Information and Inference: A Journal of the IMA (2020) **00**, 1–28
doi:10.1093/imaiai/iaaa017

The limits of distribution-free conditional predictive inference

Rina Foygel Barber[†]
Department of Statistics, University of Chicago, Chicago, IL 60637
[†]Corresponding author. Email: rina@uchicago.edu

Emmanuel J. Candès
Departments of Statistics and Mathematics, Stanford University, Stanford, CA 94305

and

Aaditya Ramdas, Ryan J. Tibshirani
Department of Statistics and Data Science and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213

[Received on 2 April 2019; revised on 14 April 2020; accepted on 9 June 2020]

We consider the problem of distribution-free predictive inference, with the goal of producing predictive coverage guarantees that hold conditionally rather than marginally. Existing methods such as conformal prediction offer marginal coverage guarantees, where predictive coverage holds on average over all possible test points, but this is not sufficient for many practical applications where we would like to know that our predictions are valid for a given individual, not merely on average over a population. On the other hand, exact conditional inference guarantees are known to be impossible without imposing assumptions on the underlying distribution. In this work, we aim to explore the space in between these two and examine what types of relaxations of the conditional coverage property would alleviate some of the practical concerns with marginal coverage guarantees while still being possible to achieve in a distribution-free setting.

Keywords: distribution-free inference; predictive inference; conformal prediction.

1. Introduction

Consider a training data set $(X_1, Y_1), \ldots, (X_n, Y_n)$, and a test point (X_{n+1}, Y_{n+1}), with the training and test data all drawn i.i.d. from the same distribution. Here, each $X_i \in \mathbb{R}^d$ is a feature vector, while $Y_i \in \mathbb{R}$ is a response variable. The problem of *predictive inference* is the following: if we observe the n training data points, and are given the feature vector X_{n+1} for a new test data point, we would like to construct a prediction interval for Y_{n+1}—that is, a subset of $\mathbb{R}$ that we believe is likely to contain the test point's true response value Y_{n+1}.

As a motivating example, suppose that each data point i corresponds to a patient, with X_i encoding relevant covariates (age, family history, current symptoms, etc.), while the response Y_i measures a quantitative outcome (e.g. reduction in blood pressure after treatment with a drug). When a new patient arrives at the doctor's office with covariate values X_{n+1}, the doctor would like to be able to predict their eventual outcome Y_{n+1} with a range, making a statement along the lines of: 'Based on your age, family history and current symptoms, you can expect your blood pressure to go down by 10–15 mmHg'. In this paper, we will study the problem of making accurate predictive statements of this sort.

To study such questions, throughout this paper we will write $\widehat{C}_n(x) \subseteq \mathbb{R}$ to denote the prediction interval[1] for Y_{n+1} given a feature vector $X_{n+1} = x$. This interval is a function of both the test point x and the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$. We will write $\widehat{C}_n$ (without specifying a test point x) to refer to the algorithm that maps the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to the resulting prediction intervals $\widehat{C}_n(x)$ indexed by $x \in \mathbb{R}^d$. (For convenience in writing our results, we assume that the X_i's lie in $\mathbb{R}^d$, although our results hold more generally for any probability space.)

For the algorithm $\widehat{C}_n$ to be useful, we would like to be assured that the resulting prediction interval is indeed likely to contain the true response value, i.e. that $Y_{n+1} \in \widehat{C}_n(X_{n+1})$ with fairly high probability. When this event succeeds, we say that the predictive interval $\widehat{C}_n(X_{n+1})$ *covers* the true response value Y_{n+1}. Defining the coverage probability is not a trivial question—do we require that coverage holds with high probability on average over the test feature vector X_{n+1}, pointwise at any value $X_{n+1} = x$ or something in between? In order to be robust to distributional assumptions, we would also like to ensure that our algorithm $\widehat{C}_n$ has good coverage properties without making any assumptions about the underlying distribution P—a 'distribution-free' guarantee.

To formalize these ideas, we will begin with a few definitions. Throughout, P will denote a joint distribution on $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, and we will write P_X to denote the induced marginal on X and $P_{Y|X}$ for the conditional distribution of $Y|X$. We say that $\widehat{C}_n$ satisfies *distribution-free marginal coverage* at the level $1 - \alpha$, denoted by $(1 - \alpha)$-MC, if[2]

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geqslant 1 - \alpha \text{ for all distributions } P. \tag{1.1}$$

In other words, the probability that $\widehat{C}_n$ covers the true test value Y_{n+1} is at least $1 - \alpha$, on average over a random draw of the training and test data from *any* distribution P. We say that $\widehat{C}_n$ satisfies *distribution-free conditional coverage* at the level $1 - \alpha$, denoted by $(1 - \alpha)$-CC, if

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} = x\right\} \geqslant 1 - \alpha \text{ for all } P \text{ and almost all } x, \tag{1.2}$$

where, fixing the distribution P, we write 'almost all x' to mean that the set of points $x \in \mathbb{R}^d$ where the bound fails to hold must have measure zero under P_X. This means that the probability that $\widehat{C}_n$ covers, at a *fixed* test point $X_{n+1} = x$, is at least $1 - \alpha$.[3]

Now, how should we interpret the difference between marginal and conditional coverage? With $\alpha = 0.05$, we expect that the doctor's statement ('...you can expect your blood pressure to go down by 10–15 mmHg') should hold with 95% probability. For marginal coverage, the probability is taken over both X_{n+1} and Y_{n+1}, while for conditional coverage, X_{n+1} is fixed, and the probability is taken over Y_{n+1} only (and over all the training data in both situations). This means that for marginal coverage, the doctor's statements have a 95% chance of being accurate *on average* over all possible patients that might arrive at the clinic (marginalizing over X_{n+1}), but might for example have 0% chance of being accurate

[1] Note that the set $\widehat{C}_n(x) \subseteq \mathbb{R}$ is not required to be an interval—it may consist of a disjoint union of multiple intervals. For simplicity, we still refer to the $\widehat{C}_n(x)$'s as 'prediction intervals'.

[2] In these definitions, and throughout the remainder of the paper, all probabilities are taken with respect to training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and test point (X_{n+1}, Y_{n+1}) all drawn i.i.d. from P, unless specified otherwise.

[3] [**13**] also considers a notion of conditional coverage, where the guarantee is required to hold after conditioning on the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ but without conditioning on the test point X_{n+1}, and thus is very different from the type of conditioning that we consider here.

for patients under the age of 25, as long as this is averaged out by a higher-than-95% chance of coverage for patients older than 25. The stronger definition of conditional coverage, on the other hand, removes this possibility and requires that whatever statement the doctor makes (different for each patient) has a 95% chance of being true for every individual patient, regardless of the patient's age, family history, etc.

For practical purposes, then, marginal coverage does not seem to be sufficient—each patient would reasonably hope that the information they receive is accurate for their specific circumstances and is not comforted by knowing that the inaccurate information they might be receiving will be balanced out by some other patient's highly precise prediction. On the other hand, the problem of conditional inference is statistically very challenging and is known to be incompatible with the distribution-free setting [**8**,**13**] (we will discuss this in more detail later on). Our goal in this paper is therefore to explore the middle ground between marginal and conditional inference, while working in the distribution-free setting in order to be robust to violations of any modeling assumptions.

1.1 *Summary of contributions*

As mentioned above, it is known to be impossible for any finite-length prediction interval to satisfy distribution-free conditional coverage in the sense of (1.2)—this is because, without assuming smoothness of the underlying distribution P, we cannot exclude the possibility that there is some sort of discontinuity at $X = x$ that leads to a failure of coverage. (Background on this type of impossibility result is described more formally in Section 2.2.)

This impossibility motivates us to consider an approximate version of the conditional coverage property. We will say that $\widehat{C}_n$ satisfies *distribution-free approximate conditional coverage* at level $1-\alpha$ and tolerance $\delta > 0$, denoted by $(1-\alpha, \delta)$-CC, if

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\} \geqslant 1-\alpha \text{ for all distributions } P \text{ and all } \mathscr{X} \subseteq \mathbb{R}^d \text{ with } P_X(\mathscr{X}) \geqslant \delta. \quad (1.3)$$

For example, at $\alpha = 0.05$ and $\delta = 0.1$, the coverage probability has to be at least 95% for any subgroup of patients that makes up at least 10% of the overall population. If $\delta > 0$ is fairly small, then this approximate conditional coverage property is quite a bit stronger than marginal coverage and may be sufficient for many applications.

However, we find that it is inherently impossible to find non-trivial algorithms that achieve even this relaxed notion of conditional coverage. Specifically, we compare against a trivial solution: we show with a simple argument that any method $\widehat{C}_n$ that satisfies $(1-\alpha\delta)$-MC will also satisfy $(1-\alpha, \delta)$-CC. In this sense, we can trivially achieve approximate conditional coverage by way of marginal coverage, but this solution is not satisfactory since, for small δ, a $(1-\alpha\delta)$-MC prediction interval will be extremely wide. However, the main result of this paper, Theorem 3.1 (see Section 3), proves that any $(1-\alpha, \delta)$-CC method is essentially no better than this kind of trivial construction (in the sense of the expected length of the resulting intervals).

Perhaps then, the definition (1.3) of approximate conditional coverage may be stronger than needed in practical applications. In a medical setting, for instance, a patient would typically want to know that coverage is accurate on average over *a subgroup of patients similar to the individual* and would not be concerned about arbitrary subgroups consisting of highly dissimilar patients. This motivates us to consider alternatives to the approximate conditional coverage property (1.3)—in Section 4, we modify (1.3) to consider only a restricted class of sets $\mathscr{X}$, for instance, only sets consisting of balls

under some metric (to represent patients similar to the individual of interest, in our example). We construct an example of an algorithm that satisfies this type of property—a modification of the split conformal method—that we analyze in Theorem 4.1. We also establish lower (Theorem 4.2) and upper (Theorem 4.3) bounds on the efficiency of any predictive method satisfying this type of property, as a function of the complexity (Vapnik–Chervonenkis (VC) dimension) of the class of sets over which coverage is required to hold.

1.2 *Notation*

Before proceeding, we establish some notation and terminology that will be used throughout the paper. All sets and functions are implicitly assumed to be measurable (e.g. 'for all $\mathcal{X} \subseteq \mathbb{R}^d$' in (1.3) should be interpreted to mean all measurable subsets of $\mathbb{R}^d$). The function leb() denotes Lebesgue measure on $\mathbb{R}$ or on $\mathbb{R}^d$. Prediction intervals are allowed to be either fixed or randomized. Specifically, a non-data-dependent prediction interval $C = C(x)$ may either be fixed (i.e. a function mapping points $x \in \mathbb{R}^d$ to subsets $C(x) \subseteq \mathbb{R}$) or random (i.e. a function mapping points $x \in \mathbb{R}^d$ to a random variable $C(x)$ taking values in the set of subsets of $\mathbb{R}$). Analogously, for a data-dependent prediction interval $\widehat{C}_n = \widehat{C}_n(x)$, fixing the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the vector $x \in \mathbb{R}^d$, this interval may be either a fixed or random subset of $\mathbb{R}$.

2. Background

In this section, we give background on the split conformal prediction method, which achieves distribution-free marginal coverage and review results in the literature establishing that distribution-free conditional coverage is not possible.

2.1 *Split conformal prediction*

The *split conformal prediction* algorithm, introduced in [10,14] (under the name 'inductive conformal prediction') and studied further by [9,13,7], is a well-known method that achieves distribution-free marginal coverage guarantees. This method makes no assumptions at all on the distribution of the data aside from requiring that the training data and the test point are exchangeable. (Of course, assuming that the training and test data are i.i.d. is simply a special case of the exchangeability assumption.)

The split conformal prediction method begins by partitioning the sample size n into two portions, $n = n_0 + n_1$, e.g. split in half. We will use the first n_0 many training points to fit an estimated regression function $\widehat{\mu}_{n_0}(x)$, and the remaining $n_1 = n - n_0$ many training points to determine the width of the prediction interval around $\widehat{\mu}_{n_0}(x)$. The estimated model $\widehat{\mu}_{n_0}$ can be fitted from $(X_1, Y_1), \ldots, (X_{n_0}, Y_{n_0})$ using any algorithm—for example, we might fit a linear model, $\widehat{\mu}_{n_0}(x) = x^\top \widehat{\beta}$ where $\widehat{\beta} \in \mathbb{R}^d$ is fitted on the data points $(X_1, Y_1), \ldots, (X_{n_0}, Y_{n_0})$ using least squares regression or any other regression method.

Next, fix a desired predictive coverage level $1 - \alpha$, for instance 95%. We then compute residuals

$$R_i = \left| Y_i - \widehat{\mu}_{n_0}(X_i) \right| \text{ for } i = n_0 + 1, \ldots, n,$$

and define[4]

$$\widehat{q}_{n_1} = \text{the } \left\lceil (1-\alpha)(n_1+1) \right\rceil \text{-th smallest value of the list } R_{n_0+1}, \ldots, R_n.$$

[4] Formally, when we write 'the k-th smallest value of the list...' for a list that has m elements, this will denote $+\infty$ in the case that $k > m$.

The predictive interval is then defined as

$$\widehat{C}_n(x) = \left[\widehat{\mu}_{n_0}(x) - \widehat{q}_{n_1}, \widehat{\mu}_{n_0}(x) + \widehat{q}_{n_1}\right]. \tag{2.1}$$

This method can also be generalized to include a local variance/scale estimate or to allow for an asymmetric construction treating the right and left tails of the residuals separately.

The split conformal algorithm is a variant of *conformal prediction*, which has a rich literature dating back many years (see, e.g. [14,11] for background). Conformal prediction similarly relies on the exchangeability of the training and test data, but rather than splitting the training data to separate the tasks of model fitting and calibrating the quantiles, conformal prediction uses the full training sample for both tasks, thus paying a higher computational cost. Here, for simplicity, we do not describe conformal prediction but focus on the split conformal algorithm, which we generalize in our own proposed methods later on.

Using the assumption that the data points are i.i.d., the proof that the split conformal prediction method satisfies $(1-\alpha)$-MC is very intuitive. For completeness, we state this known result here.

THEOREM 2.1 ([10, Proposition 1]). The split conformal prediction method defined in (2.1) satisfies the $(1-\alpha)$-MC property (1.1).

Importantly, the above guarantee holds irrespective of the regression algorithm used to fit $\widehat{\mu}_{n_0}$. Furthermore, [7] show that, in some settings, this distribution-free construction may result in an interval that is asymptotically no wider than the best possible 'oracle' interval—in other words, it is possible to provide marginal distribution-free prediction without incurring a cost in terms of overly wide intervals. (The intuition behind the proof of Theorem 2.1 will be discussed in Section 4.1 as a special case of our new results; [7]'s guarantee of optimal length will be discussed in more detail in Section 4.2.3.)

2.2 *Impossibility of distribution-free conditional coverage*

While the split conformal method satisfies distribution-free marginal coverage (1.1), as mentioned earlier, this property may not be sufficient for practical prediction tasks, as it leaves open the possibility that entire regions of test points (e.g. subgroups of patients) are receiving inaccurate predictions. To avoid this problem, we may wish to construct $\widehat{C}_n$ to guarantee coverage conditional on X_{n+1}, rather than on average over X_{n+1}. Is it possible to achieve distribution-free conditional coverage (1.2) while still constructing predictive intervals that are not too much larger than needed?

Unfortunately, it is well known that, if we do not place any assumptions on P, then estimation and inference on various functionals of P are impossible to carry out, see, e.g. [1,4] for background. More specifically, for the current problem of distribution-free conditional prediction intervals, [8,13] prove that the $(1-\alpha)$-CC property (1.2) is impossible for any algorithm $\widehat{C}_n$, unless $\widehat{C}_n$ has the property that it produces intervals with infinite expected length under *any* non-discrete distribution P, which is not a meaningful procedure.

PROPOSITION 2.2 [Rephrased from [13,8]] Suppose that $\widehat{C}_n$ satisfies $(1-\alpha)$-CC (1.2). Then for all distributions P, it holds that

$$\mathbb{E}\left[\text{leb}(\widehat{C}_n(x))\right] = \infty$$

at almost all points x aside from the atoms of P_X.

In other words, at almost all nonatomic points x, the prediction interval has infinite expected length. This means that distribution-free conditional coverage in the sense of (1.2) is impossible to attain in any meaningful sense.

Asymptotic conditional coverage. There is an extensive literature examining this problem in a setting where P is assumed to satisfy some type of smoothness condition, and conditional coverage can then be achieved asymptotically by letting the sample size n tend to infinity and using a vanishing bandwidth to compute local smoothed estimators of the conditional distribution of $Y|X$. Works in this line of the literature include [3,8], among many others. In this present work, however, we are interested in obtaining distribution-free guarantees that hold at any finite sample size n, and therefore, we aim to avoid relying on assumptions such as smoothness of P or on asymptotic arguments.

3. Approximate conditional coverage

While the results of [13] and [8] prove that distribution-free methods cannot achieve conditional predictive guarantees, in practice it may be sufficient to obtain 'approximately conditional' inference. In our doctor/patient example, we would certainly want to make sure that there is no entire subgroup of patients that are all receiving poor predictions—as in our earlier example where the predictive intervals had poor coverage for all patients below the age of 25—but we may be willing to accept that some rare groups of patients might be receiving inaccurate information.

We therefore try to relax our requirement of conditional coverage to an approximate version—recall from Section 1.1 that $\widehat{C}_n$ satisfies *distribution-free approximate conditional coverage* at level $1-\alpha$ and tolerance $\delta > 0$, denoted by $(1-\alpha,\delta)$-CC, if (1.3) holds. We can easily verify that approximate conditional coverage limits to conditional coverage by taking δ to zero:

$$\widehat{C}_n \text{ satisfies } (1-\alpha)-CC \quad \Longleftrightarrow \quad \widehat{C}_n \text{ satisfies } (1-\alpha,\delta)-CC \text{ for all } \delta > 0.$$

At the other extreme, marginal coverage is recovered by taking $\delta = 1$:

$$\widehat{C}_n \text{ satisfies } (1-\alpha)-MC \quad \Longleftrightarrow \quad \widehat{C}_n \text{ satisfies } (1-\alpha,\delta)-CC \text{ for } \delta = 1.$$

While we have seen that exact conditional coverage is impossible to meaningfully attain, does this relaxation allow us to move towards a meaningful solution? To answer this question, it is useful to first consider a simple solution obtained by way of a marginal coverage method.

3.1 *The inadequacy of reducing to marginal coverage*

The following lemma suggests that our approximate conditional coverage can be naively obtained via marginal coverage at a more stringent level.

LEMMA 3.1 Let $\widehat{C}_n$ be any method that attains distribution-free marginal coverage (1.1) with miscoverage rate $\alpha\delta$ in place of α, that is, $\widehat{C}_n$ satisfies the $(1-\alpha\delta)$-MC property. Then $\widehat{C}_n$ also satisfies $(1-\alpha,\delta)$-CC.

Proof of Lemma 3.1. Since $\widehat{C}_n$ satisfies $(1-\alpha\delta)$-MC, for any distribution P we have

$$\begin{aligned}\alpha\delta &\geqslant \mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1})\right\} \geqslant \mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), X_{n+1} \in \mathscr{X}\right\} \\ &\geqslant \delta \cdot \mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\},\end{aligned}$$

where the last step holds for any $\mathscr{X}$ with $\mathbb{P}\left\{X_{n+1} \in \mathscr{X}\right\} = P_X(\mathscr{X}) \geqslant \delta$. Rearranging yields the lemma. □

To interpret this lemma, we might apply the split conformal prediction algorithm (2.1) at the miscoverage level $\alpha\delta$, which ensures marginal coverage at this level and, therefore, ensures $(1-\alpha, \delta)$-CC. However, we would typically choose δ to be quite small, as we would like to be able to condition on small sets $\mathscr{X}$ (to ensure that there aren't any large subgroups of patients all receiving poor information). This means that any prediction intervals satisfying $(1-\alpha\delta)$-MC must generally be extremely wide, e.g. 99.5% coverage intervals instead of 95% coverage intervals when $\alpha = 0.05$ and $\delta = 0.1$. Therefore, the naive solution of using marginal coverage to ensure approximate conditional coverage is not satisfactory.

Before moving on, we extend Lemma 3.1 to generalize the naive solution given by $(1-\alpha\delta)$-MC:

LEMMA 3.2 Let $\widehat{C}_n$ be any method that satisfies $(1-c\alpha\delta)$-MC (1.1), for some $c \in [0,1]$. Let $\widehat{C}'_n$ be defined as follows: at a test point x, with probability $\frac{1-\alpha}{1-c\alpha}$, we define $\widehat{C}'_n(x) = \widehat{C}_n(x)$, or otherwise, we define $\widehat{C}'_n(x) = \varnothing$ (the empty set), where we assume that this decision is carried out independently of x and of the training data. Then $\widehat{C}'_n$ also satisfies $(1-\alpha, \delta)$-CC.

Proofs for this lemma and for all subsequent theoretical results are given in the Appendix.

To understand the role of the parameter c in this lemma, we can consider the two extremes—setting $c = 1$, we would simply output the interval $\widehat{C}_n(x)$ that satisfies $(1-\alpha\delta)$-MC, i.e. we return to the naive solution of Lemma 3.1. At the other extreme, if we set $c = 0$, at any test point $X_{n+1} = x$ the resulting prediction interval would be given by $\mathbb{R}$ with probability $1-\alpha$, or $\varnothing$ otherwise—this clearly satisfies $(1-\alpha, \delta)$-CC (and, in fact, $(1-\alpha)$-CC) but is of course meaningless as it reveals no information about the data.

3.2 *Hardness of approximate conditional coverage*

We now introduce our main result, which proves that, as in the exact conditional coverage setting, the relaxation to $(1-\alpha, \delta)$-conditional coverage is still impossible to attain meaningfully. In particular, the naive solution—obtaining $(1-\alpha, \delta)$-CC by way of marginal coverage, as in Lemmas 3.1 and 3.2—is in some sense the best possible method, in terms of the lengths of the resulting prediction intervals.

To quantify this, for any P and any marginal coverage level $1-\alpha$, consider finding the prediction interval $C_P(x)$ with the shortest possible length, subject to requiring marginal coverage to be at least $1-\alpha$ under the distribution P. As the notation suggests, the coverage properties of $C_P(x)$ are specific to P and are not distribution-free in any sense. Formally, we define the set of intervals with marginal coverage under P as

$$\mathscr{C}_P(1-\alpha) = \left\{C_P : \mathbb{P}_P\left\{Y \in C_P(X)\right\} \geqslant 1-\alpha\right\},$$

where $C_P(x)$ may denote a fixed or random interval (that is, C_P is a function mapping points $x \in \mathbb{R}^d$ to fixed or random subsets of $\mathbb{R}$). We can then define the minimum possible length as

$$L_P(1-\alpha) = \inf_{C_P \in \mathscr{C}_P(1-\alpha)} \left\{ \mathbb{E}_{P_X} \left[\text{leb}(C_P(X)) \right] \right\}. \tag{3.1}$$

If C_P is random rather than fixed, then we should interpret the expectation as being taken with respect to the random draw of X and the randomization in the construction of $C_P(X)$.

With these definitions in place, we present our main result, which proves a lower bound on the prediction interval width of any method that attains distribution-free approximate conditional coverage.

THEOREM 3.1 Suppose that $\widehat{C}_n$ satisfies $(1-\alpha, \delta)$-CC (1.3). Then for all distributions P where the marginal distribution P_X has no atoms,

$$\mathbb{E}\left[\text{leb}(\widehat{C}_n(X_{n+1}))\right] \geqslant \inf_{c \in [0,1]} \left\{ \frac{1-\alpha}{1-c\alpha} \cdot L_P(1-c\alpha\delta) \right\}.$$

How should we interpret this lower bound? Based on Lemma 3.1, we can achieve $(1-\alpha, \delta)$-CC trivially by running split conformal prediction at the marginal coverage level $1-\alpha\delta$. What would be the average width from such a procedure? As mentioned in Section 2.1, under certain assumptions on P, [7] prove that the split conformal method run at coverage level $1-\alpha\delta$ with a consistent regression algorithm $\widehat{\mu}$ will, with high probability, output a prediction interval with width that is only $o(1)$ larger than the oracle interval, which has width $L_P(1-\alpha\delta)$. More generally, for any $c \in [0,1]$, we can use the construction suggested in Lemma 3.2 combined with the split conformal method, now run at level $1-c\alpha\delta$, to instead produce expected length $\approx \frac{1-\alpha}{1-c\alpha} \cdot L_P(1-c\alpha\delta)$.

Since Theorem 3.1 demonstrates that any method satisfying $(1-\alpha, \delta)$-CC cannot beat this lower bound, this means that the $(1-\alpha, \delta)$-CC property is impossible to attain beyond the trivial solution, i.e. by applying a method that guarantees $(1-\alpha\delta)$-marginal coverage, which then yields $(1-\alpha, \delta)$-CC as a byproduct (or choosing some $c \in [0,1]$ for the more general construction). Since typically we would choose δ to be a small constant, this lower bound is indeed a substantial issue, since $L_P(1-\alpha\delta)$ will generally be much larger than the length we would need if the distribution P were known.

4. Restricted conditional coverage

Our main result, Theorem 3.1, shows that our definition of approximate conditional coverage in (1.3) is too strong; it is impossible to construct a meaningful procedure satisfying this definition. One way to weaken this condition is to restrict which sets $\mathscr{X}$ we consider, yielding a less stringent notion of approximate conditional coverage.

For example, we can require that the coverage guarantee holds 'locally' by conditioning only on any *ball* with sufficient probability δ, rather than on an arbitrary subset $\mathscr{X} \subseteq \mathbb{R}^d$. More concretely, we might require that

$$\begin{aligned} &\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; X_{n+1} \in \mathbb{B}(x,r) \right\} \\ &\quad \geqslant 1-\alpha \text{ for all distributions } P \text{ and all } x \in \mathbb{R}^d, r \geqslant 0 \text{ with } P_X(\mathbb{B}(x,r)) \geqslant \delta. \end{aligned} \tag{4.1}$$

Here, $\mathbb{B}(x, r)$ is the closed ℓ_2 ball centered at x with radius r. In the doctor/patient example, we can think of this as requiring 95% predictive accuracy on average over the subgroup of population consisting of patients similar to a given patient x, where similarity is defined with the ℓ_2 norm (of course, we can also generalize this to different metrics). As another example, [13 ,8] consider a version of conformal prediction that guarantees coverage within each one of a finite number of subgroups, i.e.

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}_k\right\} \geqslant 1-\alpha \text{ for all distributions } P \text{ and for all } k = 1, \ldots, K, \quad (4.2)$$

for some fixed partition $\mathbb{R}^d = \mathscr{X}_1 \cup \cdots \cup \mathscr{X}_K$ of the feature space. Here, we may think of predefining subgroups of patients (males below age 25, males age 25–35, etc.) and requiring 95% predictive accuracy on average over each predefined subgroup.

More generally, suppose we are given a collection $\mathfrak{X}$ of measurable subsets of $\mathbb{R}^d$. We say that $\widehat{C}_n$ satisfies distribution-free approximate conditional coverage at level $1-\alpha$ and tolerance $\delta > 0$ relative to the collection $\mathfrak{X}$, denoted by $(1-\alpha, \delta, \mathfrak{X})$-CC, if

$$\begin{aligned}&\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\}\\ &\quad\geqslant 1-\alpha \text{ for all distributions } P \text{ and all } \mathscr{X} \in \mathfrak{X} \text{ with } P_X(\mathscr{X}) \geqslant \delta. \end{aligned} \quad (4.3)$$

To avoid degenerate scenarios, we will assume that we always have $\mathbb{R}^d \in \mathfrak{X}$, meaning that requiring $(1-\alpha, \delta, \mathfrak{X})$-CC is always at least as strong as requiring $(1-\alpha)$-MC. Of course, this definition yields the original $(1-\alpha, \delta)$-CC condition if we take $\mathfrak{X}$ to be the collection of all measurable sets. If the class $\mathfrak{X}$ is too rich, then, our main result in Theorem 3.1 proves that $(1-\alpha, \delta, \mathfrak{X})$-CC is impossible to achieve beyond trivial solutions. We may ask then whether it's possible to construct meaningful prediction intervals when $\mathfrak{X}$ is sufficiently restricted.

In the following, we will first construct a concrete algorithm, based on the split conformal prediction method, that attains $(1-\alpha, \delta, \mathfrak{X})$-CC. Afterwards, we will attempt to determine how the complexity of the class $\mathfrak{X}$ determines whether this algorithm provides meaningful prediction intervals (i.e. narrower intervals than the lower bound of Theorem 3.1) and indeed if this is possible to attain with any algorithm.

4.1 *Split conformal for restricted conditional coverage*

As a concrete example, we will construct a variant of the split conformal prediction method and will generalize [7]'s results on the efficiency of split conformal prediction to establish conditions under which the resulting prediction intervals are asymptotically efficient.

Let $\widehat{\mu}_{n_0}(x)$ be some fitted regression function, which estimates the conditional mean of Y given $X = x$. As before, we require that $\widehat{\mu}_{n_0}$ is fitted on the first n_0 training samples, $(X_1, Y_1), \ldots, (X_{n_0}, Y_{n_0})$. Next, define the residual

$$R_i = \left|Y_i - \widehat{\mu}_{n_0}(X_i)\right|$$

on the remaining training samples $i = n_0 + 1, \ldots, n$ and on the test point $i = n+1$. (As for the original split conformal method, this procedure can be generalized to include a local scale estimate, $\widehat{\sigma}_{n_0}(X_i)$ or to allow for an asymmetric interval that treats the right and left tails of the residuals differently, but we do not include these generalizations here.)

The original split conformal method operates by observing that the test point residual, R_{n+1}, is equally likely to occur anywhere in the ranked list of residuals $R_{n_0+1}, \ldots, R_n, R_{n+1}$, i.e. the test residual

is exchangeable with the n_1 many residuals from the held-out portion of the training data. The split conformal prediction interval (2.1) is then constructed as

$$\widehat{C}_n(x) = \left[\widehat{\mu}_{n_0}(x) - \widehat{q}_{n_1}, \widehat{\mu}_{n_0}(x) + \widehat{q}_{n_1}\right],$$

where $\widehat{q}_{n_1}$ is the $\lceil (1-\alpha)(n_1+1) \rceil$-th smallest value among $R_{n_0+1}, \ldots, R_n$. The width of this prediction interval is determined by this residual quantile $\widehat{q}_{n_1}$, which is calculated by pooling all residuals from the holdout set $i = n_0 + 1, \ldots, n$ and is therefore calibrated to give the appropriate coverage level *on average* over the distribution P (as in Theorem 2.1).

We now need to modify this construction to guarantee a stronger notion of coverage—we need to ensure coverage on average over any $\mathscr{X} \in \mathfrak{X}$ with $P_X(\mathscr{X}) \geqslant \delta$. We will need to modify the width of the prediction interval—for example, for a set $\mathscr{X}$ where residuals tend to be large (i.e. $|Y - \widehat{\mu}_{n_0}(X)|$ is likely to be large if we condition on $X \in \mathscr{X}$), the split conformal interval constructed above is too narrow to achieve $1-\alpha$ coverage on average over this set. We will therefore construct a new interval,

$$\widehat{C}_n(x) = \left[\widehat{\mu}_{n_0}(x) - \widehat{q}_{n_1}(x), \widehat{\mu}_{n_0}(x) + \widehat{q}_{n_1}(x)\right]. \tag{4.4}$$

The width of the interval is now defined locally by the quantity $\widehat{q}_{n_1}(x)$, which we will address next. Intuitively, if x belongs to a set $\mathscr{X}$ within which residuals tend to be large, we will need $\widehat{q}_{n_1}(x)$ to be large in order to achieve the right coverage level on average over $\mathscr{X}$.

We now construct $\widehat{q}_{n_1}(x)$. First, we will narrow down the class of subsets to consider. Define

$$\widehat{N}_{n_1}(\mathscr{X}) = \sum_{i=n_0+1}^{n} \mathbf{1}\left\{X_i \in \mathscr{X}\right\},$$

the number of holdout points that lie in $\mathscr{X}$. Next, let

$$\widehat{\mathfrak{X}}_{n_1} = \left\{ \mathscr{X} \in \mathfrak{X} : \widehat{N}_{n_1}(\mathscr{X}) \geqslant \delta n_1 \left(1 - \sqrt{\frac{2\log n_1}{\delta n_1}}\right) \right\} \quad \subseteq \mathfrak{X}.$$

This definition ensures that if a given subset $\mathscr{X}$ has probability $\geqslant \delta$ under P, then we will include $\mathscr{X} \in \widehat{\mathfrak{X}}_{n_1}$ with high probability. Next, let

$$\widehat{q}_{n_1}(\mathscr{X}) = \text{the } \left\lceil \left(1 - \alpha + \frac{1}{n_1}\right) \cdot \left(\widehat{N}_{n_1}(\mathscr{X}) + 1\right) \right\rceil \text{–th smallest value of } \left\{R_i : n_0+1 \leqslant i \leqslant n, X_i \in \mathscr{X}\right\}.$$

Finally, we set

$$\widehat{q}_{n_1}(x) = \sup_{\mathscr{X} \in \widehat{\mathfrak{X}}_{n_1} : x \in \mathscr{X}} \widehat{q}_{n_1}(\mathscr{X}). \tag{4.5}$$

(Recall that $\mathbb{R}^d \in \mathfrak{X}$ by assumption, and thus $\mathbb{R}^d \in \widehat{\mathfrak{X}}_{n_1}$, so there is always at least one set $\mathscr{X}$ in this supremum.)

Our next result proves that this construction achieves the desired approximate conditional coverage property.

THEOREM 4.1 For any class $\mathfrak{X}$ of measurable subsets of $\mathbb{R}^d$, the prediction interval defined in (4.4) satisfies $(1-\alpha, \delta, \mathfrak{X})$-CC (4.3).

Of course, the supremum defined in (4.5) may be impossible to compute efficiently—this will naturally depend on the structure of the class $\mathfrak{X}$. (We expect that for simple cases, such as taking $\mathfrak{X}$ to be the set of all ℓ_2 balls as for the 'local' conditional coverage discussed earlier, we may be able to compute or approximate (4.5) more efficiently; we leave this as an open question for future work.) Furthermore, this guarantee does not yet establish that this method provides a meaningful prediction interval—it may be the case that the intervals are too wide. We will examine this question next.

4.2 *Characterizing hardness with the VC dimension*

For a class $\mathfrak{X}$ of subsets of $\mathbb{R}^d$, we write $\mathrm{VC}(\mathfrak{X})$ to denote the VC dimension of the class $\mathfrak{X}$. This measure of complexity is defined as follows. For any finite set $\mathscr{A}$ of points in $\mathbb{R}^d$, we say that $\mathscr{A}$ is *shattered* by $\mathfrak{X}$ if, for every subset of points $\mathscr{B} \subseteq \mathscr{A}$, there exists some $\mathscr{X} \in \mathfrak{X}$ with $\mathscr{X} \cap \mathscr{A} = \mathscr{B}$. The VC dimension is then defined as

$$\mathrm{VC}(\mathfrak{X}) = \max\{|\mathscr{A}| : \mathscr{A} \text{ is shattered by } \mathfrak{X}\},$$

i.e. the largest cardinality of any set shattered by $\mathfrak{X}$. Well-known examples include:

- If $\mathfrak{X}$ is the set of all ℓ_2 balls in $\mathbb{R}^d$, then $\mathrm{VC}(\mathfrak{X}) = d+1$.
- If $\mathfrak{X}$ is the set of all half-spaces in $\mathbb{R}^d$, then $\mathrm{VC}(\mathfrak{X}) = d+1$.
- If $\mathfrak{X}$ is the set of all intersections of k different half-spaces in $\mathbb{R}^d$, then $\mathrm{VC}(\mathfrak{X}) = \mathscr{O}(kd \log k)$ [**2**, Lemma 3.2.3].

While a large VC dimension of $\mathfrak{X}$ ensures that there is *some* set of points $\mathscr{A}$ that is shattered by $\mathfrak{X}$, we need a stronger formulation to establish a hardness result for restricted conditional coverage. We will consider an 'almost everywhere' version of the VC dimension, defined as follows:

$$\mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X}) = \max\left\{ m \geqslant 0 : \begin{array}{c} \text{the class of sets } \mathscr{A} = \{a_1, \ldots, a_m\} \subseteq \mathbb{R}^d \\ \text{such that } \mathfrak{X} \text{ does } \textit{not} \text{ shatter } A, \\ \text{has Lebesgue measure zero in } (\mathbb{R}^d)^m \end{array} \right\}$$

In other words, instead of searching for a *single* set $\mathscr{A}$ of size m that is shattered by $\mathfrak{X}$, we require that *almost all* sets $\mathscr{A}$ of size m are shattered by $\mathfrak{X}$. It is trivial that $\mathrm{VC}(\mathfrak{X}) \geqslant \mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X})$, but in fact, the two may coincide—for example,

- If $\mathfrak{X}$ is the set of all ℓ_2 balls in $\mathbb{R}^d$, then $\mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X}) = \mathrm{VC}(\mathfrak{X}) = d+1$.
- If $\mathfrak{X}$ is the set of all half-spaces in $\mathbb{R}^d$, then $\mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X}) = \mathrm{VC}(\mathfrak{X}) = d+1$.

In order to obtain a tight bound, we also need to define a slightly stronger notion of predictive coverage. Our previous definitions (for marginal, conditional and approximate conditional coverage) all calculated probabilities with respect to P^{n+1} for some distribution P, in other words, with the data points $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ drawn i.i.d. from an arbitrary distribution. A more general setting is where these $n+1$ data points are instead assumed to be exchangeable (which includes i.i.d. as a special case). We thus define a notion of approximate conditional coverage under exchangeability rather than the i.i.d. assumption. We say that a procedure $\widehat{C}_n$ satisfies $(1-\alpha, \delta, \mathfrak{X})$-conditional coverage under

exchangeability, denoted by $(1-\alpha, \delta, \mathfrak{X})$-CCE, if

$$\begin{aligned}&\mathbb{P}_{\tilde{P}}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\} \\ &\quad \geqslant 1-\alpha \text{ for all exchangeable distributions } \tilde{P} \text{ on } (X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \\ &\quad \text{and all } \mathscr{X} \in \mathfrak{X} \text{ with } \mathbb{P}_{\tilde{P}}\left\{X_{n+1} \in \mathscr{X}\right\} \geqslant \delta. \end{aligned} \tag{4.6}$$

Clearly, a procedure $\widehat{C}_n$ satisfying $(1-\alpha, \delta, \mathfrak{X})$-CCE will also satisfy $(1-\alpha, \delta, \mathfrak{X})$-CC by definition. It is worth noting that all proofs of predictive coverage guarantees for conformal and split conformal prediction methods do not require the i.i.d. assumption but rather only need to assume exchangeability—that is, results such as Theorem 4.1 continue to hold, meaning that our split conformal method proposed in Section 4.1 satisfies this stronger coverage property (4.6).

We will now see how the VC dimension relates to the conditional coverage problem. We will show that:

- If $\mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X}) \geqslant 2n+2$, then the $(1-\alpha, \delta, \mathfrak{X})$-CCE property cannot be obtained beyond the trivial lower bound given in Theorem 3.1.
- On the other hand, if $\mathrm{VC}(\mathfrak{X}) \ll \delta n/\log n$, then the split conformal method described in Section 4.1, which is guaranteed to satisfy $(1-\alpha, \delta, \mathfrak{X})$-CCE, produces prediction intervals of nearly optimal length under a location-family model.

An equivalent perspective is that with sufficiently many points n, the CCE property can be meaningfully attained. We now formalize these results.

4.2.1 *A lower bound* First, we will examine the setting where $\mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X}) \geqslant 2n+2$. In this setting, we will see that $(1-\alpha, \delta, \mathfrak{X})$-conditional coverage under exchangeability is incompatible with meaningful predictive intervals.

THEOREM 4.2 Suppose that $\widehat{C}_n$ satisfies $(1-\alpha, \delta, \mathfrak{X})$-CCE as defined in (4.6), where $\mathfrak{X}$ satisfies $\mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X}) \geqslant 2n+2$. Then for all distributions P where the marginal distribution P_X is continuous with respect to Lebesgue measure, we have

$$\mathbb{E}\left[\mathrm{leb}(\widehat{C}_n(X_{n+1}))\right] \geqslant \inf_{c \in [0,1]} \left\{ \frac{1-\alpha}{1-c\alpha} \cdot L_P(1-c\alpha\delta) \right\}.$$

In other words, if $\mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X}) \geqslant 2n+2$, the lower bound proved here is identical to that of Theorem 3.1, which is the trivial lower bound that can be obtained by simply requiring marginal coverage at a far stricter level. (For example, if we take $\mathfrak{X}$ to be the collection of all balls or all half-spaces in $\mathbb{R}^d$ for $d \geqslant 2n+1$, then this condition on $\mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X})$ will hold.) We remark that it is possible to prove a similar result for the $(1-\alpha, \delta, \mathfrak{X})$-CC condition (rather than the stronger $(1-\alpha, \delta, \mathfrak{X})$-CCE condition), but in that case we are only able to show this result when $\mathrm{VC}_{\mathrm{a.e.}}(\mathfrak{X}) \gg n^2$.

4.2.2 *An upper bound* Next, we prove that efficient prediction is possible when the VC dimension is low.

Since our construction given in (4.4) uses a symmetric interval around an initial model $\widehat{\mu}_{n_0}$, with the width of the interval selected to cover the worst-case scenario in terms of the choice of $\mathscr{X}$, we can

only hope for efficiency as compared to the best 'oracle' interval of this form. For a fixed function $\mu : \mathbb{R}^d \to \mathbb{R}$ and for any $\mathscr{X} \in \mathfrak{X}$ with nonzero probability under P_X, define

$$q^*_{P,\mu,\alpha}(\mathscr{X}) = \text{the } (1-\alpha) - \text{quantile of } |Y - \mu(X)|, \text{ under the distribution } (X,Y) \sim P \text{ conditional on } X \in \mathscr{X}.$$

Next, for any $x \in \mathbb{R}^d$, define

$$q^*_{P,\mu,\alpha,\delta}(x) = \sup_{\mathscr{X} \in \mathfrak{X} : x \in \mathscr{X}, P_X(\mathscr{X}) \geqslant \delta} q^*_{P,\mu,\alpha}(\mathscr{X}),$$

the maximum quantile over any set $\mathscr{X}$ containing the point x. We will then consider the 'oracle' prediction interval

$$C^*_{P,\mu,\alpha,\delta}(x) = \left[\mu(x) - q^*_{P,\widehat{\mu}_{n_0},\alpha,\delta}(x), \mu(x) + q^*_{P,\widehat{\mu}_{n_0},\alpha,\delta}(x)\right]. \tag{4.7}$$

We can easily verify that $C^*_{P,\mu,\alpha,\delta}(x)$ satisfies

$$\mathbb{P}_P\left\{Y \in C^*_{P,\mu,\alpha,\delta}(X) \;\middle|\; X \in \mathscr{X}\right\} \geqslant 1 - \alpha$$

for all $\mathscr{X} \in \mathfrak{X}$ with $P_X(\mathscr{X}) \geqslant \delta$.

Our main result proves that, if the collection $\mathfrak{X}$ has sufficiently small VC dimension, then with high probability the prediction interval $\widehat{C}_n$ constructed in (4.4) above is essentially the same as the 'oracle' interval defined in (4.7), when constructed around the pre-trained model $\mu = \widehat{\mu}_{n_0}$. To formalize this, we show that $\widehat{C}_n$ is bounded above and below by oracle intervals with slightly perturbed values of α and δ.

THEOREM 4.3 Assume that $\mathrm{VC}(\mathfrak{X}) \geqslant 1$ and $n_1 \geqslant 2$. Then for every $x \in \mathbb{R}^d$, if $\mathrm{VC}(\mathfrak{X}) \leqslant c \cdot \frac{\delta n_1}{\log n_1}$, then the split conformal prediction interval $\widehat{C}_n$ defined in (4.4) satisfies

$$\mathbb{P}_{P^n}\left\{C^*_{P,\widehat{\mu}_{n_0},\alpha_+,\delta_+}(x) \subseteq \widehat{C}_n(X_{n+1}) \subseteq C^*_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}(x)\right\} \geqslant 1 - \frac{1}{n_1},$$

where

$$\alpha_+ = \alpha + c_\alpha\sqrt{\frac{\mathrm{VC}(\mathfrak{X})\log n_1}{\delta n_1}}, \quad \alpha_- = \alpha - c_\alpha\sqrt{\frac{\mathrm{VC}(\mathfrak{X})\log n_1}{\delta n_1}}$$

and

$$\delta_+ = \delta + c_\delta\sqrt{\frac{\mathrm{VC}(\mathfrak{X})\log n_1}{n_1}}, \quad \delta_- = \delta - c_\delta\sqrt{\frac{\mathrm{VC}(\mathfrak{X})\log n_1}{n_1}}$$

and where c, c_α, c_δ are universal constants.

4.2.3 *Special case: the location family with i.i.d. noise* While the result given in Theorem 4.3 is quite general (we do not assume anything about the distribution P), we can consider a special case where, given strong conditions on P, the prediction interval $\widehat{C}_n(X_{n+1})$ nearly matches a much stronger oracle—namely, the narrowest possible valid prediction interval.

Our discussion for this setting will closely follow the work of [7] for the split conformal method. We first describe their results. Their work assumes a location-family model:

$$\begin{gathered}\text{The distribution of } Y|X \text{ is given by } Y = \mu_P(X) + \epsilon, \textit{where}\\ \mu_P(x) \text{ is a fixed function, and } \epsilon \text{ is independent from } X \text{ and has density } f_\epsilon,\\ \text{where } f_\epsilon(t) \text{ is symmetric around } t = 0 \text{ and nonincreasing for } t \geqslant 0.\end{gathered} \tag{4.8}$$

[7] additionally assume that the estimator $\widehat{\mu}_{n_0}(x)$ of the true mean function $\mu_P(x)$ is consistent—Assumption A4 in their work requires that

$$\mathbb{P}\left\{\mathbb{E}\left[\left(\widehat{\mu}_{n_0}(X) - \mu_P(X)\right)^2 \,\middle|\, \widehat{\mu}_{n_0}\right] \leqslant \eta_{n_0}\right\} \geqslant 1 - \rho_{n_0}, \tag{4.9}$$

where we should think of the quantities η_{n_0}, ρ_{n_0} as small or vanishing. To interpret this assumption, the probability on the outside is taken with respect to the training data $(X_1, Y_1), \ldots, (X_{n_0}, Y_{n_0})$ used to fit the model $\widehat{\mu}_{n_0}$, while the conditional expectation on the inside is taken with respect to a new draw $X \sim P_X$.

Under conditions (4.8) and (4.9), [7] prove that the split conformal method (2.1) is asymptotically efficient as $n_0, n_1 \to \infty$, satisfying bounds of the form

$$\operatorname{leb}\big(\widehat{C}_n(X_{n+1}) \triangle C^*_P(X_{n+1})\big) = o_P(1), \tag{4.10}$$

where $\triangle$ denotes the symmetric set difference and where $C^*_P(x)$ denotes the 'oracle' prediction interval that we would build if we knew the distribution P—under the simple model (4.8) for P above, this interval has the form

$$C^*_P(x) = \mu_P(x) \pm q^*_{\epsilon,\alpha},$$

where $q^*_{\epsilon,\alpha}$ denotes the $(1-\alpha/2)$ quantile of f_ϵ (i.e. the $(1-\alpha)$-quantile of the distribution of $|\epsilon|$).

We now extend this result to the setting of approximate conditional coverage. Specifically, working under the same assumptions, we will prove that our proposed algorithm (4.4), which is constructed to satisfy the $(1-\alpha, \delta, \mathfrak{X})$-CC property, will also return an interval that is asymptotically equivalent to the oracle interval C^*_P as long as $\mathrm{VC}(\mathfrak{X})$ is not too large.

COROLLARY 4.1 Under the conditions of Theorem 4.3 together with assumptions (4.8) and (4.9), if $\mathrm{VC}(\mathfrak{X}) \leqslant c \cdot \frac{\delta n_1}{\log n_1}$, then the split conformal prediction interval $\widehat{C}_n$ defined in (4.4) satisfies

$$\operatorname{leb}\big(\widehat{C}_n(X_{n+1}) \triangle C^*_P(X_{n+1})\big) \leqslant c' \left(\frac{\eta_{n_0}^{1/3}}{\delta^{1/2}} + \frac{\eta_{n_0}^{1/3} + \sqrt{\frac{\mathrm{VC}(\mathfrak{X}) \log n_1}{\delta n_1}}}{f_\epsilon(q^*_{\epsilon,\alpha/2})}\right)$$

with probability at least $1 - \frac{1}{n_1} - 2\rho_{n_0} - \eta_{n_0}^{1/3}$, where c, c' are universal constants.

In other words, for a location-family model with a consistent estimate of the true mean function ($\eta_{n_0}, \rho_{n_0} \to 0$), if $\mathrm{VC}(\mathfrak{X})$ is sufficiently small then the interval $\widehat{C}_n$ defined in (4.4) is able to satisfy

restricted conditional coverage in the distribution-free setting, while matching the best possible 'oracle' prediction interval length asymptotically as $n_0, n_1 \to \infty$.

5. Discussion

In this work, we have explored the possible definitions of approximate conditional coverage for distribution-free predictive inference with the goal of finding meaningful definitions that are strong enough to achieve some of the practical benefits of conditional coverage (i.e. patients feel assured that their personalized predictions have some level of accuracy), but weak enough to still allow for the possibility of meaningful distribution-free procedures. We find that requiring $(1-\alpha, \delta)$-conditional coverage to hold, i.e. coverage at level $1-\alpha$ over every subgroup with probability at least δ within the overall population, is too strong of a condition—our main result establishes a lower bound on the resulting prediction interval length and demonstrates that meaningful procedures cannot be constructed with this property. By relaxing the desired property to $(1-\alpha, \delta, \mathfrak{X})$-conditional coverage, i.e. coverage at level $1-\alpha$ over every subgroup $\mathscr{X} \in \mathfrak{X}$ that has probability at least δ, we see that sufficiently restricting the class $\mathfrak{X}$ does allow for nontrivial prediction intervals.

Many open questions remain after our preliminary findings. In particular, what types of classes $\mathfrak{X}$ are most meaningful for defining this restricted form of approximate conditional coverage? Furthermore, for nearly any class $\mathfrak{X}$, computation for the split conformal method constructed in Section 4.1 may pose a serious challenge—how can we efficiently compute predictive intervals for this problem?

Another direction for relaxing $(1-\alpha, \delta)$-CC property is to require it to hold only over some distributions P (rather than restricting to a class $\mathfrak{X}$ of sets that we condition on). Is it possible to ensure that conditional coverage at level $1-\alpha$ holds, not at some uniform tolerance level δ, but at an adaptive tolerance level $\delta(P)$ that is low for 'well-behaved' distributions P but may be as large as 1 (i.e. only ensuring marginal coverage) for degenerate distributions P? We leave these questions for future work.

Acknowledgements

The authors are grateful to the American Institute of Mathematics for supporting and hosting our collaboration. R.F.B. was partially supported by the National Science Foundation via grant DMS–1654076 and by an Alfred P. Sloan fellowship. E.J.C. was partially supported by the Office of Naval Research under grant N00014-16-1-2712, by the National Science Foundation via grant DMS–1712800, and by a generous gift from TwoSigma. R.F.B. thanks Chao Gao, Samir Khan, and Haoyang Liu for helpful suggestions on an early version of this work.

References

1. Raghu R Bahadur and Leonard J Savage. (1956) The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Stat.*, **27**, 1115–1122.
2. Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. (1989) Learnability and the Vapnik–Chervonenkis dimension. *J. ACM*, **36**, 929–965.
3. T Tony Cai, Mark Low, and Zongming Ma. (2014) Adaptive confidence bands for nonparametric regression functions. *J. Am. Stat. Assoc.*, **109**, 1054–1070.
4. David L Donoho. (1988) One-sided inference about functionals of a density. *Ann. Stat.*, **16**, 1390–1420.
5. Richard M Dudley and Rimas Norvaiša. (2011) *Concrete Functional Calculus*. Springer.
6. Vladimir Koltchinskii. (2011) *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008, vol. 2033*. Springer Science & Business Media.

7. Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. (2018) Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.*, **113**, 1094–1111.
8. Jing Lei and Larry Wasserman. (2014) Distribution-free prediction bands for non-parametric regression. *J. Royal Stat. Soc.*, **76**, 71–96.
9. Papadopoulos, H. (2008) Inductive conformal prediction: theory and application to neural networks. *Tools in Artificial Intelligence*. Croatia: InTech. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1087.5114&rep=rep1&type=pdf
10. Papadopoulos, H., Proedrou, K., Vovk, V. & Gammerman, A. (2002) Inductive confidence machines for regression. *European Conference on Machine Learning*. Springer, pp. 345–356.
11. Glenn Shafer and Vladimir Vovk. (2008) A tutorial on conformal prediction. *J. Mach. Learn. Res.*, **9**, 371–421.
12. Wacław Sierpiński. (1922) Sur les fonctions d'ensemble additives et continues. *Fundam. Math.*, **1**, 240–246.
13. Vovk, V. (2012) Conditional validity of inductive conformal predictors. *Asian Conference on Machine Learning*. Proceedings of Machine Learning Research, pp. 475–490. http://proceedings.mlr.press/v25/vovk12/vovk12.pdf
14. Vladimir Vovk, Alex Gammerman, and Glenn Shafer. (2005) *Algorithmic Learning in a Random World*. Springer Science & Business Media.

A. Proof of main impossibility result (Theorem 3.1)

A.1 *A preliminary lemma*

In order to prove our main theorem, we rely on a key lemma:

Lemma A.1 Suppose that $\widehat{C}_n$ satisfies $(1-\alpha,\delta)$-CC as defined in (1.3). Then for all distributions P where the marginal distribution P_X has no atoms, and for all measurable sets $\mathscr{B} \subseteq \mathbb{R}^d \times \mathbb{R}$ with $P(\mathscr{B}) \geqslant \delta$, we have

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} \geqslant 1-\alpha.$$

Comparing this lemma to the definition of $(1-\alpha,\delta)$-CC, we see that the definition of approximate conditional coverage requires that the result of the lemma must hold for any set of the form $\mathscr{B} = \mathscr{X}\times\mathbb{R}$, i.e. conditioning on an event $X_{n+1} \in \mathscr{X}$ (with probability at least δ). The lemma extends the property to condition also on events that are defined jointly in (X, Y).

While this may initially appear to be a simple extension of the definition of $(1-\alpha,\delta)$-CC, the proof is not trivial, and the implications of this result are very significant. To see why, suppose that we construct $\mathscr{B}$ to consist only of points (x, y) such that $Y_{n+1} = y$ is in the extreme tail of its conditional distribution given $X_{n+1} = x$—specifically, outside the range given by the $\delta/2$ and $1-\delta/2$ conditional quantiles (so that the overall probability of $\mathscr{B}$ is large enough, i.e. $\geqslant \delta$). The lemma claims that even when (X_{n+1}, Y_{n+1}) lands in this set, i.e. Y_{n+1} is in the extreme tails of its conditional distribution given X_{n+1}, this value Y_{n+1} is still quite likely to lie in $\widehat{C}_n(X_{n+1})$. This implies that $\widehat{C}_n(X_{n+1})$ must indeed be very wide.

We will next formalize this intuition to prove our theorem.

A.2 *Proof of Theorem 3.1*

First, for each $x \in \mathbb{R}^d$ and each $s \in [0,1]$, define

$$C_{P,s}(x) = \left\{y : \mathbb{P}\left\{y \in \widehat{C}_n(x)\right\} > s\right\},$$

where the probability is taken with respect to the training data. Note that $C_{P,s}(x)$ is fixed, since it is defined as a function of the *distribution* of $\widehat{C}_n(x)$, not of the random interval $\widehat{C}_n(x)$ itself.

Next, for any fixed x, in expectation over the training data, we have

$$\mathbb{E}\left[\mathrm{leb}\big(\widehat{C}_n(x)\big)\right] = \mathbb{E}\left[\int_{y\in\mathbb{R}} \mathbf{1}\left\{y \in \widehat{C}_n(x)\right\}\, \mathrm{d}y\right] = \int_{y\in\mathbb{R}} \mathbb{P}\left\{y \in \widehat{C}_n(x)\right\}\, \mathrm{d}y,$$

by Fubini's theorem. Now, we can rewrite

$$\mathbb{P}\left\{y \in \widehat{C}_n(x)\right\} = \int_{s=0}^{1} \mathbf{1}\left\{\mathbb{P}\left\{y \in \widehat{C}_n(x)\right\} > s\right\}\, \mathrm{d}s = \int_{s=0}^{1} \mathbf{1}\left\{y \in C_{P,s}(x)\right\}\, \mathrm{d}s,$$

and so plugging this in and applying Fubini's theorem again,

$$\mathbb{E}\left[\mathrm{leb}\big(\widehat{C}_n(x)\big)\right] = \int_{s=0}^{1}\int_{y\in\mathbb{R}} \mathbf{1}\left\{y \in C_{P,s}(x)\right\}\, \mathrm{d}y\, \mathrm{d}s = \int_{s=0}^{1} \mathrm{leb}\big(C_{P,s}(x)\big)\, \mathrm{d}s.$$

Next, plugging in the test point X_{n+1} and applying Fubini's theorem an additional time,

$$\begin{aligned}\mathbb{E}\left[\mathrm{leb}\big(\widehat{C}_n(X_{n+1})\big)\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathrm{leb}\big(\widehat{C}_n(X_{n+1})\big) \mid X_{n+1}\right]\right] = \mathbb{E}\left[\int_{s=0}^{1} \mathrm{leb}\big(C_{P,s}(X_{n+1})\big)\, \mathrm{d}s\right] \\ &= \int_{s=0}^{1} \mathbb{E}\left[\mathrm{leb}\big(C_{P,s}(X_{n+1})\big)\right]\, \mathrm{d}s = \int_{s=0}^{1} \mathbb{E}_{P_X}\left[\mathrm{leb}\big(C_{P,s}(X)\big)\right]\, \mathrm{d}s, \end{aligned} \tag{A.1}$$

where the last step holds since marginally $X_{n+1} \sim P_X$.

Next, we define

$$\alpha_s = \mathbb{P}_P\left\{Y \notin C_{P,s}(X)\right\},$$

the marginal miscoverage rate of the sets $C_{P,s}(x)$ (that is, we think of $C_{P,s}(x)$ as a deterministic prediction interval). Then

$$\mathbb{E}_{P_X}\left[\mathrm{leb}\big(C_{P,s}(X)\big)\right] \geqslant L_P(1-\alpha_s) \tag{A.2}$$

by the definition of the minimal prediction interval length L_P given in (3.1). Since $s \mapsto \alpha_s$ is nondecreasing and right-continuous, and satisfies $\alpha_1 = 1$, we can define

$$s_\star = \min\{s \in [0,1] : \alpha_s \geqslant \delta\}.$$

Define also

$$\mathscr{B}_+ = \left\{(x,y) : \mathbb{P}\left\{y \in \widehat{C}_n(x)\right\} \leqslant s_\star\right\} \text{ and } \mathscr{B}_- = \left\{(x,y) : \mathbb{P}\left\{y \in \widehat{C}_n(x)\right\} < s_\star\right\}.$$

Then

$$\mathbb{P}_P\left\{(X,Y) \in \mathscr{B}_+\right\} = \alpha_{s_\star} \geqslant \delta \text{ and } \mathbb{P}_P\left\{(X,Y) \in \mathscr{B}_-\right\} = \sup_{s<s_\star} \alpha_s \leqslant \delta.$$

Now, since P is assumed to have no atoms (inheriting this property from the marginal P_X), by [5, Proposition A.1] (see also [12]), we can find a measurable set $\mathscr{B}$ such that

$$\mathscr{B}_- \subseteq \mathscr{B} \subseteq \mathscr{B}_+ \text{ and } \mathbb{P}_P\{(X,Y) \in \mathscr{B}\} = \delta.$$

By definition of $\mathscr{B}$, we have

$$\begin{aligned}(x,y) \in \mathscr{B} &\Rightarrow \mathbb{P}\left\{y \in \widehat{C}_n(x)\right\} \leqslant s_\star, \\ (x,y) \notin \mathscr{B} &\Rightarrow \mathbb{P}\left\{y \in \widehat{C}_n(x)\right\} \geqslant s_\star. \end{aligned} \tag{A.3}$$

Next, writing $a \wedge b$ to denote $\min\{a, b\}$, we can calculate

$$\begin{aligned}
\int_{s=0}^{s_\star} \alpha_s \, \mathrm{d}s &= s_\star - \int_{s=0}^{s_\star} (1-\alpha_s) \, \mathrm{d}s \\
&= s_\star - \int_{s=0}^{s_\star} \mathbb{P}_P\left\{Y \in C_{P,s}(X)\right\} \, \mathrm{d}s \\
&= s_\star - \int_{s=0}^{s_\star} \mathbb{P}_P\left\{\mathbb{P}\left\{Y \in \widehat{C}_n(X) \,\middle|\, X, Y\right\} > s\right\} \, \mathrm{d}s \\
&= s_\star - \int_{s=0}^{1} \mathbb{P}_P\left\{\mathbb{P}\left\{Y \in \widehat{C}_n(X) \,\middle|\, X, Y\right\} \wedge s_\star > s\right\} \, \mathrm{d}s \\
&= s_\star - \mathbb{E}_P\left[\mathbb{P}\left\{Y \in \widehat{C}_n(X) \,\middle|\, X, Y\right\} \wedge s_\star\right] \\
&= s_\star - \left(\mathbb{E}_P\left[\mathbb{P}\left\{Y \in \widehat{C}_n(X) \,\middle|\, X, Y\right\} \cdot \mathbf{1}\left\{(X,Y) \in \mathscr{B}\right\}\right] + \mathbb{E}_P\left[s_\star \cdot \mathbf{1}\left\{(X,Y) \notin \mathscr{B}\right\}\right]\right) \\
&= s_\star - \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} - s_\star \mathbb{P}_P\left\{(X,Y) \notin \mathscr{B}\right\} \\
&= \delta\left(s_\star - \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\}\right),
\end{aligned}$$

where the last step holds since $\mathbb{P}_P\{(X,Y) \in \mathscr{B}\} = \mathbb{P}\left\{(X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} = \delta$ by construction. Next, by applying Lemma A.1 to the set $\mathscr{B}$, we have

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} \geqslant 1-\alpha$$

and therefore

$$\int_{s=0}^{s_\star} \alpha_s \, \mathrm{d}s \leqslant \delta\left(s_\star - (1-\alpha)\right). \tag{A.4}$$

In particular, since the left-hand side is non-negative, this proves that we must have $s_\star \geqslant 1-\alpha > 0$ (we can assume that $\alpha < 1$ since otherwise the theorem holds trivially).

Now, returning to (A.1) and (A.2), we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{leb}\left(\widehat{C}_n(X_{n+1})\right)\right] &\geqslant \int_{s=0}^{1} L_P(1-\alpha_s) \, \mathrm{d}s \geqslant \int_{s=0}^{s_\star} L_P(1-\alpha_s) \, \mathrm{d}s \\
&= s_\star \int_{s=0}^{s_\star} \frac{1}{s_\star} L_P(1-\alpha_s) \, \mathrm{d}s \geqslant s_\star L_P\left(1 - \int_{s=0}^{s_\star} \frac{1}{s_\star} \alpha_s \, \mathrm{d}s\right),
\end{aligned} \tag{A.5}$$

where the last step uses Jensen's inequality, together with the fact that $\alpha \mapsto L_P(1-\alpha)$ is convex. (To verify this, let $C_P \in \mathscr{C}_P(1-\alpha)$ and $C'_P \in \mathscr{C}_P(1-\alpha')$ and then define $C''_P(x)$ as the random interval that outputs $C_P(x)$ with probability $(1-t)$ and $C'_P(x)$ with probability t. Then it is easy to verify that $C''_P \in \mathscr{C}_P(1-\alpha'')$ where $\alpha'' = (1-t)\alpha + t\alpha'$ and that $\mathbb{E}_{P_X}\left[\mathrm{leb}(C''_P(X))\right] = (1-t)\mathbb{E}_{P_X}\left[\mathrm{leb}(C_P(X))\right] + t\mathbb{E}_{P_X}\left[\mathrm{leb}(C''_P(X))\right]$. This is sufficient to establish convexity.)

Combining (A.4) and (A.5), we obtain

$$\mathbb{E}\left[\mathrm{leb}\left(\widehat{C}_n(X_{n+1})\right)\right] \geqslant s_\star L_P\left(1 - \delta\left(1 - \frac{1-\alpha}{s_\star}\right)\right),$$

since L_P is non-decreasing. Finally, define

$$c = \frac{1}{\alpha} - \frac{1-\alpha}{s_\star \alpha}.$$

Since we have verified that $1-\alpha \leqslant s_\star \leqslant 1$, this means that $c \in [0,1]$, and plugging in this choice of c, we obtain

$$\mathbb{E}\left[\mathrm{leb}\big(\widehat{C}_n(X_{n+1})\big)\right] \geqslant \frac{1-\alpha}{1-c\alpha} L_P(1-c\alpha\delta),$$

which proves the theorem.

A.3 *Proof of Lemma* A.1

Let $\delta' = \mathbb{P}_P\{(X,Y) \in \mathscr{B}\} \geqslant \delta$. We will assume that $\delta' < 1$ (since the case $\delta' = 1$ is trivial). Fix a large integer $M \geqslant n+1$. First, draw M data points $(X_0^{(1)}, Y_0^{(1)}), \ldots, (X_0^{(M)}, Y_0^{(M)})$ i.i.d. from $(X,Y) \sim P$ conditional on $(X,Y) \notin \mathscr{B}$, and M additional data points $(X_1^{(1)}, Y_1^{(1)}), \ldots, (X_1^{(M)}, Y_1^{(M)})$ i.i.d. from $(X,Y) \sim P$ conditional on $(X,Y) \in \mathscr{B}$. Let $\mathscr{L}$ denote this draw of the $2M$ data points. Since P_X has no atoms, with probability 1 all the $X_0^{(i)}$'s and $X_1^{(i)}$'s are distinct, so from this point on we assume that this is true.

Next, suppose that we draw indices $m_1, \ldots, m_{n+1}$ without replacement from the set $\{1, \ldots, M\}$. Independently for each $i = 1, \ldots, n+1$, set

$$(X_i, Y_i) = \begin{cases} (X_0^{(m_i)}, Y_0^{(m_i)}), & \text{with probability } 1-\delta', \\ (X_1^{(m_i)}, Y_1^{(m_i)}), & \text{with probability } \delta'. \end{cases} \tag{A.6}$$

We can clearly see that, after marginalizing over $\mathscr{L}$, this is equivalent to drawing the data points (X_i, Y_i) i.i.d. from P. Therefore, we have

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} = \mathbb{E}\left[\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B} \,\middle|\, \mathscr{L}\right\}\right],$$

where, on the right-hand side, after conditioning on $\mathscr{L}$, the data points (X_i, Y_i) are drawn according to (A.6).

Next, consider an alternate distribution where we draw the $n+1$ data points (X_i, Y_i) from $\mathscr{L}$ but now drawing *with* replacement. Specifically, fixing $\mathscr{L}$, let $Q(\mathscr{L})$ be the discrete distribution that places probability $\frac{1-\delta'}{M}$ on each point $(X_0^{(m)}, Y_0^{(m)})$ and probability $\frac{\delta'}{M}$ on each point $(X_1^{(m)}, Y_1^{(m)})$, for $m = 1, \ldots, M$. The product distribution $\big(Q(\mathscr{L})\big)^{n+1}$ is therefore equivalent to sampling indices $m_1, \ldots, m_{n+1}$ *with* replacement from the set $\{1, \ldots, M\}$ and then defining (X_i, Y_i) again according to (A.6).

Now, if M is very large relative to n, it is extremely unlikely that we would have $m_i = m_{i'}$ for any $i \neq i'$, when drawing from $\big(Q(\mathscr{L})\big)^{n+1}$. Specifically, we can easily check that this probability is bounded by $\frac{n^2}{M}$, and so for any fixed $\mathscr{L}$, the total variation distance between the distribution given in (A.6) (i.e. sampling without replacement) and the distribution $(Q(\mathscr{L}))^{n+1}$ (i.e. sampling with replacement) is bounded by $\frac{n^2}{M}$. Therefore,

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B} \,\middle|\, \mathscr{L}\right\} \leqslant \mathbb{P}_{(Q(\mathscr{L}))^{n+1}}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} + \frac{n^2}{M},$$

where on the left-hand side, after conditioning on $\mathscr{L}$, the data points (X_i, Y_i) are drawn according to (A.6).

Next, for any $\mathscr{L}$, define the set

$$\mathscr{X}(\mathscr{L}) = \{X_1^{(1)}, \ldots, X_1^{(M)}\}.$$

Note that, for $(X, Y) \sim Q(\mathscr{L})$, by construction we have $X \in \mathscr{X}(\mathscr{L})$ if and only if $(X, Y) \in \mathscr{B}$ (since we have assumed that $\mathscr{L}$ is chosen so that $X_0^{(1)}, \ldots, X_0^{(M)}, X_1^{(1)}, \ldots, X_1^{(M)}$ are all distinct) and

$$\mathbb{P}_{Q(\mathscr{L})}\{X \in \mathscr{X}(\mathscr{L})\} = \mathbb{P}_{Q(\mathscr{L})}\{(X, Y) \in \mathscr{B}\} = \delta' \geqslant \delta.$$

Therefore, since $\widehat{C}_n$ satisfies $(1-\alpha, \delta)$-CC with respect to any distribution, we must have

$$\begin{aligned}
&\mathbb{P}_{(Q(\mathscr{L}))^{n+1}}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} \\
&\quad = \mathbb{P}_{(Q(\mathscr{L}))^{n+1}}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), X_{n+1} \in \mathscr{X}(\mathscr{L})\right\} \\
&\quad = \mathbb{P}_{(Q(\mathscr{L}))^{n+1}}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}(\mathscr{L})\right\} \cdot \delta' \leqslant \alpha\delta'
\end{aligned}$$

for every fixed $\mathscr{L}$ where $X_0^{(1)}, \ldots, X_0^{(M)}, X_1^{(1)}, \ldots, X_1^{(M)}$ are distinct. Combining everything, therefore,

$$\begin{aligned}
&\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} \\
&\quad \leqslant \mathbb{E}\left[\mathbb{P}_{(Q(\mathscr{L}))^{n+1}}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} + \frac{n^2}{M}\right] \leqslant \alpha\delta' + \frac{n^2}{M},
\end{aligned}$$

where the expectation is taken with respect to the random draw of $\mathscr{L}$. Since M can be taken to be arbitrarily large, we therefore have

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} \leqslant \alpha\delta' = \alpha \cdot \mathbb{P}\left\{(X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\},$$

which concludes the proof of the lemma.

B. Additional proofs

B.1 *Proof of Lemma* 3.2

Let $A \sim \text{Bernoulli}\left(\frac{1-\alpha}{1-c\alpha}\right)$ be the Bernoulli variable indicating whether $\widehat{C}'_n(x)$ is defined as $\widehat{C}_n(x)$ (if $A = 1$) or as the empty set (if $A = 0$). Then, for any $\mathscr{X}$ with $P_X(\mathscr{X}) \geqslant \delta$, we have

$$\begin{aligned}
\mathbb{P}\left\{Y_{n+1} \in \widehat{C}'_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\} &= \mathbb{P}\left\{A = 1, Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\} \\
&= \frac{1-\alpha}{1-c\alpha} \cdot \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\} \geqslant \frac{1-\alpha}{1-c\alpha} \cdot (1-c\alpha) = 1-\alpha,
\end{aligned}$$

where the inequality holds since $\widehat{C}_n$ satisfies $(1-c\alpha, \delta)$-CC by Lemma 3.1.

B.2 *Proof of Theorem* 4.1

Fix any distribution P and any $\mathscr{X} \in \mathfrak{X}$ with $P_X(\mathscr{X}) \geqslant \delta$. Let

$$R_{n+1} = \left|Y_{n+1} - \widehat{\mu}_{n_0}(X_{n+1})\right|$$

be the residual of the test point. By definition of the procedure, we can see that

$$\begin{aligned}
\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\} &= \mathbb{P}\left\{R_{n+1} > \widehat{q}_{n_1}(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\} \\
&\leqslant \mathbb{P}\left\{\mathscr{X} \notin \widehat{\mathfrak{X}}_{n_1} \,\middle|\, X_{n+1} \in \mathscr{X}\right\} + \mathbb{P}\left\{R_{n+1} > \widehat{q}_{n_1}(\mathscr{X}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\}.
\end{aligned}$$

The first probability depends only on the held-out portion of the training data, i.e. data points $i = n_0 + 1, \ldots, n$. We have

$$\mathscr{X} \notin \widehat{\mathfrak{X}}_{n_1} \;\Rightarrow\; \sum_{i=n_0+1}^{n} \mathbf{1}\left\{X_i \in \mathscr{X}\right\} < \delta n_1 \left(1 - \sqrt{\frac{2\log n_1}{\delta n_1}}\right).$$

Since each X_i has probability at least δ of lying in $\mathscr{X}$, therefore this probability is bounded by

$$\mathbb{P}\left\{\text{Binomial}\,(n_1, \delta) < \delta n_1 \left(1 - \sqrt{\frac{2\log n_1}{\delta n_1}}\right)\right\} \leqslant \frac{1}{n_1},$$

where the inequality holds by the multiplicative Chernoff bound. Therefore, what we have so far is

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \;\middle|\; X_{n+1} \in \mathscr{X}\right\} \leqslant \frac{1}{n_1} + \mathbb{P}\left\{R_{n+1} > \widehat{q}_{n_1}(\mathscr{X}) \;\middle|\; X_{n+1} \in \mathscr{X}\right\}.$$

Next, let $I = \{i : n_0 + 1 \leqslant i \leqslant n, X_i \in \mathscr{X}\}$. Then $|I| = \widehat{N}_{n_1}(\mathscr{X})$, and by definition of $\widehat{q}_{n_1}(\mathscr{X})$, we see that $R_{n+1} > \widehat{q}_{n_1}(\mathscr{X})$ if and only if R_{n+1} is not one of the $\left\lceil \left(1 - \alpha + \frac{1}{n_1}\right) \cdot (|I| + 1) \right\rceil$ smallest values of $\{R_i : i \in I \cup \{n+1\}\}$. Now, after conditioning on I and on the event $X_{n+1} \in \mathscr{X}$, by distribution of the data we see that these residuals are exchangeable. Therefore, this event has probability at most

$$1 - \frac{\left\lceil \left(1 - \alpha + \frac{1}{n_1}\right) \cdot (|I| + 1) \right\rceil}{|I| + 1} \leqslant \alpha - \frac{1}{n_1}$$

after conditioning on I and on the event that $X_{n+1} \in \mathscr{X}$. This bound is therefore true also after marginalizing over I, and so $\mathbb{P}\left\{R_{n+1} > \widehat{q}_{n_1}(\mathscr{X}) \;\middle|\; X_{n+1} \in \mathscr{X}\right\} \leqslant \alpha - \frac{1}{n_1}$, which concludes the proof.

B.3 *Proof of Theorem* 4.2

First, we need to show that Lemma A.1 holds in this setting.

LEMMA B.1 Suppose that $\widehat{C}_n$ satisfies $(1 - \alpha, \delta, \mathfrak{X})$-CCE as defined in (4.6), where $\mathfrak{X}$ satisfies $\mathrm{VC}_{\text{a.e.}}(\mathfrak{X}) \geqslant 2n + 2$. Then for all distributions P where the marginal distribution P_X is continuous with respect to Lebesgue measure, for all $\mathscr{B} \subseteq \mathbb{R}^d \times \mathbb{R}$ with $\mathbb{P}_P\{(X, Y) \in \mathscr{B}\} \geqslant \delta$,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} \geqslant 1 - \alpha.$$

With this lemma in place, the proof of Theorem 4.2 follows exactly as the proof of our initial result, Theorem 3.1. We now turn to proving the lemma.

Proof. of Lemma B.1 The proof of this lemma is similar to that of Lemma A.1, except that instead of taking M samples from $\mathscr{B}$ and from $\mathscr{B}^c$ for an arbitrarily large integer M, we only need to take $n + 1$ from each set.

Let $\delta' = \mathbb{P}_P\{(X, Y) \in \mathscr{B}\} \geqslant \delta$. We can assume that $\delta' < 1$ (otherwise, the bound claimed in the lemma is trivial). Draw $n + 1$ data points $(X_0^{(1)}, Y_0^{(1)}), \ldots, (X_0^{(n+1)}, Y_0^{(n+1)})$ i.i.d. from $(X, Y) \sim P$ conditional on $(X, Y) \notin \mathscr{B}$, and $n + 1$ additional data points $(X_1^{(1)}, Y_1^{(1)}), \ldots, (X_1^{(n+1)}, Y_1^{(n+1)})$ i.i.d. from $(X, Y) \sim P$ conditional on $(X, Y) \in \mathscr{B}$. Let $\mathscr{L}$ denote this draw of the $2n + 2$ data points.

Next, we draw a permutation π of the set $\{1, \ldots, n+1\}$ uniformly at random and draw $B_1, \ldots, B_{n+1} \overset{\text{iid}}{\sim} \text{Bernoulli}(\delta')$ independently of all other random variables. Define

$$(X_i, Y_i) = \begin{cases} \left(X_0^{(\pi_i)}, Y_0^{(\pi_i)}\right), & \text{if } B_i = 0, \\ \left(X_1^{(\pi_i)}, Y_1^{(\pi_i)}\right), & \text{if } B_i = 1. \end{cases}$$

We can clearly see that, after marginalizing over $\mathscr{L}$, this is equivalent to drawing the data points (X_i, Y_i) i.i.d. from P. Therefore, we have

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} = \mathbb{E}\left[\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B} \,\middle|\, \mathscr{L}\right\}\right], \tag{B.1}$$

where, on the right-hand side, after conditioning on $\mathscr{L}$, the data points (X_i, Y_i) are defined by the permutation π and the Bernoulli variables $B_1, \ldots, B_{n+1}$.

Next, consider the distribution of the data conditional on $\mathscr{L}$, which we denote by $\tilde{P}(\mathscr{L})$. Since the permutation π is drawn uniformly at random, and the B_i's are i.i.d., it is clear that the $n+1$ data points $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable under the distribution $\tilde{P}(\mathscr{L})$. Therefore, for any fixed $\mathscr{L}$ and for any set $\mathscr{X} \in \mathfrak{X}$ with $\mathbb{P}_{\tilde{P}(\mathscr{L})}\left\{X_{n+1} \in \mathscr{X}\right\} \geqslant \delta$, the $(1-\alpha, \delta, \mathfrak{X})$-CCE property ensures that

$$\mathbb{P}_{\tilde{P}(\mathscr{L})}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}\right\} \geqslant 1 - \alpha.$$

Now, fixing $\mathscr{L}$, define the set $\mathscr{X}(\mathscr{L})$ to be any element of $\mathfrak{X}$ such that

$$\mathscr{X}(\mathscr{L}) \ni X_1^{(1)}, \ldots, X_1^{(n+1)}, \quad \mathscr{X}(\mathscr{L}) \not\ni X_0^{(1)}, \ldots, X_0^{(n+1)}.$$

(Since we have assumed that $\text{VC}_{\text{a.e.}}(\mathfrak{X}) \geqslant 2n+2$, and that P_X is continuous with respect to Lebesgue measure, such a set $\mathscr{X}(\mathscr{L}) \in \mathfrak{X}$ exists with probability one for any random draw of $\mathscr{L}$.) Note that, under the distribution $\tilde{P}(\mathscr{L})$, we have $X_{n+1} \in \mathscr{X}(\mathscr{L})$ if and only if $(X_{n+1}, Y_{n+1}) \in \mathscr{B}$ and

$$\mathbb{P}_{\tilde{P}(\mathscr{L})}\left\{(X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} = \mathbb{P}_{\tilde{P}(\mathscr{L})}\left\{X_{n+1} \in \mathscr{X}(\mathscr{L})\right\} = \mathbb{P}\left\{B_{n+1} = 1\right\} = \delta' \geqslant \delta.$$

Returning to the above, we therefore have

$$\begin{aligned} &\mathbb{P}_{\tilde{P}(\mathscr{L})}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} \\ &= \mathbb{P}_{\tilde{P}(\mathscr{L})}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}), X_{n+1} \in \mathscr{X}(\mathscr{L})\right\} \\ &= \mathbb{P}_{\tilde{P}(\mathscr{L})}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} \in \mathscr{X}(\mathscr{L})\right\} \cdot \delta' \geqslant (1-\alpha) \cdot \delta'. \end{aligned}$$

Then returning to (B.1),

$$\begin{aligned} &\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} \\ &= \mathbb{E}\left[\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B} \,\middle|\, \mathscr{L}\right\}\right] \\ &= \mathbb{E}\left[\mathbb{P}_{\tilde{P}(\mathscr{L})}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}), (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\}\right] \\ &\geqslant \mathbb{E}\left[(1-\alpha) \cdot \delta'\right] = (1-\alpha) \cdot \delta'. \end{aligned}$$

Therefore,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, (X_{n+1}, Y_{n+1}) \in \mathscr{B}\right\} \geqslant \frac{(1-\alpha) \cdot \delta'}{\delta'} = 1 - \alpha,$$

which proves the lemma. □

B.4 *Proof of Theorem* 4.3

Let $\mu = \widehat{\mu}_{n_0}$. Throughout this proof, we will condition on the data $(X_1, Y_1), \ldots, (X_{n_0}, Y_{n_0})$ and will therefore treat this model as fixed—the probability bound will hold with respect to the distribution of the n_1 holdout points (and therefore, the bound also holds after marginalizing over the initial n_0 training points).

We will first see that it is sufficient to prove that, with high probability, the following two bounds hold:

$$\mathfrak{X}_{x,+} \subseteq \widehat{\mathfrak{X}}_{n_1} \subseteq \mathfrak{X}_{x,-}, \tag{B.3}$$

where we define $\mathfrak{X}_{x,+} = \{\mathscr{X} \in \mathfrak{X} : x \in \mathscr{X}, P_X(\mathscr{X}) \geqslant \delta_+\}$ and $\mathfrak{X}_{x,-} = \{\mathscr{X} \in \mathfrak{X} : x \in \mathscr{X}, P_X(\mathscr{X}) \geqslant \delta_-\}$, and

$$q^*_{P,\mu,\alpha_+}(\mathscr{X}) \leqslant \widehat{q}_{n_1}(\mathscr{X}) \leqslant q^*_{P,\mu,\alpha_-}(\mathscr{X}) \text{ for all } \mathscr{X} \in \mathfrak{X}_{x,-}. \tag{B.3}$$

If these two statements hold, then we have

$$q^*_{P,\mu,\alpha_+,\delta_+}(x) = \sup_{\mathscr{X} \in \mathfrak{X}_{x,+}} q^*_{P,\mu,\alpha_+}(\mathscr{X}) \leqslant \sup_{\mathscr{X} \in \widehat{\mathfrak{X}}_{n_1}} q^*_{P,\mu,\alpha_+}(\mathscr{X}) \leqslant \sup_{\mathscr{X} \in \widehat{\mathfrak{X}}_{n_1}} \widehat{q}_{n_1}(\mathscr{X}) = \widehat{q}_{n_1}(x),$$

and similarly

$$q^*_{P,\mu,\alpha_-,\delta_-}(x) = \sup_{\mathscr{X} \in \mathfrak{X}_{x,-}} q^*_{P,\mu,\alpha_-}(\mathscr{X}) \geqslant \sup_{\mathscr{X} \in \widehat{\mathfrak{X}}_{n_1}} q^*_{P,\mu,\alpha_-}(\mathscr{X}) \geqslant \sup_{\mathscr{X} \in \widehat{\mathfrak{X}}_{n_1}} \widehat{q}_{n_1}(\mathscr{X}) = \widehat{q}_{n_1}(x).$$

By construction of the intervals, we therefore see that $C^*_{P,\mu,\alpha_+,\delta_+}(x) \subseteq \widehat{C}_n(x) \subseteq C^*_{P,\mu,\alpha_-,\delta_-}(x)$, which is the claim in the theorem.

Now we verify that (B.2) and (B.3) both hold with high probability. First, by [6, Bousquet bound (Section 2.3) + Theorem 3.9], we can verify the following concentration result:[5]

$$\mathbb{P}\left\{\left|\frac{\widehat{N}_{n_1}(\mathscr{X})}{n_1} - P_X(\mathscr{X})\right| \leqslant \Delta_{\text{conc}}(\mathscr{X}) \text{ for all } \mathscr{X} \in \mathfrak{X}\right\} \geqslant 1 - \frac{1}{3n_1}, \tag{B.4}$$

where for each $\mathscr{X} \in \mathfrak{X}$, we define

$$\Delta_{\text{conc}}(\mathscr{X}) = c'\left[\sqrt{P_X(\mathscr{X})} \cdot \sqrt{\frac{\text{VC}(\mathfrak{X}) \log n_1}{n_1}} + \frac{\text{VC}(\mathfrak{X}) \log n_1}{n_1}\right],$$

for a universal constant c' (not dependent on $\mathscr{X}$).

Next, for any $\mathscr{X} \in \mathfrak{X}$, define

$$\tilde{\mathscr{X}} = \left\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : x \in \mathscr{X} \text{ and } |y - \mu(x)| > q^*_{P,\mu,\alpha_-}(\mathscr{X})\right\}.$$

Lemma B.2 below will verify that

$$\text{VC}\Big(\{\tilde{\mathscr{X}} : \mathscr{X} \in \mathfrak{X}\}\Big) \leqslant \text{VC}(\mathfrak{X}) + 1.$$

[5] To obtain this bound, we need to apply [6, Bousquet bound (Section 2.3) + Theorem 3.9] $\mathcal{O}(\log n_1)$ many times, once for each class $\mathfrak{X}_j = \{\mathscr{X} \in \mathfrak{X} : P_X(\mathscr{X}) \leqslant 2^{-j}\}$, for $j = 0, 1, \ldots, \mathcal{O}(\log n_1)$ (i.e. a peeling argument).

Therefore, again applying [6, Bousquet bound (Section 2.3) + Theorem 3.9] as above, if the universal constant c' is chosen appropriately in the definition of $\Delta_{\text{conc}}(\mathscr{X})$ then it holds that

$$\mathbb{P}\left\{\left|\frac{1}{n_1}\sum_{i=n_0+1}^{n}\mathbf{1}\left\{(X_i,Y_i)\in\tilde{\mathscr{X}}\right\}-\mathbb{P}_P\left\{(X,Y)\in\tilde{\mathscr{X}}\right\}\right|\leqslant\Delta_{\text{conc}}(\mathscr{X})\quad\text{for all }\ \mathscr{X}\in\mathfrak{X}\right\}\geqslant 1-\frac{1}{3n_1}.$$

Furthermore, we can calculate that

$$\mathbb{P}_P\left\{(X,Y)\in\tilde{\mathscr{X}}\right\}=P_X(\mathscr{X})\cdot\mathbb{P}_P\left\{|Y-\mu(X)|>q^*_{P,\mu,\alpha_-}(\mathscr{X})\,\middle|\,X\in\mathscr{X}\right\}\leqslant P_X(\mathscr{X})\cdot\alpha_-,$$

by definition of the quantile $q^*_{P,\mu,\alpha_-}(\mathscr{X})$. Therefore,

$$\mathbb{P}\left\{\frac{1}{n_1}\sum_{i=n_0+1}^{n}\mathbf{1}\left\{X_i\in\mathscr{X},|Y_i-\mu(X_i)|>q^*_{P,\mu,\alpha_-}(\mathscr{X})\right\}\leqslant\alpha_-P_X(\mathscr{X})+\Delta_{\text{conc}}(\mathscr{X})\quad\forall\,\mathscr{X}\in\mathfrak{X}\right\}\geqslant 1-\frac{1}{3n_1}. \tag{B.5}$$

An analogous argument can be used to prove that

$$\mathbb{P}\left\{\frac{1}{n_1}\sum_{i=n_0+1}^{n}\mathbf{1}\left\{X_i\in\mathscr{X},|Y_i-\mu(X_i)|\geqslant q^*_{P,\mu,\alpha_+}(\mathscr{X})\right\}\geqslant\alpha_+P_X(\mathscr{X})-\Delta_{\text{conc}}(\mathscr{X})\quad\forall\,\mathscr{X}\in\mathfrak{X}\right\}\geqslant 1-\frac{1}{3n_1}. \tag{B.6}$$

Now from this point on, we will assume that the events in (B.4), (B.5) and (B.6) all hold, which will occur with probability at least $1-\frac{1}{n_1}$. We now need to verify that this implies (B.2) and (B.3).

First, we verify (B.2). For any $\mathscr{X}\in\widehat{\mathfrak{X}}_{n_1}$, by definition of $\widehat{\mathfrak{X}}_{n_1}$ together with (B.4) we have

$$\delta\left(1-\sqrt{\frac{2\log n_1}{\delta n_1}}\right)\leqslant\frac{1}{n_1}\sum_{i=n_0+1}^{n}\mathbf{1}\left\{X_i\in\mathscr{X}\right\}=\frac{\widehat{N}_{n_1}(\mathscr{X})}{n_1}\leqslant P_X(\mathscr{X})+\Delta_{\text{conc}}(\mathscr{X}).$$

Examining the definition of δ_-, we see that this implies $P_X(\mathscr{X})\geqslant\delta_-$ when the universal constant c_δ is chosen to be sufficiently large. This proves that $\widehat{\mathfrak{X}}_{n_1}\subseteq\mathfrak{X}_{x,-}$. Conversely, for any $\mathscr{X}\in\mathfrak{X}_{x,+}$, applying (B5) and again assuming c_δ is chosen appropriately, we have

$$\frac{1}{n_1}\sum_{i=n_0+1}^{n}\mathbf{1}\left\{X_i\in\mathscr{X}\right\}=\frac{\widehat{N}_{n_1}(\mathscr{X})}{n_1}\geqslant P_X(\mathscr{X})-\Delta_{\text{conc}}(\mathscr{X})\geqslant\delta_+-\Delta_{\text{conc}}(\mathscr{X})\geqslant\delta\left(1-\sqrt{\frac{2\log n_1}{\delta n_1}}\right),$$

and so $\mathscr{X}\in\widehat{\mathfrak{X}}_{n_1}$. This proves that $\widehat{\mathfrak{X}}_{n_1}\supseteq\mathfrak{X}_{x,+}$.

Next, we verify (B.3). Fix any $\mathscr{X} \in \mathfrak{X}_{x,-}$. By the events in (B.4) and (B.5), we have

$$\begin{aligned}
&\sum_{i=n_0+1}^{n} \mathbf{1}\left\{X_i \in \mathscr{X},\ |Y_i - \mu(X_i)| > q^*_{P,\mu,\alpha_-}(\mathscr{X})\right\} \\
&\quad \leqslant n_1\alpha_- P_X(\mathscr{X}) + n_1\Delta_{\mathrm{conc}}(\mathscr{X}) \leqslant n_1\alpha_-\left(\frac{\widehat{N}_{n_1}(\mathscr{X})}{n_1} + \Delta_{\mathrm{conc}}(\mathscr{X})\right) + n_1\Delta_{\mathrm{conc}}(\mathscr{X}) \\
&\quad \leqslant \widehat{N}_{n_1}(\mathscr{X})\left(\alpha_- + \frac{2\Delta_{\mathrm{conc}}(\mathscr{X})}{\frac{1}{n_1}\widehat{N}_{n_1}(\mathscr{X})}\right) \leqslant \widehat{N}_{n_1}(\mathscr{X})\left(\alpha_- + \frac{2\Delta_{\mathrm{conc}}(\mathscr{X})}{P_X(\mathscr{X}) - \Delta_{\mathrm{conc}}(\mathscr{X})}\right).
\end{aligned}$$

By definition of $\Delta_{\mathrm{conc}}(\mathscr{X})$, together with the assumption that $\mathrm{VC}(\mathfrak{X}) \leqslant c \cdot \frac{\delta n_1}{\log n_1}$, if c is chosen to be sufficiently small then, since $P_X(\mathscr{X}) \geqslant \delta_-$, we have

$$\frac{2\Delta_{\mathrm{conc}}(\mathscr{X})}{P_X(\mathscr{X}) - \Delta_{\mathrm{conc}}(\mathscr{X})} \leqslant c''\sqrt{\frac{\mathrm{VC}(\mathfrak{X})\log n_1}{\delta n_1}},$$

where c'' is another universal constant. Furthermore, by definition of α_-, it holds that

$$\widehat{N}_{n_1}(\mathscr{X})\left(\alpha_- + c''\sqrt{\frac{\mathrm{VC}(\mathfrak{X})\log n_1}{\delta n_1}}\right) \leqslant \widehat{N}_{n_1}(\mathscr{X}) - \left\lceil\left(1 - \alpha + \frac{1}{n_1}\right)\cdot\left(\widehat{N}_{n_1}(\mathscr{X}) + 1\right)\right\rceil$$

as long as the constant c_α is chosen to be sufficiently large. Combining these calculations, we see that

$$\sum_{i=n_0+1}^{n} \mathbf{1}\left\{X_i \in \mathscr{X},\ |Y_i - \mu(X_i)| \leqslant q^*_{P,\mu,\alpha_-}(\mathscr{X})\right\} \geqslant \left\lceil\left(1 - \alpha + \frac{1}{n_1}\right)\cdot\left(\widehat{N}_{n_1}(\mathscr{X}) + 1\right)\right\rceil.$$

Since $\widehat{q}_{n_1}(\mathscr{X})$ is defined as the $\lceil\left(1-\alpha+\frac{1}{n_1}\right)\cdot(\widehat{N}_{n_1}(\mathscr{X})+1)\rceil$-th smallest value in the list $\{R_i : n_0+1 \leqslant i \leqslant n, X_i \in \mathscr{X}\}$, the above bound immediately verifies that

$$\widehat{q}_{n_1}(\mathscr{X}) \leqslant q^*_{P,\mu,\alpha_-}(\mathscr{X}).$$

We can similarly show that, if the events in (B.4) and (B.6) both hold, then

$$\begin{aligned}
&\sum_{i=n_0+1}^{n} \mathbf{1}\left\{X_i \in \mathscr{X},\ |Y_i - \mu(X_i)| \geqslant q^*_{P,\mu,\alpha_+}(\mathscr{X})\right\} \\
&\quad \geqslant \widehat{N}_{n_1}(\mathscr{X})\left(\alpha_+ - \frac{2\Delta_{\mathrm{conc}}(\mathscr{X})}{P_X(\mathscr{X}) - \Delta_{\mathrm{conc}}(\mathscr{X})}\right) > \widehat{N}_{n_1}(\mathscr{X})\cdot\alpha,
\end{aligned}$$

and by definition of $\widehat{q}_{n_1}(\mathscr{X})$ this is sufficient to establish that

$$\widehat{q}_{n_1}(\mathscr{X}) \geqslant q^*_{P,\mu,\alpha_+}(\mathscr{X}).$$

Therefore, combining everything, we have shown that (B.2) and (B.3) both hold whenever the events in (B.4), (B.5) and (B.6) all hold, which occurs with probability at least $1 - \frac{1}{n_1}$. This completes the proof of the theorem.

B.4.1 *Supporting lemma*

LEMMA B.2 Let $\mathfrak{X}$ be any collection of measurable subsets of $\mathbb{R}^d$, and let $c : \mathfrak{X} \to \mathbb{R}$ be any function. Fix any function $f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, and for each $\mathscr{X} \in \mathfrak{X}$ define

$$\tilde{\mathscr{X}} = \left\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : x \in \mathscr{X} \text{ and } f(x, y) > c(\mathscr{X})\right\}.$$

Then

$$\mathrm{VC}\Big(\{\tilde{\mathscr{X}} : \mathscr{X} \in \mathfrak{X}\}\Big) \leqslant \mathrm{VC}(\mathfrak{X}) + 1.$$

Proof. To see this, suppose $\mathrm{VC}(\{\tilde{\mathscr{X}} : \mathscr{X} \in \mathfrak{X}\}) = m$. If $m = 1$ then the result is trivial, so assume $m \geqslant 2$. We can then find m points $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, for $i = 1, \ldots, m$, which are shattered by $\{\tilde{\mathscr{X}} : \mathscr{X} \in \mathfrak{X}\}$. Without loss of generality, assume that $f(x_m, y_m) = \min_{i=1,\ldots,m} f(x_i, y_i)$. We will now show that the set $\{x_1, \ldots, x_{m-1}\}$ is shattered by $\mathfrak{X}$. Fix any subset $I \subseteq \{1, \ldots, m-1\}$, and let $\tilde{I} = I \cup \{m\}$. Then since $\{\tilde{\mathscr{X}} : \mathscr{X} \in \mathfrak{X}\}$ shatters $(x_1, y_1), \ldots, (x_m, y_m)$, there must be some $\mathscr{X} \in \mathfrak{X}$ such that $(x_i, y_i) \in \tilde{\mathscr{X}}$ for $i \in \tilde{I}$ and $(x_i, y_i) \notin \tilde{\mathscr{X}}$ for $i \notin \tilde{I}$. In particular, taking $i = m \in \tilde{I}$, we have

$$(x_m, y_m) \in \tilde{\mathscr{X}} \quad \Rightarrow \quad f(x_m, y_m) > c(\mathscr{X}) \quad \Rightarrow \quad f(x_i, y_i) > c(\mathscr{X}) \text{ for all } i.$$

Now, for all $i \in I$,

$$i \in \tilde{I} \quad \Rightarrow \quad (x_i, y_i) \in \tilde{\mathscr{X}} \quad \Rightarrow \quad x_i \in \mathscr{X},$$

and for all $i \in \{1, \ldots, m-1\} \backslash I$, we know that $f(x_i, y_i) > c(\mathscr{X})$ and therefore

$$i \notin \tilde{I} \quad \Rightarrow \quad x_i \notin \mathscr{X}.$$

Since we can find such a set $\mathscr{X}$ for each subset $I \subseteq \{1, \ldots, m-1\}$, this means that $\mathfrak{X}$ shatters $\{x_1, \ldots, x_{m-1}\}$, and therefore $\mathrm{VC}(\mathfrak{X}) \geqslant m - 1$, completing the proof. □

B.5 *Proof of Corollary 4.1*

Recall that the oracle interval is given by

$$C_P^*(X_{n+1}) = \mu_P(X_{n+1}) \pm q^*_{\epsilon,\alpha},$$

where $q^*_{\epsilon,\alpha}$ is the $(1 - \alpha/2)$-quantile of f_ϵ. By Theorem 4.3, for every $x \in \mathbb{R}^d$ we have

$$\mathbb{P}\left\{C^*_{P,\widehat{\mu}_{n_0},\alpha_+,\delta_+}(x) \subseteq \widehat{C}_n(x) \subseteq C^*_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}(x)\right\} \geqslant 1 - \frac{1}{n_1},$$

where $\alpha_+, \alpha_-, \delta_+, \delta_-$ are defined as in the statement of that theorem. Therefore, it must also hold that

$$\mathbb{P}\left\{C^*_{P,\widehat{\mu}_{n_0},\alpha_+,\delta_+}(X_{n+1}) \subseteq \widehat{C}_n(X_{n+1}) \subseteq C^*_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}(X_{n+1})\right\} \geqslant 1 - \frac{1}{n_1},$$

and so with probability at least $1 - \frac{1}{n_1}$, we have

$$\mathrm{leb}\big(\widehat{C}_n(X_{n+1}) \,\triangle\, C_P^*(X_{n+1})\big) \leqslant \mathrm{leb}\big(C^*_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}(X_{n+1}) \backslash C_P^*(X_{n+1})\big) + \mathrm{leb}\big(C_P^*(X_{n+1}) \backslash C^*_{P,\widehat{\mu}_{n_0},\alpha_+,\delta_+}(X_{n+1})\big).$$

Now we bound these two terms. We can calculate deterministically that

$$\mathrm{leb}\big(C^*_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}(X_{n+1}) \backslash C_P^*(X_{n+1})\big) \leqslant |\widehat{\mu}_{n_0}(X_{n+1}) - \mu_P(X_{n+1})| + 2 \max\left\{q^*_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}(X_{n+1}) - q^*_{\epsilon,\alpha}, 0\right\}$$

and

$$\mathrm{leb}\big(C_P^*(X_{n+1}) \backslash C_{P,\widehat{\mu}_{n_0},\alpha_+,\delta_+}^*(X_{n+1})\big) \leqslant |\widehat{\mu}_{n_0}(X_{n+1}) - \mu_P(X_{n+1})| + 2\max\big\{q_{\epsilon,\alpha}^* - q_{P,\widehat{\mu}_{n_0},\alpha_+,\delta_+}^*(X_{n+1}), 0\big\}.$$

Therefore, with probability at least $1 - \frac{1}{n_1}$, we have

$$\begin{aligned}\mathrm{leb}\big(\widehat{C}_n(X_{n+1}) \vartriangle C_P^*(X_{n+1})\big) \leqslant\ & 2|\widehat{\mu}_{n_0}(X_{n+1}) - \mu_P(X_{n+1})| + 2\max\big\{q_{\epsilon,\alpha}^* - q_{P,\widehat{\mu}_{n_0},\alpha_+,\delta_+}^*(X_{n+1}), 0\big\} \\ & + 2\max\big\{q_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}^*(X_{n+1}) - q_{\epsilon,\alpha}^*, 0\big\},\end{aligned}$$

so we now need to bound these remaining terms with high probability.

First, we bound $|\widehat{\mu}_{n_0}(X_{n+1}) - \mu_P(X_{n+1})|$. Define

$$\widehat{\Delta}_{n_0} = \mathbb{E}\Big[\big(\widehat{\mu}_{n_0}(X) - \mu_P(X)\big)^2 \,\Big|\, \widehat{\mu}_{n_0}\Big],$$

which satisfies $\mathbb{P}\big\{\widehat{\Delta}_{n_0} \leqslant \eta_{n_0}\big\} \geqslant 1 - \rho_{n_0}$ by (4.9). We have

$$\begin{aligned}\mathbb{P}\Big\{|\widehat{\mu}_{n_0}(X_{n+1}) - \mu_P(X_{n+1})| > \eta_{n_0}^{1/3}\Big\} &= \mathbb{E}\Big[\mathbb{P}\Big\{|\widehat{\mu}_{n_0}(X_{n+1}) - \mu_P(X_{n+1})| > \eta_{n_0}^{1/3} \,\Big|\, \widehat{\mu}_{n_0}\Big\}\Big] \\ &\leqslant \mathbb{E}\left[\min\left\{\frac{\mathbb{E}\big[(\widehat{\mu}_{n_0}(X_{n+1}) - \mu_P(X_{n+1}))^2 \,\big|\, \widehat{\mu}_{n_0}\big]}{\eta_{n_0}^{2/3}}, 1\right\}\right] \\ &= \mathbb{E}\left[\min\left\{\frac{\widehat{\Delta}_{n_0}}{\eta_{n_0}^{2/3}}, 1\right\}\right] \leqslant \rho_{n_0} + \frac{\eta_{n_0}}{\eta_{n_0}^{2/3}}.\end{aligned}$$

Therefore, with probability at least $1 - \frac{1}{n_1} - \rho_{n_0} - \eta_{n_0}^{1/3}$, we have

$$\begin{aligned}\mathrm{leb}\big(\widehat{C}_n(X_{n+1}) \vartriangle C_P^*(X_{n+1})\big) \leqslant\ & 2\eta_{n_0}^{1/3} + 2\max\big\{q_{\epsilon,\alpha}^* - q_{P,\widehat{\mu}_{n_0},\alpha_+,\delta_+}^*(X_{n+1}), 0\big\} \\ & + 2\max\big\{q_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}^*(X_{n+1}) - q_{\epsilon,\alpha}^*, 0\big\}.\end{aligned}$$

Next, since $\mathbb{R}^d \in \mathfrak{X}$ by assumption, by definition we have

$$q_{P,\widehat{\mu}_{n_0},\alpha_+,\delta_+}^*(X_{n+1}) = \sup_{\mathscr{X}\in\mathfrak{X}: X_{n+1}\in\mathscr{X}, P_X(\mathscr{X})\geqslant\delta_+} q_{P,\widehat{\mu}_{n_0},\alpha_+}^*(\mathscr{X}) \geqslant q_{P,\widehat{\mu}_{n_0},\alpha_+}^*(\mathbb{R}^d) \geqslant q_{\epsilon,\alpha_+}^*,$$

where the last step uses the location family assumption (4.8). Therefore, with probability at least $1 - \frac{1}{n_1} - \rho_{n_0} - \eta_{n_0}^{1/3}$, we have

$$\mathrm{leb}\big(\widehat{C}_n(X_{n+1}) \vartriangle C_P^*(X_{n+1})\big) \leqslant 2\eta_{n_0}^{1/3} + 2\big(q_{\epsilon,\alpha}^* - q_{\epsilon,\alpha_+}^*\big) + 2\max\big\{q_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}^*(X_{n+1}) - q_{\epsilon,\alpha}^*, 0\big\}.$$

We now address the last term. By definition, we have

$$q_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}^*(X_{n+1}) = \sup_{\mathscr{X}\in\mathfrak{X}: X_{n+1}\in\mathscr{X}, P_X(\mathscr{X})\geqslant\delta_-} q_{P,\widehat{\mu}_{n_0},\alpha_-}^*(\mathscr{X}) \leqslant \sup_{\mathscr{X}\in\mathfrak{X}: P_X(\mathscr{X})\geqslant\delta_-} q_{P,\widehat{\mu}_{n_0},\alpha_-}^*(\mathscr{X}).$$

By the location family assumption (4.8) we can see that, for any $\mathscr{X}$,

$$q_{P,\widehat{\mu}_{n_0},\alpha_-}^*(\mathscr{X}) \leqslant \inf_{0<\alpha'<\alpha_-}\left\{q_{\epsilon,\alpha_- - \alpha'}^* + \begin{array}{c}\text{the } (1-\alpha')\text{-quantile of } |\widehat{\mu}_{n_0}(X) - \mu_P(X)| \\ \text{conditional on } \widehat{\mu}_{n_0} \text{ and on the event } X \in \mathscr{X}\end{array}\right\}.$$

And, for any $\mathscr{X}$ with $P_X(\mathscr{X}) \geqslant \delta_-$, this last quantile is bounded by

$$\sqrt{\frac{\mathbb{E}\left[(\widehat{\mu}_{n_0}(X) - \mu_P(X))^2 \mid \widehat{\mu}_{n_0}, X \in \mathscr{X}\right]}{\alpha'}} \leqslant \sqrt{\frac{\widehat{\Delta}_{n_0}}{\alpha'\delta_-}}.$$

Therefore, choosing $\alpha' = \eta_{n_0}^{1/3}$,

$$q^*_{P,\widehat{\mu}_{n_0},\alpha_-,\delta_-}(X_{n+1}) \leqslant q^*_{\epsilon,\alpha_- - \eta_{n_0}^{1/3}} + \sqrt{\frac{\widehat{\Delta}_{n_0}}{\eta_{n_0}^{1/3}\delta_-}} \leqslant q^*_{\epsilon,\alpha_- - \eta_{n_0}^{1/3}} + \eta_{n_0}^{1/3}\delta_-^{-1/2},$$

where the last bound holds with probability at least $1 - \rho_{n_0}$ by (4.9). Combining everything, with probability at least $1 - \frac{1}{n_1} - 2\rho_{n_0} - \eta_{n_0}^{1/3}$, we have

$$\mathrm{leb}\big(\widehat{C}_n(X_{n+1}) \triangle C^*_P(X_{n+1})\big) \leqslant 2\eta_{n_0}^{1/3} + 2\big(q^*_{\epsilon,\alpha_- - \eta_{n_0}^{1/3}} - q^*_{\epsilon,\alpha_+}\big) + 2\eta_{n_0}^{1/3}\delta_-^{-1/2}.$$

Finally, by our assumptions (4.8) on the density f_ϵ and the definition of $q^*_{\epsilon,\cdot}$, for any $\alpha' < \alpha'' \in [0,1]$ we have

$$\frac{1}{2}(\alpha'' - \alpha') = \int_{t=q^*_{\epsilon,\alpha''}}^{q^*_{\epsilon,\alpha'}} f_\epsilon(t)\,\mathrm{d}t \geqslant f_\epsilon(q^*_{\epsilon,\alpha'}) \cdot \big(q^*_{\epsilon,\alpha'} - q^*_{\epsilon,\alpha''}\big).$$

Therefore,

$$q^*_{\epsilon,\alpha_- - \eta_{n_0}^{1/3}} - q^*_{\epsilon,\alpha_+} \leqslant \frac{\alpha_+ - \big(\alpha_- - \eta_{n_0}^{1/3}\big)}{2f_\epsilon(q^*_{\epsilon,\alpha_- - \eta_{n_0}^{1/3}})},$$

which completes the proof for constants c, c' chosen appropriately.