naval Research under grant N00014-16-1-2712, by the National Science Foundation via grant DMS–1712800, and by a generous gift from TwoSigma. The authors are grateful to an anonymous reviewer for helpful suggestions on the presentation of the proof of Theorem 1. E.J.C. thanks Yaniv Romano for help with some experiments. A.R. thanks Arun Kumar Kuchibhotla for discussions regarding cross-conformal prediction.

References

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

Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. In *Conference on Learning Theory*, pages 605–622, 2014.

Ronald Butler and Edward D Rothman. Predictive intervals based on reuse of the sample. *Journal of the American Statistical Association*, 75(372):881–889, 1980.

Krisztian Buza. Feedback prediction for blogs. In *Data analysis, machine learning and knowledge discovery*, pages 145–152. Springer, 2014.

Wenyu Chen, Kelli-Jean Chun, and Rina Foygel Barber. Discretized conformal prediction for efficient distribution-free inference. *Stat*, 7(1):e173, 2018.

Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7 (1):1–26, 1979.

Bradley Efron and Gail Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

Trena M Ezzati-Rice, Frederick Rohde, and Janet Greenblatt. Sample design of the medical expenditure panel survey household component, 1998-2007. *Agency for Healthcare Research and Quality*, MEPS Methodology Report No. 22, 2008.

Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

HG Landau. On dominance relations and the structure of animal societies: Iii the condition for a score structure. *Bulletin of Mathematical Biology*, 15(2):143–148, 1953.

Jing Lei. Fast exact conformalization of lasso using piecewise linear homotopy. *arXiv preprint arXiv:1708.00427*, 2017.

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.

Rupert G Miller. The jackknife–a review. *Biometrika*, 61(1):1–15, 1974.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. InTech, 2008.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Maurice H Quenouille. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):68–84, 1949.

Maurice H Quenouille. Notes on bias in estimation. *Biometrika*, 43(3/4):353–360, 1956.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

Lukas Steinberger and Hannes Leeb. Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*, 2016.

Lukas Steinberger and Hannes Leeb. Conditional predictive inference for high-dimensional stable algorithms. *arXiv preprint arXiv:1809.01412*, 2018.

Robert A Stine. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031, 1985.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

John Tukey. Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29:614, 1958.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490, 2012.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):9–28, 2015.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *arXiv preprint arXiv:1212.4966*, 2012.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pages 37–51, 2018.

Huan Xu, Constantine Caramanis, and Shie Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):187–193, 2012.

A Asymmetric jackknife+ and CV+

In settings where the distribution of Y given X appears to have symmetric noise, it is natural to construct predictions intervals symmetrically, which is why we can consider absolute values of residuals. If the data is likely to be skewed, however, we may want to consider an asymmetric construction. Fix any $\alpha_+, \alpha_- > 0$ with $\alpha_+ + \alpha_- = \alpha$, and let

$$\widehat{C}^{\text{jackknife+}}_{n,\alpha_\pm}(X_{n+1}) = \left[\widehat{q}^{-}_{n,\alpha_-}\left\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{\text{sgn,LOO}}\right\},\ \widehat{q}^{+}_{n,\alpha_+}\left\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{\text{sgn,LOO}}\right\}\right], \quad (19)$$

where the signed residuals are

$$R_i^{\text{sgn,LOO}} = Y_i - \widehat{\mu}_{-i}(X_i).$$

We can of course consider the analogous asymmetric version of the original jackknife,

$$\widehat{C}^{\text{jackknife}}_{n,\alpha_\pm}(X_{n+1}) = \left[\widehat{\mu}(X_{n+1}) + \widehat{q}^{-}_{n,\alpha_-}\left\{R_i^{\text{sgn,LOO}}\right\},\ \widehat{\mu}(X_{n+1}) + \widehat{q}^{+}_{n,\alpha_+}\left\{R_i^{\text{sgn,LOO}}\right\}\right]. \quad (20)$$

This type of asymmetric jackknife was considered by Steinberger and Leeb [2018]. Similarly we can define an asymmetric version of jackknife-minmax or of CV+.

We remark that, even if we were to choose $\alpha_- = \alpha_+ = \alpha/2$, these asymmetric constructions would not necessarily be equal to the original jackknife, jackknife+, jackknife-minmax, and CV+ intervals, because the empirical distribution of the signed residuals will in general be asymmetric even if only due to random chance.

All of the coverage guarantees that we have proved for the various symmetric methods, hold also for their asymmetric counterparts. For example, to verify $1-2\alpha$ coverage for the asymmetric jackknife+ in the assumption-free setting, the proof of Theorem 1 proceeds identically except that the matrix of residuals $R \in \mathbb{R}^{(n+1)\times(n+1)}$ constructed in the proof is replaced with two matrices

$$(R_\pm)_{ij} = \begin{cases} +\infty, & i = j, \\ \pm\left(Y_i - \widetilde{\mu}_{-(i,j)}(X_i)\right), & i \neq j, \end{cases}$$

where R_+ (resp. R_-) is used to bound the probability of noncoverage in the right (resp. left) tail by α_+ (resp. α_-).

B Additional proofs

B.1 Proof of jackknife-minmax (Theorem 3)

The proof for jackknife-minmax proceeds nearly identically to the proof for jackknife+ (Theorem 1). We define the residuals R_{ij} exactly as in the proof of Theorem 1, but we will use a different definition for the matrix A:

$$A_{ij} = \mathbb{1}\left\{\min_{j'} R_{ij'} > R_{ji}\right\},$$

that is, the *smallest* residual for data point i (leaving out any point j') is larger than R_{ji}, which is data point j's residual when leaving out point i. Define $\mathcal{S}(A)$ exactly as before. Now we follow essentially the same three steps as in the proof of Theorem 1:

- Step 1: we will establish deterministically that, for this new definition of the matrix A, we have $|\mathcal{S}(A)| \leq \alpha(n+1)$ (whereas, for jackknife+, the bound was $2\alpha(n+1)$).
- Step 2: using the fact that the data points are i.i.d. (or more generally exchangeable), we will show that the probability that the test point $n+1$ is strange (i.e., $n+1 \in \mathcal{S}(A)$) is therefore bounded by α.
- Step 3: finally, we will verify that the jackknife-minmax interval can only fail to cover the test response value Y_{n+1} if $n+1$ is a strange point.

Step 1: bounding the number of strange points To prove Step 1, let

$$i_\star \in \arg\min_{i \in \mathcal{S}(A)} \min_{j'} R_{i,j'}.$$

Then by definition, for all $j \in \mathcal{S}(A)$, $R_{ji_\star} \geq \min_{j'} R_{jj'} \geq \min_{j'} R_{i_\star j'}$. This means that, by definition of the new comparison matrix A, we have $A_{i_\star j} = 0$ for all $j \in \mathcal{S}(A)$, and therefore,

$$n + 1 - |\mathcal{S}(A)| \geq \sum_{j=1}^{n+1} A_{i_\star j} \geq (1-\alpha)(n+1),$$

where the last step holds by definition of $\mathcal{S}(A)$ (since $i_\star \in \mathcal{S}(A)$ is a strange point). Therefore $|\mathcal{S}(A)| \leq \alpha(n+1)$ as desired.

Step 2: exchangeability of the data points This step is identical to Step 2 in the proof of Theorem 1.

Step 3: connecting to jackknife-minmax Suppose that $Y_{n+1} \notin \widehat{C}^{\text{jack-mm}}_{n,\alpha}(X_{n+1})$. This means that either

$$Y_{n+1} > \max_{i=1,\ldots,n} \widehat{\mu}_{-i}(X_{n+1}) + \widehat{q}^{+}_{n,\alpha}\{R^{\text{LOO}}_j\},$$

which implies that $Y_{n+1} > \max_{i=1,\ldots,n} \widehat{\mu}_{-i}(X_{n+1}) + R^{\text{LOO}}_j$ for at least $(1-\alpha)(n+1)$ many indices $j \in \{1,\ldots,n\}$, or otherwise

$$Y_{n+1} < \min_{i=1,\ldots,n} \widehat{\mu}_{-i}(X_{n+1}) - \widehat{q}^{+}_{n,\alpha}\{R^{\text{LOO}}_j\},$$

which implies that $Y_{n+1} < \min_{i=1,\ldots,n} \widehat{\mu}_{-i}(X_{n+1}) - R^{\text{LOO}}_j$ for at least $(1-\alpha)(n+1)$ many indices $j \in \{1,\ldots,n\}$. In either case, then, we have

$$\begin{aligned}
(1-\alpha)(n+1) &\leq \sum_{j=1}^{n} \mathbb{1}\left\{Y_{n+1} \notin \big[\min_{i=1,\ldots,n} \widehat{\mu}_{-i}(X_{n+1}) - R^{\text{LOO}}_j, \max_{i=1,\ldots,n} \widehat{\mu}_{-i}(X_{n+1}) + R^{\text{LOO}}_j\big]\right\} \\
&= \sum_{j=1}^{n} \mathbb{1}\left\{\min_{i=1,\ldots,n} \big|Y_{n+1} - \widehat{\mu}_{-i}(X_{n+1})\big| > \big|Y_j - \widehat{\mu}_{-j}(X_j)\big|\right\} \\
&= \sum_{j=1}^{n+1} \mathbb{1}\left\{\min_{j'} R_{n+1,j'} > R_{j,n+1}\right\} \\
&= \sum_{j=1}^{n+1} A_{n+1,j'},
\end{aligned}$$

and therefore $n+1 \in \mathcal{S}(A)$.

B.2 Proofs for the CV+ method

In this section, we will give details for how the CV+ method relates to the cross-conformal prediction method, and then prove our theoretical guarantees for CV+.

B.2.1 Details for comparing to the cross-conformal method

In Section 3.2, we introduced the cross-conformal method (12) of Vovk [2015], Vovk et al. [2018], and stated two properties—first, that the cross-conformal prediction set is always contained in the CV+ prediction interval (13), and second, that the results of Vovk et al. [2018] imply a coverage guarantee (14) for the cross-conformal method. Here we give details to justify these two statements.

First, we verify that the cross-conformal prediction set $\widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1})$ is contained in the CV+ interval. To see this, suppose that $y \in \widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1})$. Then by definition of this predictive set, we have

$$\tau + \sum_{i=1}^{n} \mathbb{1}\left\{\left|y - \widehat{\mu}_{-S_{k(i)}}(X_{n+1})\right| < R_i^{\text{CV}}\right\} + \tau \mathbb{1}\left\{\left|y - \widehat{\mu}_{-S_{k(i)}}(X_{n+1})\right| = R_i^{\text{CV}}\right\} > \alpha(n+1)$$

for some $\tau \in [0,1]$. Since the left-hand side is monotone in τ, in particular this implies that the above inequality holds at $\tau = 1$, and so

$$\sum_{i=1}^{n} \mathbb{1}\left\{\left|y - \widehat{\mu}_{-S_{k(i)}}(X_{n+1})\right| \leq R_i^{\text{CV}}\right\} > \alpha(n+1) - 1.$$

Therefore,

$$\sum_{i=1}^{n} \mathbb{1}\left\{y > \widehat{\mu}_{-S_{k(i)}}(X_{n+1}) + R_i^{\text{CV}}\right\} < (1-\alpha)(n+1),$$

meaning that y is *not* larger than the $\lceil (1-\alpha)(n+1) \rceil$-th smallest value of $\widehat{\mu}_{-S_{k(i)}}(X_{n+1}) + R_i^{\text{CV}}$, $i = 1, \ldots, n$. In other words,

$$y \leq \widehat{q}^{+}_{n,\alpha}\left\{\widehat{\mu}_{-S_{k(i)}}(X_{n+1}) + R_i^{\text{CV}}\right\}.$$

An identical argument proves that

$$y \geq \widehat{q}^{-}_{n,\alpha}\left\{\widehat{\mu}_{-S_{k(i)}}(X_{n+1}) - R_i^{\text{CV}}\right\},$$

meaning that y must lie in the CV+ prediction interval $\widehat{C}^{\text{CV+}}_{n,K,\alpha}(X_{n+1})$. This completes our verification of the claim (13) that the CV+ interval contains the cross-conformal prediction set deterministically.

Next we give details for the coverage guarantee for the cross-conformal method, which is implied but not stated explicitly by Vovk et al. [2018]. Specifically, Vovk

et al. [2018] show that a modification of the K-fold cross-conformal method can lead to a $1-2\alpha$ coverage guarantee. To define the modified method, let

$$P_k(y) = \frac{\tau + \sum_{i \in S_k} \mathbb{1}\left\{\left|y - \widehat{\mu}_{-S_k}(X_{n+1})\right| < R_i^{\mathrm{CV}}\right\} + \tau \mathbb{1}\left\{\left|y - \widehat{\mu}_{-S_k}(X_{n+1})\right| = R_i^{\mathrm{CV}}\right\}}{m+1},$$

where $m = n/K$ is the number of data points in each fold (assumed to be an integer). Plugging in the true test value Y_{n+1}, we see that $P_k(Y_{n+1})$ is a rank-based p-value comparing the test residual $\left|Y_{n+1} - \widehat{\mu}_{-S_k}(X_{n+1})\right|$ of the test point against all other residuals in the kth fold. (Here $\tau \sim \mathrm{Unif}[0,1]$ corrects for discretization, so that this p-value is uniformly distributed on $[0,1]$ instead of on a discrete grid.) Vovk et al. [2018] then consider a modified cross-conformal method,

$$\widehat{C}^{\text{modified-cc}}_{n,K,\alpha}(X_{n+1}) = \left\{ y \in \mathbb{R} : \frac{1}{K} \sum_{k=1}^{K} P_k(y) > \alpha \right\}. \tag{21}$$

Vovk et al. [2018] cite earlier work by Vovk and Wang [2012, Corollary 2], which proves that an arithmetic means of p-values is itself a valid p-value up to a factor of 2—that is,

$$\mathbb{P}\left\{ \frac{1}{K} \sum_{k=1}^{K} P_k(Y_{n+1}) \leq \alpha \right\} \leq 2\alpha$$

for any $\alpha \in [0,1]$, and so the modified cross-conformal method at level α has predictive coverage at least $1-2\alpha$.

Now we relate the modified cross-conformal method to its original version. Plugging in the definition of the p-values $P_k(y)$ and comparing with the original (unmodified) cross-conformal predictive set (12), we can see that, deterministically,

$$\widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1}) \supseteq \widehat{C}^{\text{modified-cc}}_{n,K,\alpha'}(X_{n+1}) \text{ where } \alpha' = \alpha + (1-\alpha)\frac{K-1}{n+K}.$$

Since the modified cross-conformal method, run at level α', has coverage at least $1-2\alpha'$, this proves that

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1}) \right\} \geq 1 - 2\alpha' \geq 1 - 2\alpha - 2(1-\alpha) \cdot \frac{1-1/K}{n/K+1}.$$

B.2.2 Proof of Theorem 4

This proof follows essentially the same steps as the proof of Theorem 1. Suppose that we draw $n/K - 1$ additional test points, so that in total we have $m = n/K$ many test points, $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$. After partitioning the training data into sets $S_1, \ldots, S_K$ of size m, we define $S_{K+1} = \{n+1, \ldots, n+m\}$, the set of test points.

For any $k, k' \in \{1, \ldots, K+1\}$ with $k \neq k'$, let $\widetilde{\mu}_{-(S_k, S_{k'})}$ define the regression function fitted on the training plus test data, with subsets S_k and $S_{k'}$ removed. Next, we define a matrix of residuals, $R \in \mathbb{R}^{(n+m)\times(n+m)}$ with entries

$$R_{ij} = \begin{cases} +\infty, & k(i) = k(j), \\ \left|Y_i - \widetilde{\mu}_{-(S_{k(i)}, S_{k(j)})}(X_i)\right|, & k(i) \neq k(j). \end{cases}$$

Define a comparison matrix $A \in \{0,1\}^{(n+m)\times(n+m)}$ with

$$A_{ij} = \mathbb{1}\left\{R_{ij} > R_{ji}\right\},$$

and consider the set of "strange" points,

$$\mathcal{S}(A) = \left\{i \in \{1, \ldots, n+m\} : A_{i\bullet} \geq (1-\alpha)(n+1)\right\},$$

where $A_{i\bullet} = \sum_{j=1}^{n+m} A_{ij}$. Now we proceed as for the proof of Theorem 1.

- First, for Step 1, we bound the number of strange points deterministically as $|\mathcal{S}(A)| \leq 2\alpha(n+m) + (1-2\alpha)(m-1) - 1$.
- For Step 2 we see that the exchangeability of the data points implies that the probability that the test point $n+1$ is strange (i.e., $n+1 \in \mathcal{S}(A)$) is therefore bounded by $2\alpha + \frac{1-K/n}{K+1}$.
- For Step 3, we see that noncoverage of the CV+ interval implies that the test point is strange, which completes the proof of the theorem.

For Step 3, the proof that $Y_{n+1} \notin \widehat{C}^{\text{CV+}}_{n,K,\alpha}(X_{n+1})$ implies that $n+1 \in \mathcal{S}(A)$, is identical to the corresponding step in Theorem 1. We remark that in fact, our argument verifies a strictly stronger statement: if Y_{n+1} is not contained in the K-fold cross-conformal prediction set (with $n = K$, in the case of jackknife+), then the test point is strange—this is strictly stronger because the cross-conformal prediction set satisfies $\widehat{C}^{\text{cross-conf}}_{n,K,\alpha}(X_{n+1}) \subseteq \widehat{C}^{\text{CV+}}_{n,K,\alpha}(X_{n+1})$ always.

Turning to Step 1, we now bound the number of strange points. This step is very similar to the proof of Step 1 in the proof of Theorem 1. Let $s_k = |S_k \cap \mathcal{S}(A)|$ be the number of strange points in the kth fold, so that $s_1 + \cdots + s_{K+1} = s := |\mathcal{S}(A)|$. As before, each strange point $i \in \mathcal{S}(A)$ can lose against at most $\alpha(n+1) - 1$ other strange points. However, the difference relative to jackknife+ is that the data points i and j in the same fold ($k(i) = k(j)$) do not play against each other in the "tournament" so there are two types of pairs of strange points: those that are in different folds (writing $s = |\mathcal{S}(A)|$, there are at most $s \cdot \big(\alpha(n+1) - 1\big)$ such pairs, as in the jackknife+ proof), and those that are in the same fold (and therefore do not play a game). Thus we have established that

$$\frac{s(s-1)}{2} \leq s \cdot \big(\alpha(n+1) - 1\big) + \sum_k \frac{s_k(s_k-1)}{2}.$$

We can simplify the last sum as

$$\sum_k \frac{s_k(s_k-1)}{2} = \sum_k \frac{s_k^2 - s_k}{2} \leq \sum_k \frac{ms_k - s_k}{2} = \frac{m-1}{2}\sum_k s_k = \frac{s(m-1)}{2},$$

since $s_k \leq |S_k| = n/K = m$ for each k. Simplifying the expression above we have therefore proved that

$$s \leq 2\alpha(n+1) + m - 2 = 2\alpha(n+m) + (1-2\alpha)(m-1) - 1,$$

as desired.

Finally, we verify Step 2. For the K-fold setting, this is a bit more subtle than for jackknife+. This is because we cannot claim that $A \stackrel{\mathrm{d}}{=} \Pi A \Pi^\top$ for any $(n+m)\times(n+m)$ permutation matrix Π—indices i, j belonging in the same fold $(k(i) = k(j))$ behave differently than indices i, j in different folds $(k(i) \neq k(j))$. However, treating the split of the training and test data into folds $S_1, \ldots, S_K, S_{K+1}$ as fixed, we can verify that $A \stackrel{\mathrm{d}}{=} \Pi A \Pi^\top$ for any $(n+m)\times(n+m)$ permutation matrix Π that preserves the equivalence relation induced by the folds, $i \sim j$ if $k(i) = k(j)$. Now, for any $j \in \{1, \ldots, n+m\}$, there exists such a Π with $\Pi_{j,n+1} = 1$. Thus, as in the proof of Theorem 1, this implies that $\mathbb{P}\{n+1 \in \mathcal{I}(A)\} = \mathbb{P}\{j \in \mathcal{I}(A)\}$ for all $j \in \{1, \ldots, n+m\}$. Therefore, combining with the result of Step 1,

$$\mathbb{P}\{n+1 \in \mathcal{S}(A)\} \leq \frac{2\alpha(n+m) + (1-2\alpha)(m-1) - 1}{n+m} \leq 2\alpha + \frac{1 - K/n}{K+1}.$$

Combining with Step 3, we have completed the proof of the coverage guarantee.

B.3 Proof of Theorem 5

We will first consider an oracle leave-one-out method that, while impossible to implement in practice, achieves the target $1-\alpha$ coverage rate. We will then relate the ϵ-inflated jackknife and 2ϵ-inflated jackknife+ to the oracle method.

B.3.1 Oracle method

For each $i = 1, \ldots, n+1$, let $\widetilde{\mu}_{-i}$ be the regression function fitted on the training and test data with point i removed, i.e.

$$\widetilde{\mu}_{-i} = \mathcal{A}\Big((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_{n+1}, Y_{n+1})\Big).$$

Note that $\widetilde{\mu}_{-(n+1)} = \widehat{\mu}$, while for $i = 1, \ldots, n$, $\widetilde{\mu}_{-i}$ differs from $\widehat{\mu}_{-i}$ since the test point (X_{n+1}, Y_{n+1}) is included in the regression. Consider an "oracle jackknife" method, where we use $\widetilde{\mu}_{-i}$ in place of $\widehat{\mu}_{-i}$:

$$\widehat{C}^{\mathrm{oracle}}_{n,\alpha'}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}^{+}_{n,\alpha'}\big\{R^{\mathrm{oracle}}_i\big\},$$

where $\alpha' = \alpha + \sqrt{\nu}$ and

$$R_i^{\text{oracle}} = \big|Y_i - \widetilde{\mu}_{-i}(X_i)\big|.$$

Next, we will confirm that this oracle method achieves $1-\alpha'$ coverage. This fact is based on the exchangeability of the training and test data points, and can be proved using standard techniques from the conformal prediction literature (see, e.g., Vovk et al. [2005], Lei et al. [2018] for background). Writing

$$R_{n+1}^{\text{oracle}} = \big|Y_{n+1} - \widetilde{\mu}_{-(n+1)}(X_{n+1})\big| = \big|Y_{n+1} - \widehat{\mu}(X_{n+1})\big|,$$

we see that failure to cover, i.e., $Y_{n+1} \notin \widehat{C}_{n,\alpha'}^{\text{oracle}}$, occurs if and only if

$$R_{n+1}^{\text{oracle}} > \text{the } \lceil (1-\alpha')(n+1) \rceil\text{-th smallest value of } R_1^{\text{oracle}}, \ldots, R_n^{\text{oracle}}. \tag{22}$$

Now, since the training and test data are i.i.d., and the algorithm $\mathcal{A}$ is assumed to be invariant to the labeling of the points (8), this means that the resulting oracle residuals $R_1^{\text{oracle}}, \ldots, R_n^{\text{oracle}}, R_{n+1}^{\text{oracle}}$ are exchangeable. In other words, the rank of R_{n+1}^{oracle} among this list of oracle residuals is uniformly random. Therefore, the probability that the event in (22) occurs is equal to the probability that R_{n+1}^{oracle} is not one of the smallest $\lceil (1-\alpha')(n+1) \rceil$ values in the list of oracle residuals, and so

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_{n,\alpha'}^{\text{oracle}}(X_{n+1})\right\} \leq 1 - \frac{\lceil (1-\alpha')(n+1) \rceil}{n+1} \leq \alpha'. \tag{23}$$

(The first inequality cannot be replaced with an equality, due to the possibility of ties among the residuals.)

B.3.2 Bound for the jackknife

Now we relate the oracle method back to the jackknife. We will show that

$$\widehat{C}_{n,\alpha}^{\text{jackknife},\epsilon}(X_{n+1}) \supseteq \widehat{C}_{n,\alpha'}^{\text{oracle}}(X_{n+1})$$

with sufficiently high probability. To see why, suppose instead that this set inclusion does not hold. Then by definition of $\widehat{C}_{n,\alpha}^{\text{jackknife},\epsilon}(X_{n+1})$, we must have

$$\widehat{q}_{n,\alpha}^{+}\big\{R_i^{\text{LOO}}\big\} + \epsilon < \widehat{q}_{n,\alpha'}^{+}\big\{R_i^{\text{oracle}}\big\}.$$

By definition of these quantiles, we can conclude that the number of indices $i \in \{1, \ldots, n\}$ with $R_i^{\text{oracle}} > R_i^{\text{LOO}} + \epsilon$ is at least

$$\lceil (1-\alpha)(n+1) \rceil - \Big(\lceil (1-\alpha')(n+1) \rceil - 1\Big) \geq \sqrt{\nu}(n+1).$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha}(X_{n+1}) \not\supseteq \widehat{C}^{\text{oracle}}_{n,\alpha'}(X_{n+1})\right\} \\
&\qquad \leq \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{1}\left\{R_i^{\text{oracle}} > R_i^{\text{LOO}} + \epsilon\right\} \geq \sqrt{\nu}(n+1)\right\} \\
&\qquad \leq \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{1}\left\{\left|\widetilde{\mu}_{-i}(X_i) - \widehat{\mu}_{-i}(X_i)\right| > \epsilon\right\} \geq \sqrt{\nu}(n+1)\right\} \\
&\qquad \leq \frac{\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\left\{\left|\widetilde{\mu}_{-i}(X_i) - \widehat{\mu}_{-i}(X_i)\right| > \epsilon\right\}\right]}{\sqrt{\nu}(n+1)}, \quad (24)
\end{aligned}$$

where the last step holds by Markov's inequality. Observe also that, for every $i = 1, \ldots, n$,

$$\begin{aligned}
\mathbb{P}\left\{\left|\widetilde{\mu}_{-i}(X_i) - \widehat{\mu}_{-i}(X_i)\right| > \epsilon\right\} &= \mathbb{P}\left\{\left|\widetilde{\mu}_{-i}(X_i) - \widetilde{\mu}_{-(i,n+1)}(X_i)\right| > \epsilon\right\} \\
&= \mathbb{P}\left\{\left|\widetilde{\mu}_{-(n+1)}(X_{n+1}) - \widetilde{\mu}_{-(n+1,i)}(X_{n+1})\right| > \epsilon\right\} \\
&= \mathbb{P}\left\{\left|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})\right| > \epsilon\right\} \\
&\leq \nu,
\end{aligned}$$

where the first and third step hold by definition of the various regression functions; the second step holds since the data points are i.i.d. and so swapping the label of point i and point $n+1$ does not change the distribution; and the last step holds by applying out-of-sample stability (16). Therefore, returning to (24), we have

$$\mathbb{P}\left\{\widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha}(X_{n+1}) \not\supseteq \widehat{C}^{\text{oracle}}_{n,\alpha'}(X_{n+1})\right\} \leq \frac{n \cdot \nu}{\sqrt{\nu}(n+1)} \leq \sqrt{\nu}.$$

Combining this bound with (23), we have therefore proved that

$$\begin{aligned}
\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha}(X_{n+1})\right\} &\geq \mathbb{P}\left\{Y_{n+1} \notin \widehat{C}^{\text{oracle}}_{n,\alpha'}(X_{n+1})\right\} - \sqrt{\nu} \\
&\geq 1 - \alpha' - \sqrt{\nu} = 1 - \alpha - 2\sqrt{\nu}. \quad (25)
\end{aligned}$$

B.3.3 Bound for the jackknife+

The argument for the jackknife+ proceeds similarly, except that we now compare against the jackknife rather than the oracle—we will verify that

$$\widehat{C}^{\text{jackknife+},2\epsilon}_{n,\alpha}(X_{n+1}) \supseteq \widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha'}(X_{n+1})$$

holds with sufficiently high probability, where we again define $\alpha' = \alpha + \sqrt{\nu}$.

Suppose that this does not hold. Then it must either fail at the upper bounds of the intervals, i.e.,

$$\widehat{q}^{+}_{n,\alpha}\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}\} + 2\epsilon < \widehat{\mu}(X_{n+1}) + \widehat{q}^{+}_{n,\alpha'}\{R_i^{\mathrm{LOO}}\} + \epsilon$$

or at the lower bounds, i.e.,

$$\widehat{q}^{-}_{n,\alpha}\{\widehat{\mu}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}\} - 2\epsilon > \widehat{\mu}(X_{n+1}) - \widehat{q}^{+}_{n,\alpha'}\{R_i^{\mathrm{LOO}}\} - \epsilon.$$

In the first case, this implies that $\widehat{\mu}(X_{n+1}) > \widehat{\mu}_{-i}(X_{n+1}) + \epsilon$ for at least $\sqrt{\nu}(n+1)$ many indices $i \in \{1,\ldots,n\}$, while in the second case, we instead have $\widehat{\mu}(X_{n+1}) < \widehat{\mu}_{-i}(X_{n+1}) - \epsilon$ for at least $\sqrt{\nu}(n+1)$ many indices $i \in \{1,\ldots,n\}$. Combining the two, then, we have

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{C}^{\text{jackknife+},2\epsilon}_{n,\alpha}(X_{n+1}) \not\supseteq \widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha'}(X_{n+1})\right\} \\
&\qquad \leq \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{1}\left\{\left|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})\right| > \epsilon\right\} \geq \sqrt{\nu}(n+1)\right\} \\
&\qquad \leq \frac{\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\left\{\left|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})\right| > \epsilon\right\}\right]}{\sqrt{\nu}(n+1)} \leq \frac{\nu n}{\sqrt{\nu}(n+1)} \leq \sqrt{\nu}, \quad (26)
\end{aligned}$$

where we again apply Markov's inequality and the out-of-sample stability property just as for the jackknife proof. Combining the coverage result (25) (with α' in place of α) with (26), we have proved that

$$\begin{aligned}
\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{jackknife+},2\epsilon}_{n,\alpha}(X_{n+1})\right\} &\geq \mathbb{P}\left\{Y_{n+1} \notin \widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha'}(X_{n+1})\right\} - \sqrt{\nu} \\
&\geq \left[1 - \alpha' - 2\sqrt{\nu}\right] - \sqrt{\nu} = 1 - \alpha - 4\sqrt{\nu}.
\end{aligned}$$

B.4 Proof of Theorem 6

Let $R_i^{\mathrm{naive}} = \left|Y_i - \widehat{\mu}(X_i)\right|$ be the ith residual in the "naive" method while $R_i^{\mathrm{LOO}} = \left|Y_i - \widehat{\mu}_{-i}(X_i)\right|$ is the leave-one-out residual as before. Suppose that we run jackknife at level $1-\alpha'$, where $\alpha' = \alpha + \sqrt{\nu}$. Then by definition of the two methods, we have

$$\widehat{C}^{\mathrm{naive},2\epsilon}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \left(2\epsilon + \widehat{q}^{+}_{n,\alpha}\{R_i^{\mathrm{naive}}\}\right)$$

and

$$\widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha'}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \left(\epsilon + \widehat{q}^{+}_{n,\alpha'}\{R_i^{\mathrm{LOO}}\}\right).$$

We will now check that

$$\widehat{C}^{\mathrm{naive},2\epsilon}_{n,\alpha}(X_{n+1}) \supseteq \widehat{C}^{\text{jackknife},\epsilon}_{n,\alpha'}(X_{n+1})$$

with sufficiently high probability. Following similar arguments as in the proof of Theorem 5, if this does not hold, then it must be the case that

$$\sum_{i=1}^{n} \mathbb{1}\left\{R_i^{\text{LOO}} > R_i^{\text{naive}} + \epsilon\right\} \geq \lceil (1-\alpha)(n+1) \rceil - \Big(\lceil (1-\alpha')(n+1) \rceil - 1 \Big) \geq \sqrt{\nu}(n+1).$$

By the triangle inequality, this implies that

$$\sum_{i=1}^{n} \mathbb{1}\left\{\left|\widehat{\mu}(X_i) - \widehat{\mu}_{-i}(X_i)\right| > \epsilon\right\} \geq \sqrt{\nu}(n+1).$$

Therefore, by Markov's inequality,

$$\begin{aligned}\mathbb{P}\left\{\widehat{C}_{n,\alpha}^{\text{naive},2\epsilon}(X_{n+1}) \not\supseteq \widehat{C}_{n,\alpha'}^{\text{jackknife},\epsilon}(X_{n+1})\right\} \\ \leq \frac{\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\left\{\left|\widehat{\mu}(X_i) - \widehat{\mu}_{-i}(X_i)\right| > \epsilon\right\}\right]}{\sqrt{\nu}(n+1)} \leq \frac{n\nu}{\sqrt{\nu}(n+1)} \leq \sqrt{\nu},\end{aligned}$$

where the second step applies the in-sample stability property (17). Therefore,

$$\begin{aligned}\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha'}^{\text{naive},2\epsilon}(X_{n+1})\right\} &\geq \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha'}^{\text{jackknife},\epsilon}(X_{n+1})\right\} - \sqrt{\nu} \\ &\geq \left[1 - \alpha' - 2\sqrt{\nu}\right] - \sqrt{\nu} \geq 1 - \alpha - 4\sqrt{\nu},\end{aligned}$$

where for the next-to-last step we apply the result of Theorem 5 (with α' in place of α).

B.5 Proof of Theorem 2

In this section, we will prove a stronger version of Theorem 2, and will verify that the lower bounds on coverage hold even when we use the ϵ-inflated versions of each of the intervals, for any $\epsilon > 0$. That is, we will construct pathological examples for which, for the naive method and for jackknife, we have

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{naive},\epsilon}(X_{n+1})\right\} = \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{jackknife},\epsilon}(X_{n+1})\right\} = 0,$$

and for jackknife+ with $\alpha \leq \frac{1}{2}$, we have

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{jackknife+},\epsilon}(X_{n+1})\right\} \leq 1 - 2\alpha + 6\sqrt{\frac{\log n}{n}}.$$

B.5.1 Proof for jackknife and naive methods

First we construct a simple example for the jackknife and naive intervals. Define the regression method $\mathcal{A}$ as follows: given a training sample of size n, the algorithm returns the function

$$\widehat{\mu}(x) = \begin{cases} 0, & \text{if } x = X_i \text{ for any of the training points } X_i,\ i = 1, \ldots, n, \\ (1+\epsilon)n, & \text{otherwise.} \end{cases}$$

Now we define the data distribution on (X, Y): let $X \sim \mathcal{N}(0, 1)$, and let $Y \equiv 0$. Then with probability 1, the values $X_1, \ldots, X_{n+1}$ will be distinct. In this case, the leave-one-out residuals will be given by

$$R_i^{\mathrm{LOO}} = \big|Y_i - \widehat{\mu}_{-i}(X_i)\big| = \big|Y_i - (1+\epsilon)(n-1)\big| = (1+\epsilon)(n-1)$$

for all $i = 1, \ldots, n$, yielding a residual quantile $\widehat{q}^{+}_{n,\alpha}\big\{R_i^{\mathrm{LOO}}\big\} = (1+\epsilon)(n-1)$ and a prediction interval

$$\widehat{C}^{\mathrm{jackknife}}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}^{+}_{n,\alpha}\big\{R_i^{\mathrm{LOO}}\big\} = (1+\epsilon)n \pm (1+\epsilon)(n-1) = [1+\epsilon, (1+\epsilon)(2n-1)].$$

Therefore, the ϵ-inflated interval is given by

$$\widehat{C}^{\mathrm{jackknife},\epsilon}_{n,\alpha}(X_{n+1}) = [1, \epsilon + (1+\epsilon)(2n-1)].$$

However, $Y_{n+1} = 0$ with probability 1, and so

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\mathrm{jackknife},\epsilon}_{n,\alpha}(X_{n+1})\right\} = 0.$$

Similarly, for the naive method, its residuals will be given by

$$R_i^{\mathrm{naive}} = \big|Y_i - \widehat{\mu}(X_i)\big| = \big|Y_i - 0\big| = 0,$$

and so following the same argument we see that

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\mathrm{naive},\epsilon}_{n,\alpha}(X_{n+1})\right\} = 0.$$

A simple calculation shows that in this example, jackknife+ interval contains 0 at its left endpoint, and hence maintains its coverage.

B.5.2 Proof for jackknife+

Next we give the construction for the jackknife+. Our construction is similar in spirit to the example constructed by Vovk [2015, Appendix A], which gives intuition for why the leave-one-out cross-conformal predictor (described earlier in Section 2) may fail when the $n+1$ data points are only assumed to be exchangeable (specifically,

the data points and their residuals are chosen deterministically, and then randomly permuted). Here our construction is more technical as we need to work in the setting of i.i.d. data.

To give intuition we first sketch the idea. Suppose that the distribution of (X, Y) is chosen such that, with probability $\approx 2\alpha$, X is drawn from some "bad" region where predicting Y is challenging and we consistently underestimate Y, while with the remaining probability, X is drawn from some "good" region where predicting Y is easy—in fact, we can do this with zero error. Now, what is the chance that Y_{n+1} is not covered by the jackknife+? If X_{n+1} is "good", then Y_{n+1} will be covered. If X_{n+1} is "bad", then we will have

- For approximately half of the "bad" X_i ($\approx \alpha n$ data points), we will have
$$0 < Y_{n+1} - \widehat{\mu}_{-i}(X_{n+1}) < Y_i - \widehat{\mu}_{-i}(X_i).$$
- For all "good" X_i and for the remaining "bad" X_i (in total, $\approx (1-\alpha)n$ of the data points), we will have
$$Y_{n+1} - \widehat{\mu}_{-i}(X_{n+1}) > Y_i - \widehat{\mu}_{-i}(X_i) \geq 0.$$

This will be sufficient to see that Y_{n+1} is almost certainly not covered by the jackknife+ interval whenever X_{n+1} is "bad", i.e., with probability 2α.

Now we give the formal construction. Fix a small $\gamma > 0$ and a large $\tau > 0$, which we will specify later on. First, we will choose the distribution for the data. Let[9]

$$X_i = (A_i, B_i, C_i) \sim \text{Bernoulli}\big(2\alpha(1-\gamma)\big) \times \text{Unif}\{\pm 1\} \times \text{Unif}[-1, 1],$$

and let $Y_i = \tau A_i$. Next we define the regression algorithm $\mathcal{A}$ as follows: given a training sample $(X_j, Y_j) = \big((A_j, B_j, C_j), Y_j\big)$ indexed over $j = 1, \ldots, m$, the resulting fitted regression function $\widehat{\mu}$ is defined as

$$\widehat{\mu}(x) = \tau ac \cdot \prod_{j=1}^{m} B_j,$$

at any point $x = (a, b, c) \in \{0, 1\} \times \{\pm 1\} \times [-1, 1]$. Defining $B_{-i} = \prod_{j=1,\ldots,n; j\neq i} B_j$, we therefore have

$$\widehat{\mu}_{-i}(x) = \tau ac \cdot B_{-i}$$

[9]To obtain this distribution, if the dimension is $d \geq 3$ we can take the above distribution over the first 3 coefficients in X_i and simply ignore the remaining coefficients, while if $d = 1$ or $d = 2$ we can transform the data if needed; for example if $d = 1$, taking $X_i \sim \text{Unif}[0, 1]$, we can use the first k digits of X to generate a draw $A_i \sim \text{Bernoulli}(p)$ where $p \approx \big(2\alpha(1-\gamma)\big)$ (up to 10^{-k} accuracy), then we use the next digit of X_i to draw $B_i \sim \text{Unif}\{\pm 1\}$, and finally let $C_i \sim \text{Unif}[-1, 1]$ be defined by the remaining digits of X_i.

for each $i = 1, \ldots, n$. Now we check that coverage is roughly $1 - 2\alpha$. We have

$$\begin{aligned}
&\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}^{\text{jackknife+},\epsilon}_{n,\alpha}(X_{n+1})\right\} \\
&\geq \mathbb{P}\left\{Y_{n+1} > \widehat{q}^{+}_{n,\alpha}\left\{\widehat{\mu}_{-i}(X_{n+1}) + R^{\text{LOO}}_i + \epsilon\right\}\right\} \\
&\geq \mathbb{P}\left\{Y_{n+1} > \widehat{q}^{+}_{n,\alpha}\left\{\widehat{\mu}_{-i}(X_{n+1}) + R^{\text{LOO}}_i + \epsilon\right\} \text{ and } A_{n+1} = 1\right\} \\
&= 2\alpha(1-\gamma) \cdot \mathbb{P}\left\{\tau > \widehat{q}^{+}_{n,\alpha}\left\{\tau C_{n+1} B_{-i} + R^{\text{LOO}}_i + \epsilon\right\}\right\} \\
&= 2\alpha(1-\gamma) \cdot \mathbb{P}\left\{\tau > \widehat{q}^{+}_{n,\alpha}\left\{\tau C_{n+1} B_{-i} + \tau A_i\big(1 - C_i B_{-i}\big) + \epsilon\right\}\right\} \\
&= 2\alpha(1-\gamma) \cdot \mathbb{P}\left\{\widehat{q}^{+}_{n,\alpha}\left\{C_{n+1} B_{-i} + A_i\big(1 - C_i B_{-i}\big) + \frac{\epsilon}{\tau}\right\} < 1\right\} \\
&= 2\alpha(1-\gamma) \cdot \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{1}\left\{C_{n+1} B_{-i} + A_i\big(1 - C_i B_{-i}\big) + \frac{\epsilon}{\tau} < 1\right\} \geq (1-\alpha)(n+1)\right\}.
\end{aligned}$$

Next we verify that this last probability is close to 1. These indicator variables are independent conditional on $A_1, \ldots, A_n, B_1, \ldots, B_n, C_{n+1}$ (as they then depend only on C_i for each i). We denote the conditional probabilities by

$$P_i = \mathbb{P}\left\{C_{n+1} B_{-i} + A_i\big(1 - C_i B_{-i}\big) + \frac{\epsilon}{\tau} < 1 \;\middle|\; A_1, \ldots, A_n, B_1, \ldots, B_n, C_{n+1}\right\},$$

and calculate

$$P_i = \mathbb{1}\left\{C_{n+1} B_{-i} < 1 - \epsilon/\tau\right\} \cdot \begin{cases} 1, & \text{if } A_i = 0, \\ \frac{1 - B_{-i} C_{n+1} - \epsilon/\tau}{2}, & \text{if } A_i = 1. \end{cases}$$

By Hoeffding's inequality, we have

$$\begin{aligned}
\mathbb{P}\Bigg\{ &\sum_{i=1}^{n} \mathbb{1}\left\{C_{n+1} B_{-i} + A_i\big(1 - C_i B_{-i}\big) + \frac{\epsilon}{\tau} < 1\right\} \\
&\geq \sum_{i=1}^{n} P_i - t \;\Bigg|\; A_1, \ldots, A_n, B_1, \ldots, B_n, C_{n+1}\Bigg\} \geq 1 - e^{-2t^2/n},
\end{aligned}$$

for any $t \geq 0$. Choosing $t = \sqrt{\frac{n \log n}{2}}$ and marginalizing, we have

$$\mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{1}\left\{C_{n+1} B_{-i} + A_i\big(1 - C_i B_{-i}\big) + \frac{\epsilon}{\tau} < 1\right\} \geq \sum_{i=1}^{n} P_i - \sqrt{\frac{n \log n}{2}}\right\} \geq 1 - \frac{1}{n}.$$

Combining everything so far, we have therefore proved that

$$\begin{aligned}
&\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}^{\text{jackknife+},\epsilon}_{n,\alpha}(X_{n+1})\right\} \geq \\
&\qquad 2\alpha(1-\gamma) \cdot \left(1 - \frac{1}{n} - \mathbb{P}\left\{\sum_{i=1}^{n} P_i - \sqrt{\frac{n \log n}{2}} < (1-\alpha)(n+1)\right\}\right).
\end{aligned}$$

Now we bound this last probability. We have

$$\begin{aligned}\sum_{i=1}^{n} P_i &= \sum_{i=1}^{n} \mathbb{1}\left\{C_{n+1}B_{-i} < 1-\epsilon/\tau\right\} \cdot \left(\mathbb{1}\left\{A_i = 0\right\} + \mathbb{1}\left\{A_i = 1\right\} \cdot \frac{1 - B_{-i}C_{n+1} - \epsilon/\tau}{2}\right) \\ &\geq \mathbb{1}\left\{|C_{n+1}| \leq 1-\epsilon/\tau\right\} \cdot \sum_{i=1}^{n} \left((1-A_i) + A_i \cdot \frac{1 - B_{-i}C_{n+1} - \epsilon/\tau}{2}\right) \\ &\geq \mathbb{1}\left\{|C_{n+1}| \leq 1-\epsilon/\tau\right\} \cdot \left(n - \frac{1+\epsilon/\tau}{2}\sum_{i=1}^{n} A_i\right) - \frac{1}{2}\left|\sum_{i=1}^{n} A_i B_i\right|,\end{aligned}$$

where the last step holds since $A_i B_{-i} C_{n+1} = A_i B_i \cdot \big(C_{n+1}\prod_{j=1}^{n} B_j\big)$. Next, $|C_{n+1}| \leq 1-\epsilon/\tau$ holds with probability $1-\epsilon/\tau$, while by Hoeffding's inequality,

$$\mathbb{P}\left\{\sum_{i=1}^{n} A_i \leq n \cdot 2\alpha(1-\gamma) + \sqrt{\frac{n\log n}{2}}\right\} \geq 1 - \frac{1}{n}$$

and

$$\mathbb{P}\left\{\frac{1}{2}\left|\sum_{i=1}^{n} A_i B_i\right| \leq \sqrt{\frac{n\log n}{2}}\right\} \geq 1 - \frac{2}{n}.$$

Putting these calculations together,

$$\mathbb{P}\left\{\sum_{i=1}^{n} P_i \geq \left(n - \frac{1+\epsilon/\tau}{2}\left(n \cdot 2\alpha(1-\gamma) + \sqrt{\frac{n\log n}{2}}\right)\right) - \sqrt{\frac{n\log n}{2}}\right\} \geq 1 - \epsilon/\tau - \frac{3}{n}.$$

After simplifying,

$$\mathbb{P}\left\{\sum_{i=1}^{n} P_i \geq n\big(1 - \alpha(1+\epsilon/\tau)(1-\gamma)\big) - \sqrt{2n\log n}\right\} \geq 1 - \epsilon/\tau - \frac{3}{n}.$$

Now we choose $\gamma = \frac{2.15}{\alpha}\sqrt{\frac{\log(n)}{n}}$ and $\tau = \epsilon n$ (we can assume that $\gamma \leq 1$, since otherwise the theorem is trivial as it only claims that the coverage rate is no higher than 1). With this choice, we calculate

$$\mathbb{P}\left\{\sum_{i=1}^{n} P_i - \sqrt{\frac{n\log n}{2}} < (1-\alpha)(n+1)\right\} \leq \frac{4}{n},$$

and so returning to our earlier calculations we have

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}^{\text{jackknife+},\epsilon}_{n,\alpha}(X_{n+1})\right\} \geq 2\alpha(1-\gamma) \cdot \left(1 - \frac{5}{n}\right) \geq 2\alpha - 6\sqrt{\frac{\log(n)}{n}},$$

thus proving the theorem.