Electronic Journal of Statistics
Vol. 15 (2021) 4625–4658
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1910

Distribution-free conditional median inference[*]

Dhruv Medarametla[1] **and Emmanuel Candès**[2]

[1] *Department of Statistics, Stanford University*

[2] *Department of Statistics and Mathematics, Stanford University*

Abstract: We consider the problem of constructing confidence intervals for the median of a response $Y \in \mathbb{R}$ conditional on features $X \in \mathbb{R}^d$ in a situation where we are not willing to make any assumption whatsoever on the underlying distribution of the data (X, Y). We propose a method based upon ideas from conformal prediction and establish a theoretical guarantee of coverage while also going over particular distributions where its performance is sharp. Additionally, we prove an equivalence between confidence intervals for the conditional median and confidence intervals for the response variable, resulting in a lower bound on the length of any possible conditional median confidence interval. This lower bound is independent of sample size and holds for all distributions with no point masses.

MSC2020 subject classifications: 62G08, 62G15.
Keywords and phrases: Distribution-free, nonparametric inference, median regression, quantile regression, conformal inference.

Received March 2021.

Contents

1. Introduction

Consider a dataset $(X_1, Y_1), \dots, (X_n, Y_n) \subseteq \mathbb{R}^d \times \mathbb{R}$ and a test point (X_{n+1}, Y_{n+1}), with all datapoints being drawn i.i.d. from the same distribution P. Given

arXiv: 2102.07967

[*] D. M. thanks Stanford University for supporting this research as part of the Masters program in Statistics. E. C. was supported by Office of Naval Research grant N00014-20-12157, by the National Science Foundation grant OAC 1934578, and by the Army Research Office (ARO) under grant W911NF-17-1-0304.

our training data, can we provide a confidence interval for the expected value $\mu(X_{n+1}) = \mathbb{E}[Y_{n+1}|X_{n+1}]$?

Methods for inferring the conditional mean are certainly not in short supply. In fact, most existing methods predict not just the conditional mean at a datapoint $\mathbb{E}[Y_{n+1}|X_{n+1}]$, but the full conditional mean function $\mathbb{E}[Y|X = x]$ for all $x \in \mathbb{R}^d$. To the best of our knowledge, however, each approach relies on some assumptions in order to guarantee coverage. For instance, the classical linear regression model is often used but is only accurate if $Y|X$ is normal with mean $\mu(x) = \mathbb{E}[Y|X = x]$ affine in x and standard deviation independent of x. Nonparametric regressions cannot estimate the conditional mean without imposing smoothness conditions and assuming that the conditional distribution is sufficiently light tailed. Since reliable conditional mean inference is a very common problem, these methods are nevertheless used all the time, e.g. in predicting disease survival times, classifying spam emails, pricing financial assets, and more. The issue is that the assumptions these methods make rarely hold in practice. Thus, the question remains: is it possible to estimate the conditional mean at a datapoint in a *distribution-free* setting, with no assumptions on P?

It turns out that it is not only impossible to get a nontrivial confidence interval for the conditional mean $\mathbb{E}[Y|X = X_{n+1}]$, but it is actually impossible to get a confidence interval for $\mathbb{E}[Y]$ itself. This result originates in [1], where the authors show that any parameter sensitive to tails of a distribution cannot be estimated when no restrictions exist on the distribution class; an example of a distribution with a non-estimable mean is given in Appendix B.3.

Thus, within the distribution-free setting, making progress on inferring the conditional mean requires a modification to the problem statement. One strategy is to restrict the range of Y. An example of this is in [2], which introduces an algorithm that calculates a confidence interval for the conditional mean in the case where $Y \in \{0, 1\}$. However, even with this restriction, Barber shows that there exists a fundamental bound limiting how small any such confidence interval can be.

The other strategy is to modify the measure of central tendency that we study. Bahadur and Savage's result suggests that the best parameters to study are robust to distribution outliers; this observation motivates our investigation of the conditional median.

In a nutshell, the conditional median is possible to infer because of the strong and quantifiable relationship between any particular sampled datapoint (X_i, Y_i) and $\text{Median}(Y|X = X_i)$. Its robustness to outliers means that even within the distribution-free setting, there is no need to worry about 'hidden' parts of the distribution. Additionally, there already exists a well-known algorithm for estimating $\text{Median}(Y)$ given a finite number n of i.i.d. samples. Explored in [12] and covered in Appendix B.4, this algorithm produces intervals with guaranteed rates of coverage and widths going to zero as the sample size goes to infinity, suggesting that an algorithm for the conditional median might also perform well. We note that there already exists literature dealing with estimating the conditional quantile through quantile regression; [9] provides an introduction to different methods. The difference between quantile regression and our work here is that quantile regression requires continuity conditions and only provides theoretical guarantees at the asymptotic level, as seen in [5], [19], and [13]; the methods described here provide coverage guarantees on finite sample sizes and require no assumptions.

Our goal in this paper is to combine ideas from regular median inference with procedures from distribution-free inference in order to understand how well an algorithm can cover the conditional median and, more generally, conditional quantiles. In particular, we want to see if the properties of the median and quantiles lead to a valid inference method while also examining the limits of this inference.

It is important to note that the quantity we are attempting to study is $\text{Median}(Y|X = X_{n+1})$, not $\text{Median}(Y|X = x)$. The first term is a random variable dependent on X_{n+1}, whereas the second is fixed and exists for all x in the support of P. We focus on the first quantity because we are in the distribution-free setting; as we know nothing about the class of distributions that P belongs to, it is more tractable to make inferences about datapoints as opposed to pointwise across the full distribution.

Another way to frame this is to think of our goal as to cover the value of the conditional median function when weighting by the marginal distribution P_X, as opposed to covering the full conditional median function over all $x \in \mathbb{R}^d$. Because we are predicting the conditional median at an unknown value X_{n+1}, our success is not measured by whether or not we predict the conditional median correctly at all possible $x \in \mathbb{R}^d$, but by how often we predict the conditional median correctly across P_X.

The methods used in this paper are similar to those from distribution-free *predictive inference*, which focuses on predicting Y_{n+1} from a finite training set. The field of conformal predictive inference began with [20] and was built up by works such as [18] and [21]; it has been generating interest recently due to its versatility and lack of assumptions. Applications of conformal predictive inference range from criminal justice to drug discovery, as seen in [15] and [7] respectively.

While this paper relies on techniques from predictive inference, our focus is on *parameter inference*, which is quite different from prediction because it focuses on predicting a function of the conditional distribution $Y_{n+1}|X_{n+1}$ as opposed to the datapoint Y_{n+1}. For example, using sample datapoints from an unknown normal distribution to estimate the true mean of the distribution would constitute parameter inference; using the estimated quantiles of the datapoints to create a predictive confidence interval for the next datapoint would constitute predictive inference. Whereas predictive inference exploits the fact that (X_{n+1}, Y_{n+1}) is exchangeable with the sample datapoints, parameter inference requires another layer of analysis, as $(X_{n+1}, \text{Median}(Y_{n+1}|X_{n+1}))$ is not exchangeable with (X_i, Y_i). This additional complexity demands modifying approaches from predictive inference to produce valid $\underline{\text{parameter}}$ inference.

1.1. Terminology

We begin by setting up definitions to formalize the concepts above. Throughout this paper, we assume that any distribution $(X, Y) \sim P$ is over $\mathbb{R}^d \times \mathbb{R}$ unless explicitly stated otherwise. Each result in this paper holds true for all values of d and n.

Given a feature vector X_{n+1}, we let $\hat{C}_n(X_{n+1}) \subseteq \mathbb{R}$ denote a confidence interval for some functional of the conditional distribution $Y_{n+1}|X_{n+1}$. Note that we use the phrase confidence interval for convenience; in its most general form, $\hat{C}_n(X_{n+1})$ is a subset of $\mathbb{R}$. This interval is a function of the point X_{n+1} at which we seek inference as well as of our training data $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. We write $\hat{C}_n$ to refer to the general algorithm that maps $\mathcal{D}$ to the resulting confidence intervals $\hat{C}_n(x)$ for each $x \in \mathbb{R}^d$.

In order for $\hat{C}_n$ to be useful, we want it to *capture*, or contain, the parameter we care about with high probability. We formalize this as follows:

Definition 1. We say that $\hat{C}_n$ satisfies *distribution-free* ***median*** *coverage* at level $1 - \alpha$, denoted by $(1 - \alpha)$-Median, if

$$\mathbb{P}\big\{\text{Median}(Y_{n+1}|X_{n+1}) \in \hat{C}_n(X_{n+1})\big\} \geq 1 - \alpha$$
$$\text{for all distributions } P \text{ on } (X, Y) \in \mathbb{R}^d \times \mathbb{R}.$$

Definition 2. For $0 < q < 1$, let $\text{Quantile}_q(Y_{n+1}|X_{n+1})$ refer to the qth quantile of the conditional distribution $Y_{n+1}|X_{n+1}$. We say that $\hat{C}_n$ satisfies *distribution-free* ***quantile*** *coverage* for the qth quantile at level $1 - \alpha$, denoted by $(1 - \alpha, q)$-Quantile, if

$$\mathbb{P}\{\text{Quantile}_q(Y_{n+1}|X_{n+1}) \in \hat{C}_n(X_{n+1})\} \geq 1 - \alpha$$
$$\text{for all distributions } P \text{ on } (X, Y) \in \mathbb{R}^d \times \mathbb{R}.$$

The probabilities in Definitions 1 and 2 are both taken over the training data $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ and test point X_{n+1}. Thus, satisfying $(1 - \alpha)$-Median is equivalent to satisfying $(1 - \alpha, 0.5)$-Quantile.

These concepts are similar to predictive coverage, with the key difference being that our goal is now to predict a function of X_{n+1} rather than a new datapoint Y_{n+1}.

Definition 3. We say that $\hat{C}_n$ satisfies *distribution-free* ***predictive*** *coverage* at level $1 - \alpha$, denoted by $(1 - \alpha)$-Predictive, if

$$\mathbb{P}\big\{Y_{n+1} \in \hat{C}_n(X_{n+1})\big\} \geq 1 - \alpha \text{ for all distributions } P \text{ on } (X, Y) \in \mathbb{R}^d \times \mathbb{R},$$

where this probability is taken over the training data $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ and test point (X_{n+1}, Y_{n+1}).

Finally, we define the type of conformity scores that we will use in our general quantile inference algorithm.

Definition 4. We say that a function $f : (\mathbb{R}^d, \mathbb{R}) \to \mathbb{R}$ is a *locally nondecreasing conformity score* if, for all $x \in \mathbb{R}^d$ and $y, y' \in \mathbb{R}$ with $y \leq y'$, we have $f(x, y) \leq f(x, y')$.

1.2. Summary of results

We find that there exists a distribution-free predictive inference algorithm that satisfies both $(1 - \alpha/2)$-Predictive and $(1 - \alpha)$-Median. Moreover, an improved version of this algorithm also satisfies $(1 - \alpha, q)$-Quantile. Together, these prove that there exists nontrivial algorithms $\hat{C}_n$ that satisfy $(1 - \alpha)$-Median and $(1 - \alpha, q)$-Quantile for all $0 < q < 1$.

We go on to show that conditional median inference and predictive inference are nearly equivalent problems. Specifically, we show that any algorithm that contains $\text{Median}(Y_{n+1}|X_{n+1})$ with probability $1 - \alpha$ must also contain Y_{n+1} with probability at least $1 - \alpha$, and that any algorithm that contains Y_{n+1} with probability at least $1 - \alpha/2$ must contain $\text{Median}(Y_{n+1}|X_{n+1})$ with probability $1 - \alpha$.

Taken together, these results give us somewhat conflicting perspectives. On the one hand, there exist distribution-free algorithms that capture the conditional median and conditional quantile with high likelihood; on the other hand, any such algorithm will also capture a large proportion of the distribution itself, putting a hard limit on how well such algorithms can ever perform.

2. Confidence intervals for the conditional median

This section proves the existence of algorithms obeying distribution-free median and quantile coverage. We then focus on situations where these algorithms are sharp.

2.1. Basic conditional median inference

Algorithm 1 below operates by taking the training dataset and separating it into two halves of sizes $n_1 + n_2 = n$. Next, a regression algorithm $\hat{\mu}$ is trained on $\mathcal{D}_1 = \{(X_1, Y_1), \ldots, (X_{n_1}, Y_{n_1})\}$. The residuals $Y_i - \hat{\mu}(X_i)$ are calculated for $n_1 < i \leq n$, and the $1 - \alpha/2$ quantile of the absolute value of these residuals is then used to create a confidence band around the prediction $\hat{\mu}(X_{n+1})$. The expert will recognize that this is identical to a well-known algorithm from predictive inference as explained later.

Theorem 1. *For all distributions P, all regression algorithms $\hat{\mu}$, and all split sizes $n_1 + n_2 = n$, the output of Algorithm 1 contains Median$(Y_{n+1}|X_{n+1})$ with probability at least $1 - \alpha$. That is, the algorithm satisfies $(1 - \alpha)$-Median.*

The proof of Theorem 1 is covered in Appendix A.1.

```
Algorithm 1: Confidence Interval for Median($Y_{n+1}|X_{n+1}$) with Coverage $1-\alpha$
  Input:
    Number of i.i.d. datapoints $n \in \mathbb{N}$.
    Split sizes $n_1 + n_2 = n$.
    Datapoints $(X_1, Y_1), \dots, (X_n, Y_n) \sim P \subseteq (\mathbb{R}^d, \mathbb{R})$.
    Test point $X_{n+1} \sim P$.
    Regression algorithm $\hat{\mu}$.
    Coverage level $1-\alpha \in (0,1)$.
  Process:
    Randomly split $\{1, \dots, n\}$ into disjoint $\mathcal{I}_1$ and $\mathcal{I}_2$ with $|\mathcal{I}_1| = n_1$ and $|\mathcal{I}_2| = n_2$.
    Fit regression function $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ on $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$.
    For $i \in \mathcal{I}_2$ set $E_i = |Y_i - \hat{\mu}(X_i)|$.
    Compute $Q_{1-\alpha/2}(E)$, the $(1-\alpha/2)(1+1/n_2)$-th empirical quantile of $\{E_i : i \in \mathcal{I}_2\}$.
  Output:
    Confidence interval $\hat{C}_n(X_{n+1}) = [\hat{\mu}(X_{n+1}) - Q_{1-\alpha/2}(E), \hat{\mu}(X_{n+1}) + Q_{1-\alpha/2}(E)]$ for Median($Y_{n+1}|X_{n+1}$).
```

Remark 2.1. Algorithm 1 works independently of how $\hat{\mu}$ is trained; this means that any regression function may be used, from simple linear regression to more complicated machine learning algorithms. It is important to note that $\hat{\mu}(x)$ does not need to be an estimate of the true conditional mean $\mu(x)$; while one option is to train it to predict the conditional mean, it can be fit to predict the conditional median, conditional quantile, or any other measure of central tendency. The best choice for what to fit $\hat{\mu}$ to may depend on one's underlying belief about the distribution.

We return at last to the connection with predictive inference. Introduced in [14] and [20] and studied in [11], [3], [15], and several other papers, the *split conformal method* was initially created to achieve distribution-free predictive coverage guarantees. In particular, [20] shows that Algorithm 1 satisfies $(1-\alpha/2)$-Predictive, implying that in order to capture Median($Y_{n+1}|X_{n+1}$) with probability $1-\alpha$, our algorithm produces a wider confidence interval than an algorithm trying to capture Y_{n+1} with the same probability.

2.2. General conditional quantile inference

Algorithm 1 is a good first step towards a usable method for conditional median inference; however, it may be too rudimentary to be used in practice. Algorithm 2 is a more general version of Algorithm 1 that results in conditional quantile coverage and better empirical performance. This provides a better understanding of how diverse parameter inference algorithms can be.

Algorithm 2 differs from Algorithm 1 in two ways. First, we use the rq quantile of the lower scores to create the confidence interval's lower bound, and the $1 - s(1-q)$ quantile of the upper scores (corresponding to the top $s(1-q)$ of the score distribution) for the upper bound. Second, the functions we fit are no

Algorithm 2: Confidence Interval for $\text{Quantile}_q(Y_{n+1}|X_{n+1})$ with Coverage $1-\alpha$

```
Input:
    Number of i.i.d. datapoints $n \in \mathbb{N}$.
    Split sizes $n_1 + n_2 = n$.
    Datapoints $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P \subseteq (\mathbb{R}^d, \mathbb{R})$.
    Test point $X_{n+1} \sim P$.
    Locally nondecreasing conformity score algorithms $f^{\text{lo}}$ and $f^{\text{hi}}$.
    Quantile level $q \in (0, 1)$.
    Coverage level $1-\alpha \in (0, 1)$.
    Split probabilities $r + s = \alpha$.
Process:
    Randomly split $\{1, \ldots, n\}$ into disjoint $\mathcal{I}_1$ and $\mathcal{I}_2$ with $|\mathcal{I}_1| = n_1$ and $|\mathcal{I}_2| = n_2$.
    Fit conformity scores $f^{\text{lo}}, f^{\text{hi}} : (\mathbb{R}^d, \mathbb{R}) \to \mathbb{R}$ on $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$.
    For $i \in \mathcal{I}_2$ set $E_i^{\text{lo}} = f^{\text{lo}}(X_i, Y_i))$ and $E_i^{\text{hi}} = f^{\text{hi}}(X_i, Y_i))$.
    Compute $Q^{\text{lo}}_{rq}(E)$, the $rq(1 + 1/n_2) - 1/n_2$ empirical quantile of $\{E_i^{\text{lo}} : i \in \mathcal{I}_2\}$, and
    $Q^{\text{hi}}_{1-s(1-q)}(E)$, the $(1 - s(1-q))(1 + 1/n_2)$ empirical quantile of $\{E_i^{\text{hi}} : i \in \mathcal{I}_2\}$.
Output:
    Confidence interval
    $\hat{C}_n(X_{n+1}) = \{y : Q^{\text{lo}}_{rq}(E) \le f^{\text{lo}}(X_{n+1}, y), f^{\text{hi}}(X_{n+1}, y) \le Q^{\text{hi}}_{1-s(1-q)}(E)\}$ for
    $\text{Quantile}_q(Y_{n+1}|X_{n+1})$.
```

longer regression functions, but instead locally nondecreasing conformity scores. These scores are described in Definition 4; see Remark 2.3 for examples.

Theorem 2. *For all distributions P, all locally nondecreasing conformity scores f^{lo} and f^{hi}, all split sizes $n_1 + n_2 = n$, and all $0 < q < 1$, the output of Algorithm 2 contains $\text{Quantile}_q(Y_{n+1}|X_{n+1})$ with probability at least $1-\alpha$. That is, Algorithm 2 satisfies $(1-\alpha, q)$-Quantile.*

The proof of Theorem 2 is covered in Appendix A.2 and is similar to that of Theorem 1; the main modifications come from the changes described above. Regarding the first change, the asymmetrical quantiles on the lower and upper end of the E_i's balance the fact that datapoints have asymmetrical probabilities of being on either side of the conditional quantile. Regarding the second change, because the conformity scores E_i still preserve relative ordering, they do not affect the relationship between datapoints and the conditional quantile.

Remark 2.2. One possible choice for r and s is $r = s = \alpha/2$. This is motivated by the logic that r and s decide the probabilities of failure on the lower bound and the upper bound, respectively; if we want the bound to be equally accurate on both ends, it makes sense to set r and s equal. Another choice is $r = (1-q)\alpha$ and $s = q\alpha$; this results in the quantiles for Q^{lo}_{rq} and $Q^{\text{hi}}_{1-s(1-q)}$ being approximately equal, with the algorithm taking the $q(1-q)\alpha$ quantile of the scores on both the lower and upper ends.

Remark 2.3. The versatility of the conformity scores f^{lo} and f^{hi} is what differentiates Algorithm 2 from Algorithm 1 and makes it a viable option for conditional quantile inference. Below are a few examples of possible scores and the style of intervals they produce.

- $f^{\text{lo}}(X_i, Y_i) = f^{\text{hi}}(X_i, Y_i) = Y_i - \hat{\mu}(X_i)$, where $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ is trained on $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$ as a central tendency estimator. This is the conformity score used in Algorithm 1 and [20], resulting in a confidence interval of the form $[\hat{\mu}(X_{n+1}) + c_{\text{lo}}, \hat{\mu}(X_{n+1}) + c_{\text{hi}}]$ for some $c_{\text{lo}}, c_{\text{hi}} \in \mathbb{R}$. This score is best when the conditional distribution $Y|X = x$ is similar for all x and either the mean or median can be estimated with reasonable accuracy. Note that if the conditional distribution $Y - \mathbb{E}[Y|X]$ is independent of X, Algorithm 2 will output the same confidence interval for $\hat{\mu}(x) = \mathbb{E}[Y|X = x]$ and $\hat{\mu}(x) = \text{Median}(Y|X = x)$.
- $f^{\text{lo}}(X_i, Y_i) = f^{\text{hi}}(X_i, Y_i) = \frac{Y_i - \hat{\mu}(X_i)}{\hat{\sigma}(X_i)}$, where $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ and $\hat{\sigma} : \mathbb{R}^d \to \mathbb{R}^+$ are trained on $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$ as a central tendency estimator and conditional absolute deviation estimator, respectively. This score results in a confidence interval of the form $[\hat{\mu}(X_{n+1}) + c_{\text{lo}}\hat{\sigma}(X_{n+1}), \hat{\mu}(X_{n+1}) + c_{\text{hi}}\hat{\sigma}(X_{n+1})]$ for some $c_{\text{lo}}, c_{\text{hi}} \in \mathbb{R}$. Unlike the previous example, this score no longer results in a fixed-length confidence interval; it is best used when there is high heteroskedasticity in the underlying distribution. This is the conformity scores used to create adaptive predictive intervals in [10]. Note that a normalization constant $\gamma > 0$ can be added to the denominator $\hat{\sigma}(X_i)$ to create stable confidence intervals.
- $f^{\text{lo}}(X_i, Y_i) = Y_i - \hat{Q}^{\text{lo}}(X_i)$ and $f^{\text{hi}}(X_i, Y_i) = Y_i - \hat{Q}^{\text{hi}}(X_i)$, where $\hat{Q}^{\text{lo}}, \hat{Q}^{\text{hi}} : \mathbb{R}^d \to \mathbb{R}$ are trained on $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$ to estimate the rq quantile and the $1 - s(1 - q)$ quantile of the conditional distribution, respectively. This choice results in a confidence interval of the form $[\hat{Q}^{\text{lo}}(X_{n+1}) + c_{\text{lo}}, \hat{Q}^{\text{hi}}(X_{n+1}) + c_{\text{hi}}]$ for some $c_{\text{lo}}, c_{\text{hi}} \in \mathbb{R}$. These scores are best when one can estimate the conditional quantiles reasonably well and the conditional distribution $Y|X = x$ is heteroskedastic. Note that if $\hat{Q}^{\text{lo}}$ and $\hat{Q}^{\text{hi}}$ are trained well, then the resulting confidence interval will be approximately $[\hat{Q}^{\text{lo}}(X_{n+1}), \hat{Q}^{\text{hi}}(X_{n+1})]$. These are the scores used to create the predictive intervals seen in [16] and [17].
- $f^{\text{lo}}(X_i, Y_i) = f^{\text{hi}}(X_i, Y_i) = \hat{F}_{Y|X=X_i}(Y_i)$, where $\hat{F}_{Y|X=x} : \mathbb{R}^d \times \mathbb{R} \to [0, 1]$ is trained on $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$ to be the estimated cumulative distribution function of the conditional distribution $Y|X$. Using this score will result in a confidence interval $[\hat{F}^{-1}_{Y|X=X_{n+1}}(c_{\text{lo}}), \hat{F}^{-1}_{Y|X=X_{n+1}}(c_{\text{hi}})]$ for some $c_{\text{lo}}, c_{\text{hi}} \in [0, 1]$, similar to the predictive intervals in [6] and [8]. This can be a good approach when the conditional distribution $Y|X$ is particularly complex.
- $f^{\text{lo}}(X_i, Y_i) = f^{\text{hi}}(X_i, Y_i) = \log Y_i - \hat{\mu}(X_i)$, where $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ is trained on $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$ as a log central tendency estimator. This results in a confidence interval of the form $[c_{\text{lo}} \exp(\hat{\mu}(X_{n+1})), c_{\text{hi}} \exp(\hat{\mu}(X_{n+1}))]$ for some $c_{\text{lo}}, c_{\text{hi}} \in \mathbb{R}^+$. This score works well when Y is known to be positive and one wants to minimize the approximation ratio; it is equivalent to taking a log transformation of the data.

In general, a good choice for $f^{\text{lo}}(X_i, Y_i)$ and $f^{\text{hi}}(X_i, Y_i)$ depends on one's underlying belief about the distribution as well as on the sample size n, though some scores perform better in practice. The choice of n_1 and n_2 represents a balance between model training and interval tightness; increasing n_1 increases the amount of data for f^{lo} and f^{hi}, and increasing n_2 results in a better quantile for the predictive interval. [17] contains more information on the effect of the conformity score on the size of predictive intervals as well as the impact of the ratio n_1/n on interval width and coverage. We also simulate the impact of different scores on conditional quantile intervals in Section 4.

Remark 2.4. Algorithm 2 can be generalized to use more than one split, resulting in multiple confidence intervals that can then be combined into one output. For more information, [22] describes how to apply k-fold cross validation to split conformal inference, and [4] describes the special case of having n different splits using the jackknife+ procedure.

2.3. Algorithm sharpness

Now that we have seen that Algorithms 1 and 2 achieve coverage, an important question to ask is whether or not the terms for the error quantile can be improved. Do our methods consistently overcover the conditional median, and if so, is it possible to take a lower quantile of the error terms and still have Theorems 1 and 2 hold? In this section, we prove that this is impossible by going over a particular distribution P^δ for which the $1-\alpha/2$ term in Algorithm 1 is necessary. Additionally, we go over a choice $\hat{\mu}_c$ with the property that Algorithm 1 *always* results in $1-\alpha/2$ coverage when ran with input $\hat{\mu}_c$; this implies that there does not exist a distribution with the property that Algorithm 1 will always provide a sharp confidence interval for the conditional median regardless of the regression algorithm.

For each $\delta > 0$, consider $(X,Y) \sim P^\delta$ over $\mathbb{R}\times\mathbb{R}$, where $P^\delta_X = \mathrm{Unif}[-0.5, 0.5]$ and $Y|X \overset{\mathrm{d}}{=} X\,B$; $B \in \{0,1\}$ is here an independent Bernoulli variable with $\mathbb{P}(B=1) = 0.5+\delta$. That is, $0.5+\delta$ of the distribution is on the line segment $Y = X$ from $(-0.5,-0.5)$ to $(0.5,0.5)$, and $0.5-\delta$ of the distribution is on the line segment $Y = 0$ from $(-0.5,0)$ to $(0.5,0)$. Thus, $\mathrm{Median}(Y|X=x) = x$. A visualization of P^δ is shown in Figure 1.

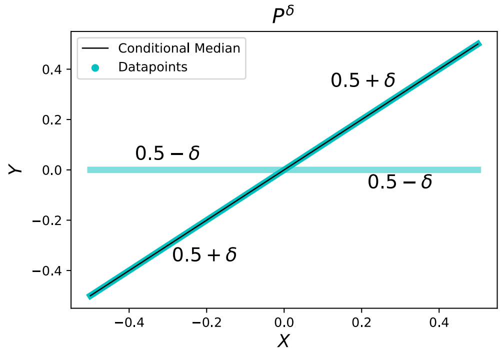


FIG 1. *A distribution for which it is difficult to estimate the conditional median.*

We know that Algorithm 1 is accurate for all distributions P and all algorithms $\hat{\mu}$. Consider the regression algorithm $\hat{\mu} : \mathbb{R} \to \mathbb{R}$ such that $\hat{\mu}(x) = 0$ for all $x \in \mathbb{R}$; in other words, $\hat{\mu}$ predicts $Y_i = 0$ for all X_i. We show that Algorithm 1 returns a coverage almost exactly equal to $1 - \alpha$.

Theorem 3. *For all $\epsilon > 0$, there exist N and $\delta > 0$ such that if we sample $n > N$ datapoints from the distribution P^δ and use Algorithm 1 with $\hat{\mu} = 0$ as defined above and $n_1 = n_2 = n/2$ to get a confidence interval for the conditional median,*

$$\mathbb{P}\big\{\text{Median}(Y_{n+1}|X_{n+1}) \in \hat{C}_n(X_{n+1})\big\} \leq 1 - \alpha + \epsilon.$$

The proof is in Appendix A.3. Theorem 3 does not directly prove that the $1 - \alpha/2$ term in Algorithm 1 is sharp. However, we can see that if Algorithm 1 used the $1 - \alpha'/2$ quantile of the residuals with $\alpha' > \alpha$, then by Theorem 3 there would exist a choice for δ and n where the probability of conditional median coverage would be less than $1 - \alpha$. Therefore, the $1 - \alpha/2$ term is required for the probability of coverage to always be at least $1 - \alpha$.

Remark 2.5. It is possible to generalize Theorem 3 to Algorithm 2 as well; we can change P^δ to have $Y|X \sim X\,B$ with $B \sim \text{Bernoulli}(q + \mathbf{1}[X \geq 0](1 - 2q) + \delta)$. This results in $\text{Quantile}_q(Y|X = x) = x$ for all $x \in [-0.5, 0.5]$. Then, if we consider the conformity scores $f^{\text{lo}}(X_i, Y_i) = f^{\text{hi}}(X_i, Y_i) = Y_i$, it can be shown for large n and small δ that Algorithm 2 returns a confidence interval that has a conditional quantile coverage of at most $1 - \alpha + \epsilon$, meaning that the rq and $1 - s(1 - q)$ terms in the quantiles for the error scores are sharp.

These results may seem somewhat pedantic because we are restricting $\hat{\mu}(x)$ to be the zero function and $f^{\text{lo}}(X_i, Y_i)$ and $f^{\text{hi}}(X_i, Y_i)$ to be Y_i; this simplification is done to better illustrate our point. Even when $f^{\text{lo}}(X_i, Y_i)$ and $f^{\text{hi}}(X_i, Y_i)$ are trained using more complicated approaches, there still exist distributions that result in only $1 - \alpha$ coverage for Algorithm 2. For an example of a distribution where Algorithm 2 only achieves $1 - \alpha$ coverage for standard conformity scores $f^{\text{lo}}(X_i, Y_i)$ and $f^{\text{hi}}(X_i, Y_i)$, refer to P_3 in Section 4. The existence of P^δ and similarly 'confusing' distributions helps to show why capturing the conditional median can be tricky in a distribution-free setting.

At the same time, there exist conformity scores for which Algorithms 1 and 2 have rates of coverage that are always near $1 - \alpha/2$. For $c > 0$, define the randomized regression function $\hat{\mu}_c$ as follows: set $M = \max_{i \in \mathcal{I}_1} |Y_i|$ and $\hat{\mu}_c(x) = A_x$ for all $x \in \mathbb{R}^d$, where $A_x \overset{i.i.d.}{\sim} \mathcal{N}(0, (cM)^2)$. We prove the following:

Theorem 4. *For all $\epsilon > 0$, there exists c and N such that for all $n > N$, there is a split $n_1 + n_2 = n$ such that when Algorithm 1 is ran using the regression function $\hat{\mu}_c$ on n datapoints with $\mathcal{I}_1$ of size n_1 and $\mathcal{I}_2$ of size n_2, the resulting interval will be finite and will satisfy*

$$\mathbb{P}\big\{\text{Median}(Y_{n+1}|X_{n+1}) \in \hat{C}_n(X_{n+1})\big\} \geq 1 - \alpha/2 - \epsilon$$

for ***any*** *distribution P.*

The proof for this theorem is in Appendix A.4.

Remark 2.6. Theorem 4 can be extended to Algorithm 2 by taking $f^{\mathrm{lo}}(X_i, Y_i) = f^{\mathrm{hi}}(X_i, Y_i) = Y_i - \hat{\mu}_c(X_i)$. The corresponding result shows that given a large enough number of datapoints and a particular data split, there exist conformity scores that result in Algorithm 2 capturing the conditional quantile nontrivially with probability at least $1 - rq - s(1-q)$ for all distributions P.

Due to the definition of $\hat{\mu}_c$, the resulting confidence intervals will be near-useless; the predictions will be so far off that the intervals will have width several times the range of the (slightly clipped) marginal distribution P_Y. However, they still will be finite, and will still achieve predictive inference at a rate roughly equal to $1 - \alpha/2$. The existence of scores that always result in higher-than-needed rates of coverage means that any result like Theorem 3 that provides a nontrivial upper bound for either Algorithm 1 or Algorithm 2's coverage of the conditional median on a specific distribution will have to restrict the class of regression functions and/or conformity scores.

3. Median intervals and predictive intervals are equivalent

Up until this point, we have looked at the existence and accuracy of algorithms for estimating the conditional median. This section shows that any algorithm for the conditional median is also a predictive algorithm, and vice versa. As a consequence, this means there exists a strong lower bound on the size of any conditional median confidence interval.

Theorem 5. *Let $\hat{C}_n$ be any algorithm that satisfies $(1-\alpha)$-Median. Then, for any nonatomic distribution P on $\mathbb{R}^d \times \mathbb{R}$, we have that*

$$\mathbb{P}\big\{Y_{n+1} \in \hat{C}_n(X_{n+1})\big\} \geq 1 - \alpha.$$

That is, $\hat{C}_n$ satisfies $(1-\alpha)$-Predictive for all nonatomic distributions P.

Proof. The proof above uses the same approach from the proof of Theorem 1 in [2]. Consider an arbitrary $\hat{C}_n$ that satisfies $(1-\alpha)$-Median, and let P be any distribution over $\mathbb{R}^d \times \mathbb{R}$ for which P_X is nonatomic. Pick some $M \geq n+1$, and sample $\mathcal{L} = \{(X^j, Y^j) : 1 \leq j \leq M\} \overset{\text{i.i.d.}}{\sim} P$. We define two different ways of sampling our data from $\mathcal{L}$.

Fix $\mathcal{L}$ and pick $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ **without** replacement from $\mathcal{L}$. Call this method of sampling Q_1. It is clear that after marginalizing over $\mathcal{L}$, the (X_i, Y_i)'s are effectively drawn i.i.d. from P; thus, we have that

$$\mathbb{P}_P\{Y_{n+1} \in \hat{C}_n(X_{n+1})\} = \mathbb{E}_{\mathcal{L}}\big[\mathbb{P}_{Q_1}\{Y_{n+1} \in \hat{C}_n(X_{n+1}) | \mathcal{L}\}\big].$$

Now, pick $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ **with** replacement from $\mathcal{L}$, and call this method of sampling Q_2. Note that because P_X is nonatomic, the X^j's are distinct with probability 1, which means that $\mathrm{Median}_{Q_2}(Y | X = X^j) = Y^j$. Then, as $\hat{C}_n$ applies to all distributions, it applies to our point distribution over $\mathcal{L}$; thus, we have that for all $\mathcal{L}$,

$$\begin{aligned}\mathbb{P}_{Q_2}\{Y_{n+1} \in \hat{C}_n(X_{n+1})|\mathcal{L}\} =&\mathbb{P}_{Q_2}\{\text{Median}_{Q_2}(Y|X = X_{n+1}) \in \hat{C}_n(X_{n+1})|\mathcal{L}\} \\ \geq& 1-\alpha.\end{aligned}$$

Let R be the event that any two of the (X_i, Y_i) are equal to each other; that is, $R = \{(X_a, Y_a) = (X_b, Y_b) \text{ for any } a < b\}$. Note that under Q_2, for $1 \leq a < b \leq n+1$, the probability of (X_a, Y_a) and (X_b, Y_b) being equal is $1/M$. Then, by the union bound,

$$\mathbb{P}_{Q_2}\{R\} \leq \sum_{1\leq a<b\leq n+1} \mathbb{P}_{Q_2}\{(X_a, Y_a) = (X_b, Y_b)\} \leq \frac{n^2}{M},$$

where the last step is from the fact that the number of possible pairs (a, b) is bounded above by n^2. Meanwhile, we know that $\mathbb{P}_{Q_1}\{R\} = 0$ by the definition of Q_1.

We can use this to bound the total variation distance between Q_1 and Q_2. For any fixed $\mathcal{L}$ and any event E, note that $\mathbb{P}_{Q_1}\{E\} = \mathbb{P}_{Q_1}\{E|R^C\} = \mathbb{P}_{Q_2}\{E|R^C\}$. Then, we can calculate

$$\begin{aligned}&|\mathbb{P}_{Q_1}\{E\} - \mathbb{P}_{Q_2}\{E\}| \\ =&|\mathbb{P}_{Q_1}\{E|R^C\} - (\mathbb{P}_{Q_2}\{E|R^C\}\mathbb{P}_{Q_2}\{R^C\} + \mathbb{P}_{Q_2}\{E|R\}\mathbb{P}_{Q_2}\{R\})| \\ =&|\mathbb{P}_{Q_2}\{E|R^C\}\mathbb{P}_{Q_2}\{R\} - \mathbb{P}_{Q_2}\{E|R\}\mathbb{P}_{Q_2}\{R\}| \\ =&\mathbb{P}_{Q_2}\{R\}|\mathbb{P}_{Q_2}\{E|R^C\} - \mathbb{P}_{Q_2}\{E|R\}| \\ \leq& n^2/M.\end{aligned}$$

Therefore, for any fixed $\mathcal{L}$, the total variation distance between the distributions Q_1 and Q_2 is at most n^2/M, implying that

$$\mathbb{P}_{Q_1}\{Y_{n+1} \in \hat{C}_n(X_{n+1})|\mathcal{L}\} \geq \mathbb{P}_{Q_2}\{Y_{n+1} \in \hat{C}_n(X_{n+1}))|\mathcal{L}\} - n^2/M \geq 1-\alpha-n^2/M,$$

which means that

$$\mathbb{P}_P\{Y_{n+1} \in \hat{C}_n(X_{n+1})\} = \mathbb{E}_{\mathcal{L}}\left[\mathbb{P}_{Q_1}\{Y_{n+1} \in \hat{C}_n(X_{n+1})|\mathcal{L}\}\right] \geq 1-\alpha-n^2/M.$$

Taking the limit as M goes to infinity gives the result. □

Remark 3.1. Theorem 5 also applies to all algorithms that satisfy $(1-\alpha, q)$-Quantile; for our uniform distribution over $\mathcal{L}$, $\text{Quantile}_q(Y|X = X^j) = Y^j$, so the proof translates exactly. As a result, this means that all algorithms that satisfy $(1-\alpha, q)$-Quantile also satisfy $(1-\alpha)$-Predictive for all nonatomic distributions P.

Remark 3.2. The approach taken in the proof of Theorem 5 is similar to those used to show the limits of distribution-free inference in other settings. As mentioned earlier, [2] shows that in the setting of distributions P over $\mathbb{R}^d \times \{0, 1\}$, any confidence interval $\hat{C}_n(X_{n+1})$ for $\mathbb{E}[Y|X = X_{n+1}]$ with coverage $1-\alpha$ must contain Y_{n+1} with probability $1-\alpha$ for all nonatomic distributions P, and goes on to provide a lower bound for the length of the confidence interval. Additionally, [3] proves a similar theorem about predictive algorithms $\hat{C}_n(X_{n+1})$ for Y_{n+1} that are required to have a weak form of conditional coverage. The proof for the result from [2] involves the same idea of marginalizing over a large finite sampled subset $\mathcal{L}$ in order to apply $\hat{C}_n$ to the distribution over $\mathcal{L}$; the proof for the result from [3] focuses on sampling a large number of datapoints conditioned on whether or not they belong to a specific subset $\mathcal{B} \subseteq \mathbb{R}^d \times \mathbb{R}$. In both cases, studying two sampling distributions and measuring the total variation distance between them was crucial. Thus, it seems that this strategy may have further use in the future when studying confidence intervals for other parameters or data in a distribution-free setting.

We now prove a similar result in the opposite direction.

Theorem 6. *Let $\hat{C}_n$ be any algorithm that only outputs confidence intervals and satisfies $(1-\alpha/2)$-Predictive. Then, $\hat{C}_n$ satisfies $(1-\alpha)$-Median.*

Proof. Consider any distribution P. For all $x \in \mathbb{R}^d$ in the support of P, set $m(x) = \text{Median}(Y|X = x)$.

We know that under the distribution P, $\mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1})\} \geq 1-\alpha/2$. Then, we can condition on whether or not the conditional median is contained in $\hat{C}_n$ as follows:

$$\begin{aligned}
&1-\alpha/2\\
\leq&\mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1})\}\\
=&\mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1})|m(X_{n+1}) \in \hat{C}_n(X_{n+1})\}\mathbb{P}\{m(X_{n+1}) \in \hat{C}_n(X_{n+1})\}\\
&+\mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1})|m(X_{n+1}) \notin \hat{C}_n(X_{n+1})\}\mathbb{P}\{m(X_{n+1}) \notin \hat{C}_n(X_{n+1})\}\\
\leq&\mathbb{P}\{m(X_{n+1}) \in \hat{C}_n(X_{n+1})\} + \frac{1}{2}\mathbb{P}\{m(X_{n+1}) \notin \hat{C}_n(X_{n+1})\}\\
=&\frac{1}{2} + \frac{1}{2}\mathbb{P}\{m(X_{n+1}) \in \hat{C}_n(X_{n+1})\}.
\end{aligned}$$

The key insight here is that because $\hat{C}_n$ only outputs confidence intervals, $\hat{C}_n(X_{n+1})$ can cover at most half of the conditional distribution of $Y|X = X_{n+1}$ if $m(X_{n+1})$ is not contained in the confidence interval.

Shifting the constant over and multiplying by 2 tells us that $\mathbb{P}\{m(X_{n+1}) \in \hat{C}_n(X_{n+1})\} \geq 1-\alpha$, as desired. □

Remark 3.3. Theorem 6 can be modified to replace $(1-\alpha)$-Median with $(1-\alpha, q)$-Quantile. In particular, let $\hat{C}_n(x) = [\hat{L}_n(x), \hat{H}_n(x)]$ be any algorithm that only outputs confidence intervals and satisfies $\mathbb{P}\{Y_{n+1} \geq \hat{L}_n(X_{n+1})\} \geq 1-rq$ and $\mathbb{P}\{Y_{n+1} \leq \hat{H}_n(X_{n+1})\} \geq 1-s(1-q)$ for some $r+s=\alpha$. Then, $\hat{C}_n$ satisfies $(1-\alpha, q)$-Quantile.

Alternatively, we can also conclude that if $\hat{C}_n$ only outputs confidence intervals and satisfies $(1-\min\{q, 1-q\}\alpha)$-Predictive, then it satisfies $(1-\alpha, q)$-Quantile. The proofs of both of these statements are in Appendix A.5.

Theorem 5 tells us that conditional median inference is at least as imprecise as predictive inference. As a result, because all predictive intervals have nonvanishing widths (assuming nonzero conditional variance) no matter the sample size n, it is not possible to write down conditional median algorithms with widths converging to 0. Thus, it may be better to study other distribution parameters similar to the conditional median if we are looking for better empirical performance. For discussion on related distribution parameters that are worth studying and may result in stronger inference, refer to Section 5.2.

Theorem 6 tells us that one way to approach conditional median inference is to apply strong predictive algorithms. It also suggests that improvements in predictive inference may translate to conditional median inference.

Lastly, we know that Algorithm 1 captures Y_{n+1} with probability $1 - \alpha/2$. Because there is space between the bounds of Theorems 5 and 6, there may exist a better conditional median algorithm that only captures Y_{n+1} with probability $1-\alpha$. Based on our result from Section 2.3, any such algorithm will likely follow a format different than the split conformal approach. Studying this problem in more detail, particularly on difficult distributions P, might lead to more accurate conditional median algorithms.

4. Simulations

In this section, we analyze the impact of different conformity scores on the outcome of Algorithm 2. Specifically, we look at the four following conformity scores:

Score 1: $f_1^{\mathrm{lo}}(X_i, Y_i) = f_1^{\mathrm{hi}}(X_i, Y_i) = Y_i - \hat{\mu}(X_i)$. We train $\hat{\mu}$ to predict the conditional mean using quantile regression forests on the dataset $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$.

Score 2: $f_2^{\mathrm{lo}}(X_i, Y_i) = f_2^{\mathrm{hi}}(X_i, Y_i) = \frac{Y_i - \hat{\mu}(X_i)}{\hat{\sigma}(X_i)}$. We train $\hat{\mu}$ and $\hat{\sigma}$ jointly using random forests on the dataset $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$.

Score 3: $f_3^{\mathrm{lo}}(X_i, Y_i) = Y_i - \hat{Q}^{\mathrm{lo}}(X_i)$ and $f_3^{\mathrm{hi}}(X_i, Y_i) = Y_i - \hat{Q}^{\mathrm{hi}}(X_i)$. We train $\hat{Q}^{\mathrm{lo}}$ and $\hat{Q}^{\mathrm{hi}}$ to predict the conditional $\alpha/2$ quantile and $1 - \alpha/2$ quantile, respectively, using quantile regression forests on the dataset $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$.

Score 4: $f_4^{\mathrm{lo}}(X_i, Y_i) = f_4^{\mathrm{hi}}(X_i, Y_i) = \hat{F}_{Y|X=X_i}(Y_i)$. We create $\hat{F}_{Y|X=X_i}(Y_i)$ by using 101 quantile regression forests trained on $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$ to estimate the conditional q^{th} quantile for $q \in \{0, 0.01, \ldots, 0.99, 1\}$ and use linear interpolation between quantiles to estimate the conditional CDF. Our training method ensures that quantile predictions will never cross.

To compare the performance of Algorithm 2 against a method that does not have a distribution-free guarantee, we also consider a nonconformalized algorithm that outputs $\hat{C}_n(X_{n+1}) = [\hat{Q}^{\mathrm{lo}}(X_{n+1}), \hat{Q}^{\mathrm{hi}}(X_{n+1})]$, where $\hat{Q}^{\mathrm{lo}}$ and $\hat{Q}^{\mathrm{hi}}$ are trained on $\mathcal{D}$ to predict the conditional $\alpha/2$ quantile and $1 - \alpha/2$ quantile, respectively, using quantile regression forests. We refer to this algorithm as QRF.

In order to test the conditional median coverage rate, we must look at distributions for which the conditional median is known and, therefore, focus on simulated datasets. We consider the performance of Algorithm 2 and QRF on these three distributions:

Distribution 1: We draw $(X, Y) \sim P_1$ from $\mathbb{R}^d \times \mathbb{R}$, where $d = 10$. Here, $X = (X^1, \ldots, X^d)$ is an equicorrelated multivariate Gaussian vector with mean zero and $\operatorname{Var}(X^i) = 1$, $\operatorname{Cov}(X^i, X^j) = 0.25$ for $i \neq j$. We set $Y = (X^1 + X^2)^2 - X^3 + \sigma(X)\epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ is independent of X and $\sigma(x) = 0.1 + 0.25\|x\|_2^2$ for all $x \in \mathbb{R}^d$.

Distribution 2: We draw $(X, Y) \sim P_2$ from $\mathbb{R} \times \mathbb{R}$. Draw $X \sim \operatorname{Unif}[-4\pi, 4\pi]$ and $Y = U^{1/4} f(X)$, where $U \sim \operatorname{Unif}[0, 1]$ is independent of X and $f(x) = 1 + |x| \sin^2(x)$ for all $x \in \mathbb{R}$.

Distribution 3: We draw $(X, Y) \sim P_3$ from $\mathbb{R} \times \mathbb{R}$. Draw $X \sim \operatorname{Unif}[-1, 1]$ and set $Y = B \cdot f(X)$, where $B \sim \operatorname{Bernoulli}(0.5 + 2\delta)$ is independent of X and $f(x) = \gamma\{Mx\} - \frac{\gamma}{2} - (-1)^{\lfloor Mx \rfloor}(1 - \frac{\gamma}{2})$ for all $x \in \mathbb{R}$. Note that $\{r\}$ is the fractional part of r. We set $\delta = 0.0001$ and $M = 1/\gamma = 25$.

Distributions 2 and 3 are shown in Figure 2.

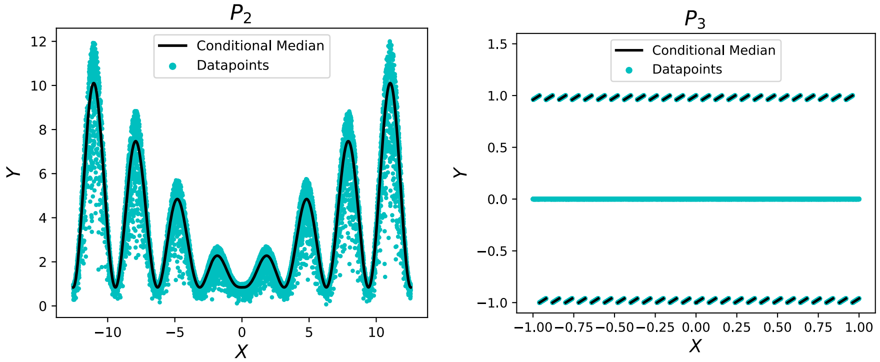


FIG 2. *Plots of* $n = 10{,}000$ *datapoints from the two distributions* P_2 *and* P_3 *overlaid with the conditional medians. The left panel is a case where the conditional distribution* $Y|X$ *has high heteroskedasticity. The right is a case where it is nearly impossible to tell the location of the conditional median.*

For each distribution, we run Algorithm 2 using each conformity score, as well as QRF, to get a confidence interval for the conditional median. We run 500 trials; in each trial, we set $n = 5{,}000$ with $n_1 = n_2 = n/2$ and $\alpha = 0.1$ with $r = s = \alpha/2$. For a separate study on the impact of n on the confidence interval width and coverage, we refer the reader to [17]. We test coverage on $5{,}000$ datapoints for each trial. The average coverage rate, average interval width, and other statistics for each distribution and conformity score are shown in Figure 3. An example of the resulting confidence intervals for a single trial on Distribution 2 are displayed in Figure 4.

Distribution	Score	AC	SDAC	MCC	AW	SDAW
1	1	99.08%	0.20%	0.0%	14.08	0.573
	2	99.78%	0.09%	5.0%	13.03	0.405
	3	99.72%	0.10%	8.4%	12.86	0.358
	4	99.77%	0.09%	12.0%	13.02	0.398
	QRF	99.73%	0.09%	1.6%	11.10	0.191
2	1	99.85%	0.20%	92.4%	4.537	0.174
	2	99.48%	0.29%	78.4%	3.604	0.093
	3	99.89%	0.14%	95.6%	3.619	0.060
	4	99.87%	0.14%	93.6%	3.700	0.087
	QRF	99.91%	0.12%	93.4%	3.48	0.051
3	1	90.05%	0.86%	15.4%	2.122	0.044
	2	89.97%	0.87%	19.8%	2.084	0.051
	3	90.14%	0.87%	4.4%	1.989	0.003
	4	90.00%	0.88%	0.0%	1.990	0.002
	QRF	83.98%	0.92%	0.0%	1.962	0.017

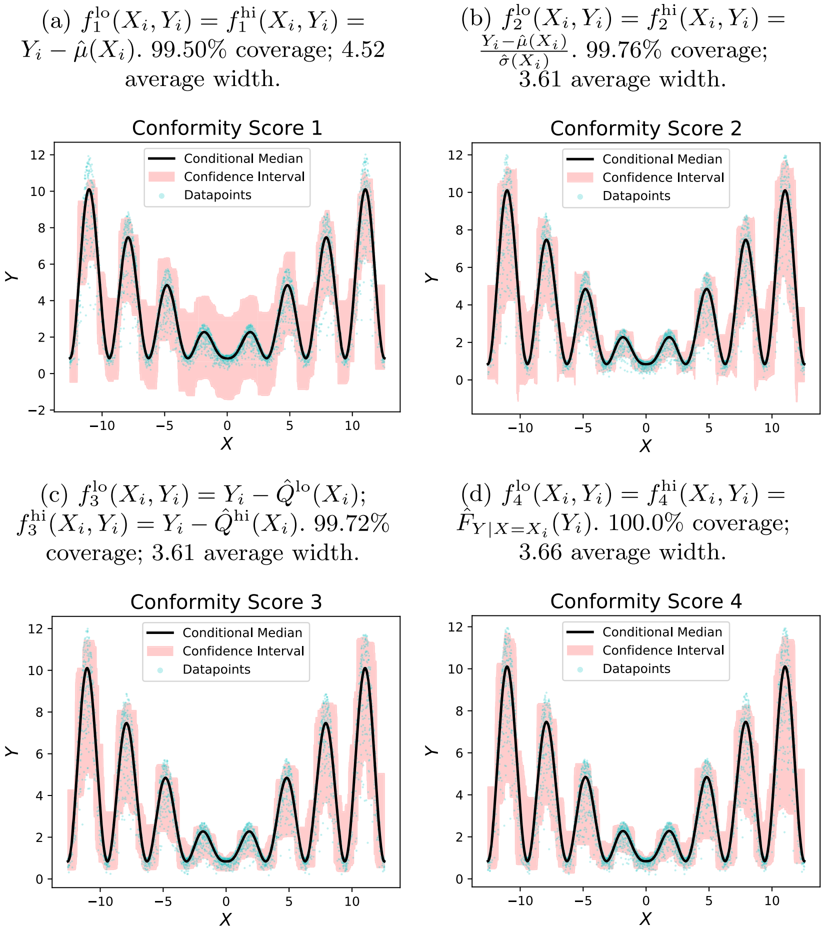
FIG 3. *For each distribution and conformity score, we calculate: average coverage (AC), an estimate of* $\mathbb{P}\{Median(Y_{n+1}|X_{n+1}) \in \hat{C}_n(X_{n+1})\}$*; standard deviation of average coverage (SDAC), an estimation of* $Var(\mathbb{P}\{Median(Y_{n+1}|X_{n+1}) \in \hat{C}_n(X_{n+1})|\mathcal{D}\})^{1/2}$*, where* $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$*; minimum conditional coverage (MCC), an estimate of* $\min_x \mathbb{P}\{Median(Y|X = x) \in \hat{C}_n(x)\}$*; average width (AW), an estimate of* $\mathbb{E}[len(\hat{C}_n(X_{n+1}))]$*; and standard deviation of average width (SDAW), an estimate of* $Var(\mathbb{E}[len(\hat{C}_n(X_{n+1}))|\mathcal{D}])^{1/2}$*. Estimations are averaged over 500 trials.* $1 - \alpha = 0.9$ *for all trials.*

Comparing the QRF algorithm with Algorithm 2, we see that while the widths are significantly lower for all three distributions, the coverage for Distribution 3 is much less than $1 - \alpha$. In particular, compared against Conformity Score 3, which has the same framework but includes a constant buffer on both ends due to the calibration set, we see how much extra width Algorithm 2 adds in the calibration step for Conformity Score 3.

Looking first at rates of coverage, we see that all scores have coverage much greater than $1 - \alpha$ on Distributions 1 and 2. However, Distribution 3 is a case where all scores have near-identical rates of coverage at just about $1 - \alpha$. Further investigation into the confidence intervals produced for Distribution 3 suggests that the algorithms are often failing to capture the conditional median when its absolute value is almost exactly 1.

The minimum conditional coverage on Distributions 1 and 3 is near 0 for each conformity score. Interestingly, the scores with the worst minimum conditional coverage on Distribution 1 have the relative best minimum conditional coverage on Distribution 3, and vice versa. Scores 1, 3, and 4 have a minimum conditional coverage greater than $1 - \alpha$ on Distribution 2, implying that these scores achieve point-wise conditional coverage.

Regarding interval width, Score 1 performs significantly worse on Distributions 1 and 2 than all other scores; meanwhile, the 3 other scores have roughly



(a) $f_1^{\text{lo}}(X_i, Y_i) = f_1^{\text{hi}}(X_i, Y_i) = Y_i - \hat{\mu}(X_i)$. 99.50% coverage; 4.52 average width.

(b) $f_2^{\text{lo}}(X_i, Y_i) = f_2^{\text{hi}}(X_i, Y_i) = \frac{Y_i - \hat{\mu}(X_i)}{\hat{\sigma}(X_i)}$. 99.76% coverage; 3.61 average width.

(c) $f_3^{\text{lo}}(X_i, Y_i) = Y_i - \hat{Q}^{\text{lo}}(X_i)$; $f_3^{\text{hi}}(X_i, Y_i) = Y_i - \hat{Q}^{\text{hi}}(X_i)$. 99.72% coverage; 3.61 average width.

(d) $f_4^{\text{lo}}(X_i, Y_i) = f_4^{\text{hi}}(X_i, Y_i) = \hat{F}_{Y|X=X_i}(Y_i)$. 100.0% coverage; 3.66 average width.

FIG 4. *Confidence intervals (pink regions) from one trial for each conformity score on Distribution 2. Note that all scores result in a coverage well over* $1 - \alpha = 0.9$.

equal average widths. On Distribution 3, Scores 3 and 4 produce intervals with significantly less width than Scores 1 and 2.

Overall, we see that Score 1 is significantly worse than the other scores on distributions with a wide range in conditional variance; Scores 3 and 4 behave very similarly on all distributions and perform slightly better than Score 2 on Distribution 3.

5. Discussion

This paper introduced two algorithms for capturing the conditional median of a datapoint within the distribution-free setting, as well as a particular distribution where the performance of these algorithms was sharp. Our lower bounds prove that in the distribution-free setting, conditional median inference is fundamentally as difficult as prediction itself, thereby setting a concrete limit to how well any median inference algorithm can ever perform. We also showed that any predictive algorithm can be used as a median algorithm at a different coverage level, suggesting that the two problems are near-equivalent.

5.1. Takeaways

A few observations may prove useful. For one, distributions such as P^δ from Section 2.3 and P_3 from Section 4 will likely show up again. Because each distribution is a mixture of two disjoint distributions with roughly equal weights, it is hard to identify which half contains the median. It is likely that similar distributions will show up as the performance-limiting distribution for distribution-free parameter inference. Further, the proof technique of sampling a large finite number of datapoints and then marginalizing (Section 3) is similar to those in [2] and [3], pointing out to possible future use. Lastly, our results and those of [2] indicate that the value of conditional parameters cannot be known with higher accuracy than the values of future samples.

5.2. Further work

We hope that this paper motivates further work on conditional parameter inference. We see three immediate potential avenues:

- One direction is to extend our methods to study other conditional parameters similar to the conditional median. For example, the smoothed conditional median, equal to the conditional median convolved with a kernel, may be easier to infer than the conditional median. This parameter would allow for smarter inference in the case of smooth distributions without making smoothness a requirement for inference. Similarly, the truncated mean and other measures of central tendency may be amenable to model-free inference and analysis, as may the conditional interquartile range and other robust measures of scale.
- Another direction is to get tighter bands by imposing mild shape constraints on the conditional median function. For instance, if we know that $\mathrm{Median}(Y|X = x)$ is convex, then the results from Section 3 no longer apply. Similarly, assuming that $\mathrm{Median}(Y|X = x)$ is decreasing in x or Lipschitz would yield intervals with vanishing widths in the limit of large samples. For instance, when predicting economic damages caused by tornadoes using wind speed as a covariate, one may assume that the median damage is nondecreasing as wind speed increases.
- A third subject of study is creating *full conformal inference* methods based off of our split conformal algorithms. Unlike split conformal inference, the full conformal method does not rely on splitting the dataset into a fitting half and a ranking half; instead, it calculates the conformity of

a potential datapoint (X_{n+1}, y) to the full dataset $\mathcal{D}$ and includes y in its confidence region only if (X_{n+1}, y) is similar enough to the observed datapoints. The study of full conformal inference has grown alongside that of split conformal inference; the method can be seen in [20], [18], and [10]. Standard full conformal algorithms do not guarantee coverage of the conditional median; however, there may exist modifications similar to locally nondecreasing conformity scores that result in a full conformal algorithm that captures the conditional median.

Appendix A: Theorem proofs

A.1. Proof of Theorem 1

Theorem 1 follows directly from Theorem 6. In particular, [20] shows that Algorithm 1 satisfies $(1-\alpha/2)$-Predictive. Because Algorithm 1 always outputs a confidence interval, we have that Algorithm 1 satisfies $(1-\alpha)$-Median by Theorem 6.

A.2. Proof of Theorem 2

We can prove Theorem 2 from the extension of Theorem 6 described in Remark 3.3. We know that Algorithm 2 always outputs a confidence interval because it is the intersection of 2 confidence intervals. Define $E^{\text{lo}}_{n+1} = f^{\text{lo}}(X_{n+1}, Y_{n+1}))$ and $E^{\text{hi}}_{n+1} = f^{\text{hi}}(X_{n+1}, Y_{n+1}))$. We can bound the probability of Y_{n+1} being at least the lower bound by $\mathbb{P}\{E^{\text{lo}}_{n+1} \geq Q^{\text{lo}}_{rq}(E)\}$ and bound the probability of Y_{n+1} being at most the upper bound by $\mathbb{P}\{E^{\text{hi}}_{n+1} \leq Q^{\text{hi}}_{1-s(1-q)}(E)\}$.

$Q^{\text{lo}}_{rq}(E)$ is defined as the $rq(1+1/n_2) - 1/n_2$-th quantile of $\{E^{\text{lo}}_i : i \in \mathcal{I}_2\}$, which is equal to the $\lceil n_2(rq(1+1/n_2) - 1/n_2)\rceil = \lceil rq(n_2+1) - 1\rceil$ smallest value of $\{E^{\text{lo}}_i : i \in \mathcal{I}_2\}$. Then, because $\{E^{\text{lo}}_i : i \in \mathcal{I}_2\} \cup \{E^{\text{lo}}_{n+1}\}$ are exchangeable, as f^{lo} is only fit on $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$, we have that the distribution of $|\{E^{\text{lo}}_i < E^{\text{lo}}_{n+1} : i \in \mathcal{I}_2\}|$ is bounded above by the uniform distribution on $\{0, 1, \ldots, n_2\}$. Therefore,

$$\begin{aligned}
\mathbb{P}\{E^{\text{lo}}_{n+1} \geq Q^{\text{lo}}_{rq}(E)\} \geq & \sum_{\lceil rq(n_2+1)-1\rceil+1}^{n_2} \frac{1}{n_2+1} \\
= & \frac{n_2 + 1 - \lceil rq(n_2+1)\rceil}{n_2+1} \\
\geq & \frac{(n_2+1)(1-rq)}{n_2+1} \\
= & 1 - rq.
\end{aligned}$$

Similarly, $Q^{\text{hi}}_{1-s(1-q)}(E)$ is defined as the $(1-s(1-q))(1+1/n_2)$-th quantile of $\{E^{\text{hi}}_i : i \in \mathcal{I}_2\}$, which equals the $\lceil n_2(1-s(1-q))(1+1/n_2)\rceil = \lceil (1-s(1-q))(n_2+$

1)⌉ smallest value of $\{E_i^{\text{hi}} : i \in \mathcal{I}_2\}$. Then, because $\{E_i^{\text{hi}} : i \in \mathcal{I}_2\} \cup \{E_{n+1}^{\text{hi}}\}$ are exchangeable, we have that the distribution of $|\{E_i^{\text{hi}} \le E_{n+1}^{\text{hi}} : i \in \mathcal{I}_2\}|$ is bounded below by the uniform distribution on $\{0, 1, \ldots, n_2\}$. Therefore,

$$\begin{aligned}\mathbb{P}\{E_{n+1}^{\text{hi}} \le Q_{1-s(1-q)}^{\text{hi}}(E)\} \ge &\sum_{0}^{\lceil(1-s(1-q))(n_2+1)\rceil-1} \frac{1}{n_2+1} \\ =&\frac{\lceil(1-s(1-q))(n_2+1)\rceil}{n_2+1} \\ \ge& 1 - s(1-q).\end{aligned}$$

Combining these two results and applying the extension of Theorem 6 tells us that Algorithm 2 satisfies $(1-\alpha, q)$-Quantile as desired.

A.3. Proof of Theorem 3

We show that given ϵ, there exists δ and N such that for all $n > N$, running Algorithm 1 on P^δ with our chosen $\hat{\mu}$ results in a confidence interval that contains the conditional median with probability at most $1 - \alpha + \epsilon$. Our approch is similar to that in Appendix B.1; however, we apply the inequalities in the opposite directions and use some analysis in order to get an upper bound as opposed to a lower bound.

First, note that $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$ is irrelevant to our algorithm, as $\hat{\mu}$ is set to be the zero function. Then, for each $i \in \mathcal{I}_2$, $E_i = |Y_i|$. Thus, $Q_{1-\alpha/2}(E)$ is the $(1-\alpha/2)(1+1/n_2)$-th empirical quantile of $\{|Y_i| : i \in \mathcal{I}_2\}$, and our confidence interval is $\hat{C}_n(X_{n+1}) = [-Q_{1-\alpha/2}(E), Q_{1-\alpha/2}(E)]$. Because the parameter we want to cover is $\text{Median}(Y|X = X_{n+1}) = X_{n+1}$,

$$\text{Median}(Y_{n+1}|X_{n+1}) \in \hat{C}_n(X_{n+1}) \text{ if and only if } |X_{n+1}| \le Q_{1-\alpha/2}(\{|Y_i| : i \in \mathcal{I}_2\}).$$

Define $M_R = \#\{i \in \mathcal{I}_2 : |X_i| \ge |X_{n+1}|\}$ and $M_E = \#\{i \in \mathcal{I}_2 : |Y_i| \ge |X_{n+1}|\}$. Now, note that $Q_{1-\alpha/2}(\{|Y_i| : i \in \mathcal{I}_2\})$ is the $\lceil n_2(1-\alpha/2)(1+1/n_2)\rceil = \lceil(1-\alpha/2)(n_2+1)\rceil$ smallest value of $\{|Y_i| : i \in \mathcal{I}_2\}$, which equals the $n_2 + 1 - \lceil(1-\alpha/2)(n_2+1)\rceil$ largest value. Letting $m = n_2 + 1 - \lceil(1-\alpha/2)(n_2+1)\rceil$,

$$|X_{n+1}| \le Q_{1-\alpha/2}(\{|Y_i| : i \in \mathcal{I}_2\}) \text{ if and only if } M_E \ge m.$$

We now build up the following two lemmas.

Lemma A.1. *For all* $0 \le M \le n_2$,

$$\mathbb{P}\{M_R = M\} = \frac{1}{n_2+1}.$$

Proof. This holds from the fact that the values $\{|X_i| : i \in \mathcal{I}_2\} \cup \{|X_{n+1}|\}$ are i.i.d. and have a distribution over $[0, 0.5]$ with no point masses. As a result, M_R is uniformly distributed over $\{0, 1, \ldots, n_2\}$. □

Lemma A.2. $M_E|M_R \sim \mathrm{Binom}(M_R, 0.5+\delta)$.

Proof. First, note that for all $i \in \mathcal{I}_2$, if $|X_i| < |X_{n+1}|$, then $\mathbb{P}\{|Y_i| \geq |X_{n+1}|\} = 0$. This is because if $|X_i| < |X_{n+1}|$, then $|Y_i| \leq |X_i| < |X_{n+1}|$. Additionally, if $|X_i| \geq |X_{n+1}|$, then $\mathbb{P}\{|Y_i| \geq |X_{n+1}|\} = 0.5+\delta$. This is due to the fact that if $|X_i| \geq |X_{n+1}|$, we have that $|Y_i| = 0$ with probability $0.5-\delta$ and $|Y_i| = |X_i|$ with probability $0.5+\delta$. With probability 1, $|X_{n+1}| > 0$. Therefore, $|Y_i| \geq |X_{n+1}|$ if and only if $|Y_i| = |X_i|$, which occurs with probability $0.5+\delta$.

Furthermore, the events $\{|Y_i| = |X_i|\}$ are mutually independent for all $i \in \mathcal{I}_2$ (the pairs (X_i, Y_i) are i.i.d.). Then, since

$$M_E = \sum_{i\in\mathcal{I}_2} \mathbf{1}[|Y_i| \geq |X_{n+1}|] = \sum_{i\in\mathcal{I}_2} \mathbf{1}[|X_i| \geq |X_{n+1}|]\mathbf{1}[|Y_i| = |X_i|]$$

and each term $\mathbf{1}[|Y_i| = |X_i|]$ is i.i.d. Bernoulli with probability $0.5+\delta$, and $M_R = \sum_{i\in\mathcal{I}_2} \mathbf{1}[|X_i| \geq |X_{n+1}|]$, the result follows. □

We now apply our two lemmas:

$$\begin{aligned}\mathbb{P}\{M_E \geq m\} =& \sum_{j=0}^{n_2} \mathbb{P}\{M_R = j\}\mathbb{P}\{M_E \geq m | M_R = j\} \\ =& \frac{1}{n_2+1}\sum_{j=0}^{n_2}\sum_{k=m}^{j} \mathbb{P}\{M_E \geq m | M_R = j\} \\ =& \frac{1}{n_2+1}\sum_{j=0}^{n_2}\sum_{k=m}^{j} \binom{j}{k}(0.5+\delta)^k(0.5-\delta)^{j-k};\end{aligned}$$

the second equality follows from Lemma A.1, and the last from Lemma A.2. Now, by applying the same train of logic as used in Appendices B.1 and B.2,

$$\begin{aligned}\mathbb{P}\{M_E \geq m\} =& \frac{1}{n_2+1}\sum_{j=0}^{n_2}\left(1 - \sum_{k=0}^{m-1}\binom{j}{k}(0.5+\delta)^k(0.5-\delta)^{j-k}\right) \\ =& 1 - \frac{1}{n_2+1}\sum_{j=0}^{n_2}\sum_{k=0}^{m-1}\binom{j}{k}(0.5+\delta)^k(0.5-\delta)^{j-k} \\ =& 1 - \frac{1}{n_2+1}\sum_{k=0}^{m-1}\left(\frac{0.5+\delta}{0.5-\delta}\right)^k \sum_{j=0}^{n_2}\binom{j}{k}(0.5-\delta)^j \\ =& 1 - \frac{1}{n_2+1}\sum_{k=0}^{m-1}\left(\frac{0.5+\delta}{0.5-\delta}\right)^k \left(\frac{(0.5-\delta)^k}{(0.5+\delta)^{k+1}} - \sum_{j=n_2+1}^{\infty}\binom{j}{k}(0.5-\delta)^j\right),\end{aligned}$$

where the last equality is from evaluating the generating function $G(\binom{t}{k}; z) = \sum_{t=0}^{\infty}\binom{t}{k}z^t$ at $z = 0.5-\delta$.

Finally, in order to bring this into a coherent bound, we expand the equation to bring out the $1-\alpha$ term and isolate the remainder, which we can then show goes to 0:

$$\begin{aligned}\mathbb{P}\{M_E \geq m\} =& 1 - \frac{1}{n_2+1}\sum_{k=0}^{m-1}\Big(\frac{1}{0.5+\delta} - \Big(\frac{0.5+\delta}{0.5-\delta}\Big)^k \sum_{j=n_2+1}^{\infty}\binom{j}{k}(0.5-\delta)^j\Big)\\ =& 1 - \frac{m}{n_2+1}\cdot\frac{1}{0.5+\delta} + \sum_{k=0}^{m-1}\Big(\frac{0.5+\delta}{0.5-\delta}\Big)^k \sum_{j=n_2+1}^{\infty}\binom{j}{k}(0.5-\delta)^j\\ \leq& 1 - \frac{m}{n_2+1}\cdot\frac{1}{0.5+\delta}\\ &+ \sum_{k=0}^{m-1}\Big(\frac{0.5+\delta}{0.5-\delta}\Big)^k\binom{n_2+1}{k}(0.5-\delta)^{n_2+1}\frac{1}{1-(0.5-\delta)\frac{n_2+1}{n_2+1-k}},\end{aligned}$$

where the inequality arises from upper bounding the summation $\sum_{j=n_2+1}^{\infty}\binom{j}{k}\times(0.5-\delta)^j$ by

$$\binom{n_2+1}{k}(0.5-\delta)^{n_2+1}\sum_{j=0}^{\infty}\Big(\frac{n_2+1}{n_2+1-k}(0.5-\delta)\Big)^j$$

using the maximum ratio of consecutive terms. Applying $m \leq (n_2+1)\alpha/2$ twice gives

$$\begin{aligned}\mathbb{P}\{M_E \geq m\} =& 1 - \frac{m}{n_2+1}\cdot\frac{1}{0.5+\delta}\\ &+ \Big(\frac{0.5+\delta}{0.5-\delta}\Big)^{m-1}\frac{1}{1-(0.5-\delta)\frac{n_2+1}{n_2+1-m}}\sum_{k=0}^{m-1}\binom{n_2+1}{k}(0.5-\delta)^{n_2+1}\\ \leq& 1 - \frac{m}{n_2+1}\cdot\frac{1}{0.5+\delta}\\ &+ \Big(\frac{0.5+\delta}{0.5-\delta}\Big)^{m-1}\Big(2+\frac{\alpha}{1-\alpha}\Big)\sum_{k=0}^{m-1}\binom{n_2+1}{k}(0.5-\delta)^{n_2+1}\\ \leq& 1 - \alpha + \frac{2\alpha\delta}{1+2\delta}\\ &+ \Big(2+\frac{\alpha}{1-\alpha}\Big)\Big(\frac{0.5+\delta}{0.5-\delta}\Big)^{(n_2+1)\alpha/2}\cdot\frac{\sum_{k=0}^{\lfloor(n_2+1)\alpha/2\rfloor}\binom{n_2+1}{k}}{2^{n_2+1}}.\end{aligned}$$

Because $\alpha < 1$,

$$\frac{\sum_{k=0}^{\lfloor(n_2+1)\alpha/2\rfloor}\binom{n_2+1}{k}}{2^{n_2+1}} \to 0 \quad \text{as} \quad n_2 \to \infty;$$

this is due to the fact that

$$\frac{\sum_{k=0}^{n_2+1}\binom{n_2+1}{k}}{2^{n_2+1}} = 1$$

and that the standard deviation of $\mathrm{Binom}(n_2, 0.5)$ is $\mathcal{O}(\sqrt{n_2})$, meaning that

$$\frac{\sum_{k=\lfloor(n_2+1)\alpha/2\rfloor+1}^{n_2+1-\lfloor(n_2+1)\alpha/2\rfloor}\binom{n_2+1}{k}}{2^{n_2+1}} \to 1.$$

Furthermore, as $\mathrm{Binom}(n_2, 0.5)$ approaches a normal distribution as $n_2 \to \infty$ and $\Phi(-c\sqrt{n_2})$ is $\mathcal{O}(d^{-n_2})$ for some $d > 1$, for small enough δ,

$$\left(\frac{0.5+\delta}{0.5-\delta}\right)^{(n_2+1)\alpha/2} \frac{\sum_{k=0}^{\lfloor(n_2+1)\alpha/2\rfloor}\binom{n_2+1}{k}}{2^{n_2+1}} \to 0 \quad \text{as} \quad n_2 \to \infty.$$

Thus, we can pick D and N such that for all $\delta < D$ and $n \geq N$,

$$\left(\frac{0.5+\delta}{0.5-\delta}\right)^{(n_2+1)\alpha/2} \frac{\sum_{k=0}^{\lfloor(n_2+1)\alpha/2\rfloor}\binom{n_2+1}{k}}{2^{n_2+1}} \leq \epsilon/2,$$

noting that $n_2 = n/2$. Then, setting $\delta = \min\{\frac{\epsilon}{4\alpha-2\epsilon}, D\}$ and $\frac{2\alpha\delta}{1+2\delta} \leq \epsilon/2$ yields

$$\mathbb{P}\big\{M_E \geq m\big\} \leq 1-\alpha+\epsilon/2+\epsilon/2 = 1-\alpha+\epsilon.$$

This says that the probability $\mathbb{P}\{M_E \geq m\}$ of the confidence interval containing the conditional median is at most $1-\alpha+\epsilon$.

A.4. Proof of Theorem 4

We show that given ϵ there exists c, N, and $n_1 + n_2 = n$ for all $n > N$ such that running Algorithm 1 on $n > N$ datapoints from an arbitrary distribution P with regression function $\hat{\mu}_c$ and split sizes $n_1 + n_2 = n$ results in a finite confidence interval that contains the conditional median with probability at least $1 - \alpha/2 - \epsilon$.

For each x in the support of P, define $m(x) = \mathrm{Median}(Y|X = x)$ and recall that $M = \max\limits_{i\in\mathcal{I}_1} |Y_i|$. We begin with two lemmas.

Lemma A.3. *For all $i \in \mathcal{I}_2$, $\mathbb{P}\{|Y_i| \leq M\} \geq 1 - \frac{1}{n_1+1}$.*

Proof. This results from the fact that $|Y_i|$ is exchangeable with $|Y_j|$ for all $j \in \mathcal{I}_1$; thus, the probability that $|Y_i|$ is the unique maximum of the set $\{Y_j : j \in \mathcal{I}_1 \cup \{i\}\}$ is bounded above by $\frac{1}{n_1+1}$. Taking the complement yields the desired result. □

Lemma A.4. *For all $i \in \mathcal{I}_2 \cup \{n+1\}$, $\mathbb{P}\{|m(X_i)| \leq M\} \geq 1 - \frac{2}{n_1+1}$.*

Proof. Note that $|m(X_i)|$ is exchangeable with $|m(X_j)|$ for all $j \in \mathcal{I}_1$. Letting $M_R = \#\{|m(X_j)| \geq |m(X_i)| : j \in \mathcal{I}_1\}$, exchangeability gives that the CDF of M_R is bounded below by the CDF of the uniform distribution over $\{0, 1, \ldots, n_1\}$. For each $j \in \mathcal{I}_1$, the event $\{|Y_j| \geq |m(X_j)|\}$ occurs with probability at least 1/2 by definition of the median; moreover, these events are mutually independent. Therefore, if we condition on M_R, we have that

$$\mathbb{P}\{|m(X_i)| > \max_{j \in \mathcal{I}_1} |Y_j| \big| M_R = k\} \leq \prod_{\substack{j \in \mathcal{I}_1 \\ |m(X_j)| \geq |m(X_i)|}} \mathbb{P}\{|Y_j| < |m(X_j)|\} \leq 2^{-k}.$$

Putting this together, we see that

$$\begin{aligned}\mathbb{P}\{|m(X_i)| > \max_{j \in \mathcal{I}_1} |Y_j|\} &= \sum_{k=0}^{n_1} \mathbb{P}\{M_R = k\} \mathbb{P}\{|m(X_i)| > \max_{j \in \mathcal{I}_1} |Y_j| \big| M_R = k\} \\ &\leq \sum_{k=0}^{n_1} \mathbb{P}\{M_R = k\} \frac{1}{2^k} \\ &\leq \frac{1}{n_1 + 1} \sum_{k=0}^{n_1} \frac{1}{2^k} \\ &\leq \frac{2}{n_1 + 1}.\end{aligned}$$

Taking the complement yields the desired result. □

Let A be the event $\{|Y_i| \leq M$ for all $i \in \mathcal{I}_2$ and $|m(X_i)| \leq M$ for all $i \in \mathcal{I}_2 \cup \{n+1\}\}$. By Lemmas A.3 and A.4, $\mathbb{P}\{A\} \geq 1 - \frac{3n_2 + 2}{n_1 + 1}$. Select $N = \left\lfloor \frac{12/\alpha + 10}{\epsilon} \right\rfloor + \lfloor 2/\alpha \rfloor + 1$, and for all $n > N$, set $n_2 = \lfloor 2/\alpha \rfloor + 1$ and $n_1 = n - n_2$. As a result, we have that $1/n_2 < \alpha/2$ and $\frac{3n_2 + 2}{n_1 + 1} < \epsilon/2$, so $\mathbb{P}\{A\} \geq 1 - \epsilon/2$.

Next, let B be the event $\{|\hat{\mu}_c(X_{i_1}) - \hat{\mu}_c(X_{i_2})| > 2M$ for all $i_1 \neq i_2 \in \mathcal{I}_2 \cup \{n+1\}\}$. Note that $\lim_{c \to \infty} \mathbb{P}\{B\} = 1$ by definition of $\hat{\mu}_c$. Select c such that $\mathbb{P}\{B\} \geq 1 - \epsilon/2$. By the union bound, $\mathbb{P}\{A \cap B\} \geq 1 - \epsilon$.

Lemma A.5. *On the event $A \cap B$, for all $i \in \mathcal{I}_2$,*

$$|m(X_i) - \hat{\mu}_c(X_i)| \geq |m(X_{n+1}) - \hat{\mu}_c(X_{n+1})| \quad \textit{if and only if} \quad |Y_i - \hat{\mu}_c(X_i)| \geq |m(X_{n+1}) - \hat{\mu}_c(X_{n+1})|.$$

Proof. Notice that on the event $A \cap B$, $|m(X_i) - \hat{\mu}_c(X_i)|, |Y_i - \hat{\mu}_c(X_i)| \in [|\hat{\mu}_c(X_i)| - M, |\hat{\mu}_c(X_i)| + M]$. This holds because $|m(X_i)|, |Y_i| \leq M$ on the event A. Similarly, $|m(X_{n+1}) - \hat{\mu}_c(X_{n+1})| \in [|\hat{\mu}_c(X_{n+1})| - M, |\hat{\mu}_c(X_{n+1})| + M]$. These two intervals both have length $2M$, but their centers are at a distance greater than $2M$ on the event B, meaning that the intervals are disjoint. Therefore, $|m(X_i) - \hat{\mu}_c(X_i)| \geq |m(X_{n+1}) - \hat{\mu}_c(X_{n+1})|$ implies that all elements of the first interval are greater than all elements of the second, so $|Y_i - \hat{\mu}_c(X_i)| \geq |m(X_{n+1}) - \hat{\mu}_c(X_{n+1})|$; similarly, $|Y_i - \hat{\mu}_c(X_i)| \geq |m(X_{n+1}) - \hat{\mu}_c(X_{n+1})|$ also implies that all elements of the first interval are greater than all elements of the second, so $|m(X_i) - \hat{\mu}_c(X_i)| \geq |m(X_{n+1}) - \hat{\mu}_c(X_{n+1})|$. □

Looking at Algorithm 1, we have that $m(X_{n+1}) \in \hat{C}_n(X_{n+1})$ if $|m(X_{n+1}) - \hat{\mu}_c(X_{n+1})| \leq Q_{1-\alpha/2}(E)$, where $E_i = |Y_i - \hat{\mu}_c(X_i)|$ for all $i \in \mathcal{I}_2$. Because $1/n_2 < \alpha/2$, $Q_{1-\alpha/2}(E)$ is finite and thus the confidence interval is bounded. By Lemma A.5, on the event $A \cap B$, $|m(X_{n+1}) - \hat{\mu}_c(X_{n+1})| \leq Q_{1-\alpha/2}(E)$ if and only if $|m(X_{n+1}) - \hat{\mu}_c(X_{n+1})| \leq Q_{1-\alpha/2}(F)$, where $F_i = |m(X_i) - \hat{\mu}_c(X_i)|$ for all $i \in \mathcal{I}_2$.

Define C to be the event $\{|m(X_{n+1}) - \hat{\mu}_c(X_{n+1})| \leq Q_{1-\alpha/2}(F)\}$. We have just shown that on the event $A \cap B \cap C$, we have $m(X_{n+1}) \in \hat{C}_n(X_{n+1})$. Additionally, because the elements of $\{|m(X_i) - \hat{\mu}_c(X_i)| : i \in \mathcal{I}_2 \cup \{n+1\}\}$ are exchangeable, we have that $\mathbb{P}\{C\} \geq 1 - \alpha/2$. Then, by the union bound,

$$\mathbb{P}\{m(X_{n+1}) \in \hat{C}_n(X_{n+1})\} \geq \mathbb{P}\{A \cap B \cap C\} \geq 1 - \alpha/2 - \epsilon,$$

proving the desired result.

A.5. Proving extensions of Theorem 6

We show the following two results:

Let $\hat{C}_n(x) = [\hat{L}_n(x), \hat{H}_n(x)]$ be any algorithm that only outputs confidence intervals and satisfies $\mathbb{P}\{Y_{n+1} \geq \hat{L}_n(X_{n+1})\} \geq 1 - rq$ and $\mathbb{P}\{Y_{n+1} \leq \hat{H}_n(X_{n+1})\} \geq 1 - s(1-q)$ for some $r + s = \alpha$. Then, $\hat{C}_n$ satisfies $(1-\alpha, q)$-Quantile. Secondly, if $\hat{C}_n$ only outputs confidence intervals and satisfies $(1 - \min\{q, 1-q\}\alpha)$-Predictive, then it satisfies $(1-\alpha, q)$-Quantile.

Consider a distribution P, and for all $x \in \mathbb{R}^d$ in the support of P, let $q(x)$ be $\text{Quantile}_q(Y|X = x)$. We prove the first result in two parts.

Conditioning on whether or not $q(X_{n+1})$ is greater than or equal to $\hat{L}(X_{n+1})$, note that

$$\begin{aligned}
1 - rq \leq& \mathbb{P}\{Y_{n+1} \geq \hat{L}_n(X_{n+1})\} \\
=& \mathbb{P}\{Y_{n+1} \geq \hat{L}_n(X_{n+1}) | q(X_{n+1}) \geq \hat{L}_n(X_{n+1})\} \mathbb{P}\{q(X_{n+1}) \geq \hat{L}_n(X_{n+1})\} \\
&+ \mathbb{P}\{Y_{n+1} \geq \hat{L}_n(X_{n+1}) | q(X_{n+1}) < \hat{L}_n(X_{n+1})\} \mathbb{P}\{q(X_{n+1}) < \hat{L}_n(X_{n+1})\} \\
\leq& \mathbb{P}\{q(X_{n+1}) \geq \hat{L}_n(X_{n+1})\} + (1-q)\mathbb{P}\{q(X_{n+1}) < \hat{L}_n(X_{n+1})\} \\
=& (1-q) + q \cdot \mathbb{P}\{q(X_{n+1}) \geq \hat{L}_n(X_{n+1})\}
\end{aligned}$$

where we use the fact that if $q(X_{n+1}) < \hat{L}_n(X_{n+1})$, at most $1 - q$ of the conditional distribution of $Y|X = X_{n+1}$ can be at least $\hat{L}_n(X_{n+1})$. Subtracting $1 - q$ from both sides and dividing by q tells us that $1 - r \leq \mathbb{P}\{q(X_{n+1}) \geq \hat{L}_n(X_{n+1})\}$.

Similarly,

$$\begin{aligned}
&1 - s(1-q) \\
\leq& \mathbb{P}\{Y_{n+1} \leq \hat{H}_n(X_{n+1})\} \\
=& \mathbb{P}\{Y_{n+1} \leq \hat{H}_n(X_{n+1}) | q(X_{n+1}) \leq \hat{H}_n(X_{n+1})\} \mathbb{P}\{q(X_{n+1}) \leq \hat{H}_n(X_{n+1})\} \\
&+ \mathbb{P}\{Y_{n+1} \leq \hat{H}_n(X_{n+1}) | q(X_{n+1}) > \hat{H}_n(X_{n+1})\} \mathbb{P}\{q(X_{n+1}) > \hat{H}_n(X_{n+1})\}
\end{aligned}$$

$$\begin{aligned}&\leq \mathbb{P}\{q(X_{n+1}) \leq \hat{H}_n(X_{n+1})\} + q \cdot \mathbb{P}\{q(X_{n+1}) > \hat{H}_n(X_{n+1})\}\\ &= q + (1-q)\mathbb{P}\{q(X_{n+1}) \leq \hat{H}_n(X_{n+1})\}.\end{aligned}$$

Subtracting q from both sides and dividing by $1-q$ tells us that $1 - s \leq \mathbb{P}\{q(X_{n+1}) \leq \hat{H}_n(X_{n+1})\}$.

Then, by the union bound, we have that $1-\alpha = 1-(r+s) \leq \mathbb{P}\{q(X_{n+1}) \geq \hat{L}_n(X_{n+1}) \cap q(X_{n+1}) \leq \hat{H}_n(X_{n+1})\} = \mathbb{P}\{q(X_{n+1} \in \hat{C}_n(X_{n+1})\}$, proving the first result.

We can prove the second result in the exact same fashion as the proof of Theorem 6. We have that

$$\begin{aligned}&1 - \min\{q, 1-q\}\alpha\\ &\leq \mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1})\}\\ &= \mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1}) | q(X_{n+1}) \in \hat{C}_n(X_{n+1})\}\mathbb{P}\{q(X_{n+1}) \in \hat{C}_n(X_{n+1})\}\\ &\quad + \mathbb{P}\{Y_{n+1} \in \hat{C}_n(X_{n+1}) | q(X_{n+1}) \notin \hat{C}_n(X_{n+1})\}\mathbb{P}\{q(X_{n+1}) \notin \hat{C}_n(X_{n+1})\}\\ &\leq \mathbb{P}\{q(X_{n+1}) \in \hat{C}_n(X_{n+1})\} + \max\{q, 1-q\}\mathbb{P}\{q(X_{n+1}) \notin \hat{C}_n(X_{n+1})\}\\ &= \max\{q, 1-q\} + \min\{q, 1-q\}\mathbb{P}\{q(X_{n+1}) \in \hat{C}_n(X_{n+1})\}.\end{aligned}$$

Subtracting $\max\{q, 1-q\}$ and dividing by $\min\{q, 1-q\}$ yields the desired result.

Appendix B: Additional results

B.1. *Alternate Proof of Theorem 1*

The proof of the theorem relies on two lemmas: the first establishes a connection between $\text{Median}(Y_{n+1}|X_{n+1})$ and $\text{Median}(Y|X = X_i)$'s using exchangeability, and the second gives us a relationship between $\text{Median}(Y|X = X_i)$'s and the Y_i's.

We begin with some notation. For all $x \in \mathbb{R}^d$ in the support of P, set $m(x) = \text{Median}(Y|X = x)$. Also, for $1 \leq i \leq n+1$, let $R(X_i) = |m(X_i) - \hat{\mu}(X_i)|$, and for $1 \leq i \leq n$ let $E(X_i) = |Y_i - \hat{\mu}(X_i)|$. Finally, put $M_R = \#\{i \in \mathcal{I}_2 : R(X_i) \geq R(X_{n+1})\}$ as the number of $i \in \mathcal{I}_2$ for which $R(X_i) \geq R(X_{n+1})$, and $M_E = \#\{i \in \mathcal{I}_2 : E(X_i) \geq R(X_{n+1})\}$ as the number of $i \in \mathcal{I}_2$ for which $E(X_i) \geq R(X_{n+1})$.

Note that $\text{Median}(Y_{n+1}|X_{n+1}) \in \hat{C}_n(X_{n+1}) = [\hat{\mu}(X_{n+1}) - Q_{1-\alpha/2}(E), \hat{\mu}(X_{n+1}) + Q_{1-\alpha/2}(E)]$ if and only if $|\text{Median}(Y_{n+1}|X_{n+1}) - \hat{\mu}(X_{n+1})| = R(X_{n+1})$ is at most $Q_{1-\alpha/2}(E)$. Thus, we must study the value of $R(X_{n+1})$ in relation to the elements of $\{E(X_i) : i \in \mathcal{I}_2\}$.

The first lemma relates $R(X_{n+1})$ to the other $R(X_i)$'s.

Lemma B.1. *For all* $0 \leq m \leq n_2$,

$$\mathbb{P}\big\{M_R \geq m\big\} \geq 1 - \frac{m}{n_2+1}.$$

Proof. Our statement follows from the fact that our samples are i.i.d. Because $\hat{\mu}$ is independent of (X_i, Y_i) for $i \in \mathcal{I}_2$ and independent of X_{n+1}, the values $\{R(X_i) : i \in \mathcal{I}_2\} \cup \{R(X_{n+1})\}$ are i.i.d. as well. Then, the probability of less than m values in $\{R(X_i) : i \in \mathcal{I}_2\}$ being at least $R(X_{n+1})$ is bounded above by $\frac{m}{|\mathcal{I}_2|+1}$, as the ordering of these values is uniformly random. Taking the complement of both sides gives the result. □

The second lemma gives a direct relationship between $E(X_i)$ and $R(X_i)$. (Note that the events $\{E(X_i) \geq R(X_i)\}$ below are mutually independent.)

Lemma B.2. *For all $i \in \mathcal{I}_2$,*

$$\mathbb{P}\big\{E(X_i) \geq R(X_i)\big\} \geq 1/2.$$

Proof. We see that $\mathbb{P}\{Y_i \geq m(X_i)\} \geq 1/2$ and $\mathbb{P}\{Y_i \leq m(X_i)\} \geq 1/2$ by the definition of the conditional median. Furthermore, the events $\{Y_i \geq m(X_i)\}$ and $\{Y_i \leq m(X_i)\}$ are independent of the events $\{m(X_i) \geq \hat{\mu}(X_i)\}$ and $\{m(X_i) \leq \hat{\mu}(X_i)\}$ given X_i, as $\hat{\mu}$ is a function of $\{(X_i, Y_i) : i \in \mathcal{I}_1\}$ and the datapoints are i.i.d. Then, conditioned on X_i, if $m(X_i) \geq \hat{\mu}(X_i)$, with probability $1/2$ we have that $m(X_i) \leq Y_i$, in which case $|m(X_i) - \hat{\mu}(X_i)| = m(X_i) - \hat{\mu}(X_i) \leq Y_i - \hat{\mu}(X_i) = |Y_i - \hat{\mu}(X_i)|$. Similarly, conditioned on X_i again, if $m(X_i) < \hat{\mu}(X_i)$, with probability $1/2$ we have that $m(X_i) \geq Y_i$, in which case $|m(X_i) - \hat{\mu}(X_i)| = \hat{\mu}(X_i) - m(X_i) \leq \hat{\mu}(X_i) - Y_i = |Y_i - \hat{\mu}(X_i)|$. The conclusion holds conditionally in both cases; marginalizing out X_i yields the desired result. □

We now study the number of datapoints obeying $E(X_i) \geq R(X_{n+1})$ by combining these lemmas together. Consider any $0 \leq m \leq n_2$. Note that by conditioning on M_R,

$$\begin{aligned}
\mathbb{P}\big\{M_E \geq m\big\} &= \sum_{j=0}^{n_2} \mathbb{P}\big\{M_R = j\big\}\mathbb{P}\big\{M_E \geq m | M_R = j\big\} \\
&\geq \sum_{j=0}^{n_2} \mathbb{P}\big\{M_R = j\big\} \sum_{k=m}^{j} \binom{j}{k} 2^{-j} \\
&\geq \sum_{j=0}^{n_2} \frac{1}{n_2+1} \sum_{k=m}^{j} \binom{j}{k} 2^{-j}.
\end{aligned}$$

The first inequality holds true by Lemma B.2, which implies that $\mathbb{P}\{M_E \geq m | M_R = j\}$ can be bounded below by the probability that $M \geq m$ for $M \sim \mathrm{Binom}(j, 0.5)$. The second inequality is due to Lemma B.1. We know that $\sum_{k=m}^{j} \binom{j}{k} 2^{-j}$ is a nondecreasing function of j; by Lemma B.1, the CDF of the distribution of M_R is lower bounded by the CDF of the uniform distribution over $\{0, 1, \ldots, n_2\}$, meaning that

$$\mathbb{E}_{M_R}\left[\sum_{k=m}^{M_R} \binom{M_R}{k} 2^{-M_R}\right] \geq \frac{1}{n_2+1} \sum_{j=0}^{n_2} \sum_{k=m}^{j} \binom{j}{k} 2^{-j}.$$

This gives

$$\begin{aligned}\sum_{j=0}^{n_2}\frac{1}{n_2+1}\sum_{k=m}^{j}\binom{j}{k}2^{-j} =&\sum_{j=0}^{n_2}\frac{1}{n_2+1}\Big(1-\sum_{k=0}^{m-1}\binom{j}{k}2^{-j}\Big)\\ =&1-\frac{1}{n_2+1}\sum_{j=0}^{n_2}\sum_{k=0}^{m-1}\binom{j}{k}2^{-j}\\ =&1-\frac{1}{n_2+1}\sum_{k=0}^{m-1}\sum_{j=0}^{n_2}\binom{j}{k}2^{-j}\\ \geq&1-\frac{1}{n_2+1}\sum_{k=0}^{m-1}\sum_{j=0}^{\infty}\binom{j}{k}2^{-j}\\ =&1-\frac{1}{n_2+1}\sum_{k=0}^{m-1}2\\ =&1-\frac{2m}{n_2+1}.\end{aligned}$$

The second-to-last equality comes from evaluating the generating function $G(\binom{t}{k};z)=\sum_{t=0}^{\infty}\binom{t}{k}z^t$ at $z=0.5$.

Putting it together, in Algorithm 1, $Q_{1-\alpha/2}(E)$ is set to be the $(1-\alpha/2)(1+1/n_2)$-th quantile of $\{E_i : i\in\mathcal{I}_2\}$. This is equal to the $\lceil n_2(1-\alpha/2)(1+1/n_2)\rceil = \lceil(1-\alpha/2)(n_2+1)\rceil$ smallest value of $\{E_i : i\in\mathcal{I}_2\}$. This means that if $M_E\geq n_2-\lceil(1-\alpha/2)(n_2+1)\rceil+1$, then $R(X_{n+1})$ will be at most the $n_2-\lceil(1-\alpha/2)(n_2+1)\rceil+1$ largest value of $\{E_i : i\in\mathcal{I}_2\}$, which is equal to the $\lceil(1-\alpha/2)(n_2+1)\rceil$ smallest value, or $Q_{1-\alpha/2}(E)$.

However, we have calculated a lower bound for the inverse CDF of M_E earlier. Substituting this in, we get that

$$\begin{aligned}\mathbb{P}\{R(X_{n+1})\leq Q_{1-\alpha/2}(E)\}\geq&\mathbb{P}\{M_E\geq n_2-\lceil(1-\alpha/2)(n_2+1)\rceil+1\}\\ \geq&\mathbb{P}\{M_E\geq n_2+1-(1-\alpha/2)(n_2+1)\}\\ =&\mathbb{P}\{M_E\geq\alpha/2(n_2+1)\}\\ \geq&1-\alpha\end{aligned}$$

by our previous calculation, completing our proof.

B.2. Alternate Proof of Theorem 2

Our approach is similar to that in Appendix B.1. The main difference in the proof arises from the fact that Algorithm 2 no longer uses the absolute value and uses two separate fitted functions, meaning that it is important to bound the probability of the confidence interval covering the desired value from both sides.

We begin with some definitions. For all $x\in\mathbb{R}^d$ in the support of P, let $q(x)$ be $\text{Quantile}_q(Y|X=x)$. For all $1\leq i\leq n+1$, define:

- $R^{\text{lo}}(X_i) = f^{\text{lo}}(X_i, q(X_i))$ and $R^{\text{hi}}(X_i) = f^{\text{hi}}(X_i, q(X_i))$.
- $E^{\text{lo}}(X_i) = f^{\text{lo}}(X_i, Y_i)$ and $E^{\text{hi}}(X_i) = f^{\text{hi}}(X_i, Y_i)$.
- $M_R^{\text{lo}} = \#\{i \in \mathcal{I}_2 : R^{\text{lo}}(X_i) \leq R^{\text{lo}}(X_{n+1})\}$ and $M_R^{\text{hi}} = \#\{i \in \mathcal{I}_2 : R^{\text{hi}}(X_i) \geq R^{\text{hi}}(X_{n+1})\}$ as the number of $i \in \mathcal{I}_2$ for which $R^{\text{lo}}(X_i) \leq R^{\text{lo}}(X_{n+1})$ and $R^{\text{hi}}(X_i) \geq R^{\text{hi}}(X_{n+1})$ respectively.
- $M_E^{\text{lo}} = \#\{i \in \mathcal{I}_2 : E^{\text{lo}}(X_i) \leq R^{\text{lo}}(X_{n+1})\}$ and $M_E^{\text{hi}} = \#\{i \in \mathcal{I}_2 : E^{\text{hi}}(X_i) \geq R^{\text{hi}}(X_{n+1})\}$ as the number of $i \in \mathcal{I}_2$ for which $E^{\text{lo}}(X_i) \leq R^{\text{lo}}(X_{n+1})$ and $E^{\text{hi}}(X_i) \geq R^{\text{hi}}(X_{n+1})$ respectively.

Now, note that $\text{Quantile}_q(Y|X = X_{n+1}) \in \hat{C}_n(X_{n+1})$ precisely when $f^{\text{lo}}(X_{n+1}, \text{Quantile}_q(Y|X = X_{n+1})) = R^{\text{lo}}(X_{n+1}) \geq Q_{rq}^{\text{lo}}$ and $f^{\text{hi}}(X_{n+1}, \text{Quantile}_q(Y|X = X_{n+1})) = R^{\text{hi}}(X_{n+1}) \leq Q_{1-s(1-q)}^{\text{hi}}$. As such, we proceed to develop two lemmas which extend those from Section B.1: the first helps us to understand the distributions of M_R^{hi} and M_R^{lo}, and the second studies the individual relationships between $E^{\text{lo}}(X_i)$ and $R^{\text{lo}}(X_i)$ and between $E^{\text{hi}}(X_i)$ and $R^{\text{hi}}(X_i)$. With both of these lemmas, we are able to bound the probability of each event $\{R^{\text{lo}}(X_{n+1}) \geq Q_{rq}^{\text{lo}}\}$ and $\{R^{\text{hi}}(X_{n+1}) \leq Q_{1-s(1-q)}^{\text{hi}}\}$.

Lemma B.3. *For all* $0 \leq m \leq n_2$,

$$\mathbb{P}\{M_R^{\text{lo}} \geq m\} \geq 1 - \frac{m}{n_2 + 1}$$

and

$$\mathbb{P}\{M_R^{\text{hi}} \geq m\} \geq 1 - \frac{m}{n_2 + 1}.$$

Proof. We prove the result for M_R^{hi}; the same approach holds for M_R^{lo}. Because f^{hi} is independent of (X_i, Y_i) for $i \in \mathcal{I}_2$ and independent of X_{n+1}, the values in $\{R^{\text{hi}}(X_i) : i \in \mathcal{I}_2\} \cup \{R^{\text{hi}}(X_{n+1})\}$ are i.i.d.. Thus, the probability of less than m values in $\{R^{\text{hi}}(X_i) : i \in \mathcal{I}_2\}$ being at least $R^{\text{hi}}(X_{n+1})$ is bounded above by $\frac{m}{|\mathcal{I}_2|+1}$, as the ordering of these values is uniformly random. Taking the complement of both sides establishes the claim. □

Lemma B.4. *For all* $i \in \mathcal{I}_2$,

$$\mathbb{P}\{E^{\text{lo}}(X_i) \leq R^{\text{lo}}(X_i)\} \geq q$$

and

$$\mathbb{P}\{E^{\text{hi}}(X_i) \geq R^{\text{hi}}(X_i)\} \geq 1 - q.$$

Proof. We see that $\mathbb{P}\{Y_i \leq q(X_i)\} \geq q$ and $\mathbb{P}\{Y_i \geq q(X_i)\} \geq 1 - q$ by the definition of the conditional quantile. Then, with probability at least q we have that $E^{\text{lo}}(X_i) = f^{\text{lo}}(X_i, Y_i) \leq f^{\text{lo}}(X_i, q(X_i)) = R^{\text{lo}}(X_i)$ by the definition of a locally nondecreasing conformity score. Similarly, with probability at least $1 - q$ we have that $E^{\text{hi}}(X_i) = f^{\text{hi}}(X_i, Y_i) \geq f^{\text{hi}}(X_i, q(X_i)) = R^{\text{hi}}(X_i)$, thereby concluding the proof. □

We now study the number of $i \in \mathcal{I}_2$ such that $E^{\text{hi}}(X_i) \geq R^{\text{hi}}(X_{n+1})$ by combining these two lemmas together. Consider any $0 \leq m \leq n_2$, and note that by conditioning on M_R^{hi}, we have

$$\begin{aligned}\mathbb{P}\{M_E^{\text{hi}} \geq m\} &= \sum_{j=0}^{n_2} \mathbb{P}\{M_R^{\text{hi}} = j\}\mathbb{P}\{M_E^{\text{hi}} \geq m | M_R^{\text{hi}} = j\} \\ &\geq \sum_{j=0}^{n_2} \mathbb{P}\{M_R^{\text{hi}} = j\} \sum_{k=m}^{j} \binom{j}{k}(1-q)^k q^{j-k} \\ &\geq \sum_{j=0}^{n_2} \frac{1}{n_2+1} \sum_{k=m}^{j} \binom{j}{k}(1-q)^k q^{j-k}.\end{aligned}$$

The first inequality holds because $\mathbb{P}\{M_E^{\text{hi}} \geq m | M_R^{\text{hi}} = j\}$ can be bounded below by the probability that $M \geq m$ for $M \sim \text{Binom}(j, 1-q)$ by Lemma B.4. The second inequality holds since the CDF M_R^{hi} is greater than or equal to the CDF of the uniform distribution over $\{0, 1, \ldots, n_2\}$ by Lemma B.3. Then, as $\sum_{k=m}^{j} \binom{j}{k}(1-q)^k q^{j-k}$ is a nondecreasing function of j, we have

$$\mathbb{E}_{M_R^{\text{hi}}}\left[\sum_{k=m}^{M_R^{\text{hi}}} \binom{M_R^{\text{hi}}}{k}(1-q)^k q^{M_R^{\text{hi}}-k}\right] \geq \frac{1}{n_2+1} \sum_{j=0}^{n_2} \sum_{k=m}^{j} \binom{j}{k}(1-q)^k q^{j-k}.$$

We now solve the summation, which gives

$$\begin{aligned}\sum_{j=0}^{n_2} \frac{1}{n_2+1} \sum_{k=m}^{j} \binom{j}{k}(1-q)^k q^{j-k} &= \sum_{j=0}^{n_2} \frac{1}{n_2+1}\left(1 - \sum_{k=0}^{m-1} \binom{j}{k}(1-q)^k q^{j-k}\right) \\ &= 1 - \frac{1}{n_2+1} \sum_{j=0}^{n_2} \sum_{k=0}^{m-1} \binom{j}{k}(1-q)^k q^{j-k} \\ &= 1 - \frac{1}{n_2+1} \sum_{k=0}^{m-1} \sum_{j=0}^{n_2} \binom{j}{k}(1-q)^k q^{j-k} \\ &\geq 1 - \frac{1}{n_2+1} \sum_{k=0}^{m-1} \left(\frac{1-q}{q}\right)^k \sum_{j=0}^{\infty} \binom{j}{k} q^j \\ &= 1 - \frac{1}{n_2+1} \sum_{k=0}^{m-1} \frac{1}{1-q} \\ &= 1 - \frac{1}{1-q} \cdot \frac{m}{n_2+1},\end{aligned}$$

where the second-to-last equality is from evaluating the generating function $G(\binom{t}{k}; z) = \sum_{t=0}^{\infty} \binom{t}{k} z^t$ at $z = q$.

Note that this same calculation works for counting the $i \in \mathcal{I}_2$ with $E^{\text{lo}}(X_i) \leq R^{\text{lo}}(X_{n+1})$ using the same lemmas, substituting M_E^{lo} for M_E^{hi}, M_R^{lo} for M_R^{hi}, and q for $1-q$ within the calculation. This gives

$$\mathbb{P}\{M_E^{\text{lo}} \geq m\} \geq 1 - \frac{1}{q} \cdot \frac{m}{n_2+1}.$$

Now, in Algorithm 2, $Q^{\text{hi}}_{1-s(1-q)}(E)$ is defined as the $(1-s(1-q))(1+1/n_2)$-th quantile of $\{E_i^{\text{hi}} : i \in \mathcal{I}_2\}$. This is equal to the $\lceil n_2(1-s(1-q))(1+1/n_2)\rceil = \lceil(1-s(1-q))(n_2+1)\rceil$ smallest value of $\{E_i^{\text{hi}} : i \in \mathcal{I}_2\}$. Thus, if $M_E^{\text{hi}} \geq n_2 - \lceil(1-s(1-q))(n_2+1)\rceil+1$, then $R^{\text{hi}}(X_{n+1})$ will be at most the $n_2-\lceil(1-s(1-q))(n_2+1)\rceil+1$ largest value of $\{E_i^{\text{hi}} : i \in \mathcal{I}_2\}$, which is equal to the $\lceil(1-s(1-q))(n_2+1)\rceil$ smallest value, or $Q^{\text{hi}}_{1-s(1-q)}(E)$. Then, using our earlier lower bound for the inverse CDF of M_E^{hi}, we get that

$$\begin{aligned}\mathbb{P}\{R^{\text{hi}}(X_{n+1}) \leq Q^{\text{hi}}_{1-s(1-q)}(E)\} \geq& \mathbb{P}\{M_E^{\text{hi}} \geq n_2 - \lceil(1-s(1-q))(n_2+1)\rceil + 1\} \\ \geq& \mathbb{P}\{M_E^{\text{hi}} \geq n_2 + 1 - (1-s(1-q))(n_2+1)\} \\ =& \mathbb{P}\{M_E^{\text{hi}} \geq s(1-q)(n_2+1)\} \\ \geq& 1-s.\end{aligned}$$

Similarly, $Q^{\text{lo}}_{rq}(E)$ is defined as the $rq-(1-rq)/n_2$-th quantile of $\{E_i^{\text{lo}} : i \in \mathcal{I}_2\}$. This is equal to the $\lceil n_2(rq-(1-rq)/n_2)\rceil = \lceil(n_2+1)rq-1\rceil$ smallest value of $\{E_i^{\text{lo}} : i \in \mathcal{I}_2\}$. Thus, if $M_E^{\text{lo}} \geq \lceil(n_2+1)rq-1\rceil$, then $R^{\text{lo}}(X_{n+1})$ will be at least $Q^{\text{lo}}_{rq}(E)$. Then, our lower bound for the M_E^{lo} inverse CDF tells us that

$$\begin{aligned}\mathbb{P}\{R^{\text{lo}}(X_{n+1}) \geq Q^{\text{lo}}_{rq}(E)\} \geq& \mathbb{P}\{M_E^{\text{lo}} \geq \lceil(n_2+1)rq-1\rceil\} \\ \geq& \mathbb{P}\{M_E^{\text{lo}} \geq (n_2+1)rq\} \\ \geq& 1-r.\end{aligned}$$

Finally, by the union bound, we have that

$$\mathbb{P}\{Q^{\text{lo}}_{rq}(E) \leq R^{\text{lo}}(X_{n+1}) \text{ and } R^{\text{hi}}(X_{n+1}) \leq Q^{\text{hi}}_{1-s(1-q)}(E)\} \geq 1-r-s = 1-\alpha$$

completing our proof.

B.3. Impossibility of capturing the distribution mean

Instead of proving the impossibility of capturing the conditional mean of a distribution, we prove a more general result: we show that there does not exist an algorithm to capture the mean of a distribution $Y \sim P$ given no assumptions about P. This is a more general form of our result because if we set $X \perp\!\!\!\perp Y$ in $(X,Y) \sim P$, then $\mathbb{E}[Y|X] = \mathbb{E}[Y]$, meaning that the impossibility of capturing the mean results in the conditional mean being impossible to capture as well.

Consider an algorithm $\hat{C}_n$ that, given i.i.d. samples $Y_1, \ldots, Y_n \sim P$, returns a (possibly randomized) confidence interval $\hat{C}_n(\mathcal{D})$, $\mathcal{D} = \{Y_i, 1 \leq i \leq n\}$, with length bounded by some function of P that captures $\mathbb{E}[Y]$ with probability at least $1-\alpha$, i.e. $\mathbb{P}\{\mathbb{E}[Y] \in \hat{C}_n(\mathcal{D})\} \geq 1-\alpha$. Pick $a > \alpha$, with $a < 1$. Consider a distribution P where for $Y \sim P$, $\mathbb{P}\{Y = 0\} = a^{1/n}$ and $\mathbb{P}\{Y = u\} = 1 - a^{1/n}$ for some u. Then for $Y_1, \ldots, Y_n \sim P$, $\mathbb{P}\{Y_1 = \cdots = Y_n = 0\} \geq a$. Consider $\hat{C}_n(\{0, \ldots, 0\})$; by our assumption on $\hat{C}_n$, there must exist some $m \in \mathbb{R}$ for which $\mathbb{P}\{m \in \hat{C}_n(\{0, \ldots, 0\})\} < 1 - \alpha/a$. Then, setting $u = \frac{m}{1-a^{1/n}}$ yields

$\mathbb{E}[Y] = m$. With probability a, $\hat{C}_n(\mathcal{D}) = \hat{C}_n(\{0, \dots, 0\})$, so

$$\mathbb{P}\{\mathbb{E}[Y] = m \notin \hat{C}_n(\mathcal{D})\} > a \cdot \frac{\alpha}{a} = \alpha.$$

This implies that $\mathbb{P}\{\mathbb{E}[Y] \in \hat{C}_n(\mathcal{D})\} < 1 - \alpha$ as desired, completing the proof.

B.4. Capturing the distribution median

```
Algorithm 3: Confidence Interval for Median($Y$) with Coverage $1 - \alpha$
  Input:
    Number of i.i.d. datapoints $n \in \mathbb{N}$.
    Datapoints $Y_1, \dots, Y_n \sim P \subseteq \mathbb{R}$.
    Coverage level $1 - \alpha \in (0, 1)$.
  Process:
    Order the $Y_i$ as $Y_{(1)} \leq \cdots \leq Y_{(n)}$.
    Calculate the largest $k \geq 0$ such that for $X \sim \mathrm{Binom}(n, 0.5)$, we have
    $\mathbb{P}\{X < k\} \leq \alpha/2$.
  Output:
    Confidence interval $\hat{C}_n = [Y_{(k)}, Y_{(n+1-k)}]$ for Median($Y$).
    (Note that $Y_{(0)} = -\infty$ and $Y_{(n+1)} = \infty$)
```

We now show that Algorithm 3 captures the median of P with probability at least $1 - \alpha$. Let $m = \mathrm{Median}(P)$, let $M^{\mathrm{lo}} = \#\{Y_i \leq m : 1 \leq i \leq n\}$ be the number of Y_i at most m, and let $M^{\mathrm{hi}} = \#\{Y_i \geq m : 1 \leq i \leq n\}$ be the number of Y_i at least m. Note that by the definition of m, we have that for all i, $\mathbb{P}\{Y_i \leq m\} \geq 0.5$, and $\mathbb{P}\{Y_i \geq m\} \geq 0.5$ as well. Additionally, the events $\{Y_i \leq m\}$ are mutually independent for all i, as are the events $\{Y_i \geq m\}$. This implies that both M^{lo} and M^{hi} follow a $\mathrm{Binom}(n, 0.5)$ distribution.

Since $\hat{C}_n = [Y_{(k)}, Y_{(n+1-k)}]$, we have that $m \in \hat{C}_n$ if and only if $M^{\mathrm{lo}} \geq k$ and $M^{\mathrm{hi}} \geq k$. Then,

$$\begin{aligned} \mathbb{P}\{m \in \hat{C}_n\} =& \mathbb{P}\{M^{\mathrm{lo}} \geq k \text{ and } M^{\mathrm{hi}} \geq k\} \\ =& 1 - \mathbb{P}\{M^{\mathrm{lo}} < k \text{ or } M^{\mathrm{hi}} < k\} \\ \geq& 1 - (\mathbb{P}\{M^{\mathrm{lo}} < k\} + \mathbb{P}\{M^{\mathrm{hi}} < k\}) \\ \geq& 1 - (\alpha/2 + \alpha/2) \\ =& 1 - \alpha. \end{aligned}$$

Acknowledgments

We thank Lihua Lei for advice on different approaches and recommended resources on this topic, as well as the Stanford Statistics department for listening to a preliminary version of this work. We also thank a referee for encouraging us to deepen the connection between predictive and conditional median confidence intervals.

References

[1] Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics* **27** 1115–1122. MR0084241

[2] Barber, R. F. (2020). Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics* **14** 3487–3524. MR4154845

[3] Barber, R. F., Candes, E. J., Ramdas, A. and Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA* **10** 455–482. MR4270755

[4] Barber, R. F., Candes, E. J., Ramdas, A. and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics* **49** 486–507. MR4206687

[5] Belloni, A., Chernozhukov, V., Chetverikov, D. and Fernández-Val, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics* **213** 4–29. MR4013213

[6] Chernozhukov, V., Wüthrich, K. and Zhu, Y. (2019). Distributional conformal prediction. *arXiv preprint arXiv:1909.07889*.

[7] Cortés-Ciriano, I. and Bender, A. (2019). Concepts and applications of conformal prediction in computational drug discovery. *arXiv preprint arXiv:1908.03569*.

[8] Kivaranovic, D., Johnson, K. D. and Leeb, H. (2020). Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics* 4346–4356. PMLR.

[9] Koenker, R. (2005). *Quantile Regression. Econometric Society Monographs*. Cambridge University Press. MR2268657

[10] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113** 1094–1111. MR3862342

[11] Lei, J., Robins, J. and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association* **108** 278–287. MR3174619

[12] Noether, G. E. (1972). Distribution-free confidence intervals. *The American Statistician* **26** 39–41.

[13] Oberhofer, W. and Haupt, H. (2016). Asymptotic theory for nonlinear quantile regression under weak dependence. *Econometric Theory* **32** 686–713. MR3506436

[14] Papadopoulos, H., Proedrou, K., Vovk, V. and Gammerman, A. (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning* 345–356. Springer. MR2050303

[15] Romano, Y., Barber, R., Sabatti, C. and Candès, E. (2020). With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*.

[16] Romano, Y., Patterson, E. and Candes, E. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems* 3543–3553.

[17] Sesia, M. and Candès, E. J. (2020). A comparison of some conformal quantile regression methods. *Stat* **9** e261. MR4104217

[18] Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research* **9** 371–421. MR2417240

[19] Takeuchi, I., Le, Q. V., Sears, T. D. and Smola, A. J. (2006). Nonparametric Quantile Estimation. *Journal of Machine Learning Research* **7** 1231–1264. MR2274404

[20] Vovk, V., Gammerman, A. and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media. MR2161220

[21] Vovk, V., Nouretdinov, I., Gammerman, A. et al. (2009). Online predictive linear regression. *The Annals of Statistics* **37** 1566–1590. MR2509084

[22] Vovk, V., Nouretdinov, I., Manokhin, V. and Gammerman, A. (2018). Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications* 37–51. PMLR.