Predictive inference in multi-environment scenarios

John C. Duchi[1], Suyash Gupta[2], Kuanhao Jiang[3], and Pragya Sur[3]

[1]Departments of Statistics and Electrical Engineering, Stanford University
[2]Amazon
[3]Department of Statistics, Harvard University

March 21, 2024

Abstract

We address the challenge of constructing valid confidence intervals and sets in problems of prediction across multiple environments. We investigate two types of coverage suitable for these problems, extending the jackknife and split-conformal methods to show how to obtain distribution-free coverage in such non-traditional, hierarchical data-generating scenarios. Our contributions also include extensions for settings with non-real-valued responses and a theory of consistency for predictive inference in these general problems. We demonstrate a novel resizing method to adapt to problem difficulty, which applies both to existing approaches for predictive inference with hierarchical data and the methods we develop; this reduces prediction set sizes using limited information from the test environment, a key to the methods' practical performance, which we evaluate through neurochemical sensing and species classification datasets.

1 Introduction

In the predictive inference problem, a statistician observes a training sample $\{(X_i, Y_i)\}_{i=1}^n$ of size n and wishes to predict the unknown value of Y_{n+1} at a test point X_{n+1}, where in the classical setting, $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable random variables. Vovk et al.'s *conformal prediction* [26] addresses this problem, even in finite sample and distribution-free settings, constructing a prediction band $\widehat{C}$ such that $\widehat{C}(X_{n+1})$ covers Y_{n+1} with a desired probability on average over the draw (X_{n+1}, Y_{n+1}).

In contemporary problems, however, the statistician rarely observes data from a single set of identically distributed training examples. She often has access to data that implicitly or explicitly arises from multiple environments. For instance, a neuroscientist investigating diseases of the nervous system may use multiple electrodes to measure neurotransmitter levels, with a goal to predict these levels at future time points. Variations—whether in voltage potentials, experimental conditions, build of the electrode, or otherwise—yields data from different electrodes that follow distinct distributions [18]. To understand the neurobiological underpinnings of decision making, the statistician must leverage information from multiple electrodes to develop a robust prediction model that alleviates spurious electrode-to-electrode variations. Even in cases in which one tries to exactly replicate data generating methodology, distribution shift effectively means that prediction methods lose substantial accuracy [20, 23]. In this paper, we investigate and develop methodology for constructing prediction intervals for such multi-environment problems.

1.1 Problem Setting

Data from multiple environments should improve predictions on a target only if the training and test data share common characteristics. To model this, we operate under a framework of hierarchical sampling [19, 16, 9], where one assumes that data from different environments arise from a common hierarchical model. Specifically, let $\mathcal{P}$ be a set of distributions on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is the outcome/response space and $\mathcal{X}$ is the space of covariates. Let μ denote a probability distribution on $\mathcal{P} \times \mathbb{N}$. Consider a sequence of exchangeable random pairs

$$(P^1_{XY}, n_1), \ldots, (P^{m+1}_{XY}, n_{m+1}) \tag{1}$$

generated from μ, where in training we observe i.i.d. samples

$$(X^i, Y^i) \coloneqq \{X^i_j, Y^i_j\}_{j=1}^{n_i}, \quad (X^i_j, Y^i_j) \overset{\text{iid}}{\sim} P^i_{XY}, \; j \in [n_i],$$

for $i \in [m]$, treating P^i_{XY} as the ith environment. In the test, we observe an i.i.d. sample $\{X^{m+1}_j\}_{j=1}^{n_{m+1}}$ generated from the marginal distribution on X according to the $(m+1)$st P^{m+1}_{XY}, and we wish to construct confidence sets for the unknown responses $\{Y^{m+1}_j\}_{j=1}^{n_{m+1}}$ with valid coverage. We depart from the traditional predictive inference literature in the sense that individual observations are not exchangeable. Instead, we focus on a more flexible assumption: within each environment, the observations are exchangeable, and the environments themselves are exchangeable as well. This weaker assumption allows addressing scenarios where the data may exhibit variations across different environments, but defining valid coverage therefore requires careful consideration.

To that end, we consider two plausible coverage notions for multi-environment settings. The first considers properties close to those conformal coverage guarantees [26, 2] provide: we seek a confidence set $\widehat{C}$ that covers a single example with a prescribed probability.

Definition 1.1. *A confidence set mapping $\widehat{C} : \mathcal{X} \rightrightarrows \mathcal{Y}$ provides $1-\alpha$ hierarchical coverage in the setting (1) if for the single observation $(X^{m+1}_1, Y^{m+1}_1) \sim P^{m+1}_{XY}$,*

$$\mathbb{P}\left(Y^{m+1}_1 \in \widehat{C}(X^{m+1}_1)\right) \geq 1 - \alpha. \tag{2}$$

Dunn et al. [9] and Lee et al. [16] both adopt the guarantee (2) as a notion of coverage in hierarchical data generation scenarios. Instead of this marginal guarantee over a single observation from the new environment $m+1$—and given the setting (1) that we expect to collect multiple observations from each environment—it is also natural to consider coverage notions over entire new samples $(X^m, Y^m) = \{(X^{m+1}_j, Y^{m+1}_j)\}_{j=1}^{n_{m+1}}$. We therefore define the following stronger coverage property.

Definition 1.2. *A confidence set mapping $\widehat{C} : \mathcal{X} \rightrightarrows \mathcal{Y}$ provides distribution-free level (α, δ)-coverage if for $(X^{m+1}_j, Y^{m+1}_j) \overset{\text{iid}}{\sim} P^{m+1}_{XY}$, $j = 1, \ldots, n$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n} 1\left\{Y^{m+1}_j \in \widehat{C}(X^{m+1}_j)\right\} \geq 1 - \alpha\right) \geq 1 - \delta. \tag{3}$$

That is, the confidence set covers a $1-\alpha$ fraction of observed examples in the new environment with probability at least $1-\delta$; taking $n = 1$ shows that (2) follows (3).

1.2 Main Contributions

Our main contributions include the following.

1. We introduce the multi-environment jackknife and split conformal methods (Algorithms 1 and 2, respectively). Theorems 1 and 2 establish that these provide distribution-free level (α, δ)-coverage (3) when the response space $\mathcal{Y} = \mathbb{R}$. These algorithms, as well as other procedures for hierarchical predictive inference, straightforwardly extend (see Algorithms 5 and 6) to general response spaces $\mathcal{Y}$, and they continue to provide (α, δ)-coverage (see Theorems 3 and 4).

2. Our experiments indicate that multi-environment algorithms can be conservative. To mitigate this, we propose a novel resizing method (Algorithm 7) to reduce the size of prediction sets given access to a limited amount of information from the test environment. This strategy appears to be useful for predictive inference problems more generally; for example, applying the resizing idea to Lee et al.'s algorithms [16], we observe notable set size reduction.

3. We develop new consistency theory (see Theorems 6 and 7) for multi-environment predictive inference, showing how both multi-environment jackknife and split conformal methods produce consistent confidence sets.

4. We investigate the behavior of both our algorithms, which target (α, δ)-coverage (3), and previous work on hierarchical conformal inference [16], evaluating methods on neurochemical sensing [18] and species classification [15, 6] data. Our experiments reveal that the multi-environment jackknife tends to yield smaller confidence sets than the split-conformal methodology when there are fewer training environments. Conversely, the multi-environment split conformal method demonstrates better performance when the number of training environments is large.

1.3 Related Work

Standard predictive inference methods include split-conformal [26, 17, 7, 22] and modified jackknife procedures [4]. The current paper extends these to multi-environment problems. Split conformal prediction separates the data into a training and a calibration set, using the training data to fit a model and the calibration data to set a threshold for constructing prediction intervals. Since it only splits the data once, the method may sacrifice statistical efficiency for computational gains. Addressing this issue, jackknife-style procedures use all available data for training and calibration by fitting leave-one-out models, increasing computational cost for accuracy. Both methods require exchangeability of the entire observed data to ensure valid coverage, while multi-environment methods work under the weaker assumption that within (but not across) environments, observations are exchangeable, and the environments themselves are exchangeable.

Recognizing the challenges inherent in collecting data, implicitly or explicitly, across multiple environments, a recent literature considers conformal prediction under hierarchical models, assuming the multi-environment setting (1). Among these, both Dunn et al. [9] and Lee et al. [16] study conformal prediction under hierarchical sampling and propose methods satisfying the marginal coverage guarantee (2), as well as a few other distribution-free guarantees, which may be a satisfying coverage guarantee for many applications. The (α, δ)-coverage condition (3) requires coverage for multiple observations in the test environment simultaneously, necessitating new development, as it is unclear if existing multi-level conformal approaches [9, 16] satisfy it.

1.4 Paper Outline

The remainder of this paper is organized as follows. Section 2 introduces multi-environment jackknife-minmax and multi-environment split conformal. Section 3 presents a general formulation of confidence sets and extends our methods to settings beyond regression. Section 4 proposes a resizing method for reducing the average length of prediction sets. Section 5 develops a consistency theory for the general formulation of confidence sets introduced in Section 3. Section 6 applies our methods to neurochemical sensing and species classification data [18, 6]. Appendix A contains technical proofs and Appendix B contains additional simulations.

2 Methods for regression

To introduce our basic methods, we assume the target space $\mathcal{Y} = \mathbb{R}$. In this case, we wish to return a confidence set $C : \mathcal{X} \rightrightarrows \mathbb{R}$, and typically $C(x)$ is an interval. We define a fitting algorithm A to be a function that takes a collection of samples as input and outputs an element of $\mathcal{F} \subset \mathcal{X} \to \mathcal{Y}$. To describe our algorithms formally, we introduce the following mappings.

Definition 2.1. *For values v_i, $i = 1, \ldots, n$, define the quantile mappings*

$$\begin{aligned}
\widehat{q}^{+}_{n,\alpha}(\{v_i\}) &:= \textit{the } \lceil (1-\alpha)(n+1) \rceil \textit{ th smallest value of } v_1, \ldots, v_n, \\
\widehat{q}^{-}_{n,\alpha}(\{v_i\}) &:= \textit{the } \lfloor \alpha(n+1) \rfloor \textit{ th smallest value of } v_1, \ldots, v_n = -\widehat{q}^{+}_{n,\alpha}(\{-v_i\}).
\end{aligned}$$

We also recall the notation (1) that the ith sample $(X^i, Y^i) = \{(X^i_j, Y^i_j)\}_{j=1}^{n_i}$.

2.1 Multi-environment Jackknife-minmax

We first introduce a multi-environment version of Barber et al.'s jackknife-minmax [4]. The idea is simple: we repeatedly fit a predictor $\widehat{f}_{-i}$ to all environments *except* environment i, then evaluate residuals on environment i to gauge the typical variability while predicting on a new environment. We define $\widehat{f} = \mathsf{A}((X^1, Y^1), \ldots, (X^m, Y^m))$ to be the predictor we "would" fit given all environments and consider the leave-one-out predictors

$$\widehat{f}_{-i} := \mathsf{A}\left((X^1, Y^1), \ldots, (X^{i-1}, Y^{i-1}), (X^{i+1}, Y^{i+1}), \ldots, (X^m, Y^m)\right).$$

From these, we construct the leave-one-out residuals for each example $j = 1, \ldots, n_i$ in environment i, letting

$$R^i_j = |Y^i_j - \widehat{f}_{-i}(X^i_j)| \;\text{ for } j \in [n_i],\; i \in [m].$$

We then pursue a blocked confidence set construction. Within each environment, we let $S^i_{1-\alpha} = \widehat{q}^{+}_{n_i,\alpha}(\{R^i_j\}_{j=1}^{n_i})$ be the $1-\alpha$ quantile of the residuals for predicting in environment i using $\widehat{f}_{-i}$. To obtain an interval that is likely to cover, we use quantiles of these residuals *across* environments, as the environments are exchangeable. We therefore construct intervals of the form

$$C(x) := \left[f_{\text{low}}(x) - \widehat{q}^{+}_{m,\delta}(\{S^i_{1-\alpha}\}_i), f_{\text{high}}(x) + \widehat{q}^{+}_{m,\delta}(\{S^i_{1-\alpha}\}_i)\right],$$

where all that remains is to choose f_{low} and f_{high} to obtain valid coverage. Algorithm 1 achieves this, using the minimum and maximum values of the held-out predictions to construct the multi-environment jackknife-minmax confidence set.

```
Algorithm 1: Multi-environment Jackknife-minmax: the regression case
Input: samples $\{X_j^i, Y_j^i\}_{j=1}^{n_i}$, $i = 1, \ldots, m$, confidence levels $\alpha, \delta$
For $i = 1, \ldots, m$, set
    $\widehat{f}_{-i} = \mathsf{A}\left((X^1, Y^1), \ldots, (X^{i-1}, Y^{i-1}), (X^{i+1}, Y^{i+1}), \ldots, (X^m, Y^m)\right),$
and construct residual quantiles
    $R_j^i = |Y_j^i - \widehat{f}_{-i}(X_j^i)|, \quad j = 1, \ldots, n_i, \quad \text{and} \quad S_{1-\alpha}^i = \widehat{q}_{n_i,\alpha}^{+}\left(R_1^i, R_2^i, \ldots, R_{n_i}^i\right).$
Return confidence interval mapping
    $\widehat{C}_{m,\alpha,\delta}^{\text{jk-minmax}}(x) := \left[\min_{i \in [m]} \widehat{f}_{-i}(x) - \widehat{q}_{m,\delta}^{+}\left(\{S_{1-\alpha}^i\}_{i=1}^m\right), \max_{i \in [m]} \widehat{f}_{-i}(x) + \widehat{q}_{m,\delta}^{+}\left(\{S_{1-\alpha}^i\}_{i=1}^m\right)\right].$
```

Theorem 1. *The multi-environment confidence mapping $\widehat{C}_{m,\alpha,\delta}^{\text{jk-minmax}}$ Algorithm 1 returns provides level (α, δ)-coverage (3).*

See Section A.1 for the proof. In words, with probability at least $1 - \delta$, the prediction intervals from Algorithm 1 cover at least $(1 - \alpha) \times 100\%$ of the examples in the test environment.

Remark In Algorithm 1, it may appear that taking the minimum and maximum of the held-out predictions could yield conservative prediction intervals; intuitively, suitably corrected δ and $(1 - \delta)$ quantiles (as in the jackknife+ [4]) of $\widehat{f}_{-i}(x) \pm S_{1-\alpha}^i$ should yield a confidence set satisfying the guarantee (3). Unfortunately, this approach fails to provide (α, δ)-coverage; see Appendix B.

2.2 A multi-environment split conformal method

We also introduce a multi-environment version of split conformal inference. Our algorithm partitions the environment index set $\{1, \ldots, m\}$ into subsets D_1 and D_2 randomly. We use the data in environments indexed by D_1 to fit a model $\widehat{f}_{D_1} = \mathsf{A}(\{X^j, Y^j\}_{j \in D_1})$ with which we construct residuals for each example $j = 1, \ldots, n_i$ in each environment $i \in D_2$. Then for each environment in D_2, we construct the $(1 - \alpha)$-th quantile of its n_i residuals. This yields a set $S_{1-\alpha}^i$, $i \in D_2$, of quantiles. To obtain a likely-to-cover interval, we consider as before quantiles of these quantiles, constructing a prediction interval of the form

$$\left[\widehat{f}_{D_1}(x) - \widehat{q}_{m,\delta}^{+}\left(\left\{S_{1-\alpha}^i\right\}_{i \in D_2}\right), \widehat{f}_{D_1}(x) + \widehat{q}_{m,\delta}^{+}\left(\left\{S_{1-\alpha}^i\right\}_{i \in D_2}\right)\right].$$

```
Algorithm 2: Multi-environment Split Conformal Inference: the regression case
Input: samples $\{X_j^i, Y_j^i\}_{j=1}^{n_i}$, $i \in [m]$, confidence levels $\alpha, \delta$, split ratio $\gamma$
Randomly partition $\{1, 2, \dots, m\}$ into two sets $D_1$ and $D_2$ such that $\frac{|D_1|}{m} = \gamma$.
Set
    $\widehat{f}_{D_1} = \mathsf{A}(\{X^j, Y^j\}_{j \in D_1})$.
For $i \in D_2$, construct residual quantiles
    $R_j^i = \left| Y_j^i - \widehat{f}_{D_1}\left(X_j^i\right) \right|, \quad j = 1, \dots, n_i, \text{ and } S_{1-\alpha}^i = \widehat{q}_{n_i,\alpha}^{+}\left(R_1^i, R_2^i, \dots, R_{n_i}^i\right).$
Return confidence interval mapping
    $\widehat{C}_{m,\alpha,\delta}^{\text{split}}(x) := \left[\widehat{f}_{D_1}(x) - \widehat{q}_{m,\delta}^{+}\left(\left\{S_{1-\alpha}^i\right\}_{i \in D_2}\right), \widehat{f}_{D_1}(x) + \widehat{q}_{m,\delta}^{+}\left(\left\{S_{1-\alpha}^i\right\}_{i \in D_2}\right)\right].$
```

Theorem 2. *The multi-environment confidence mapping $\widehat{C}_{m,\alpha,\delta}^{\text{split}}$ Algorithm 2 returns provides level (α, δ)-coverage (3). If additionally the observations Y_j^i are almost surely distinct,*

$$\mathbb{P}\left[\sum_{j=1}^{n_{m+1}} 1\left\{Y_j^{m+1} \in \widehat{C}_{m,\alpha,\delta}^{\text{split}}\left(X_j^{m+1}\right)\right\} \geq \lceil (1-\alpha)\left(n_{m+1}+1\right) \rceil\right] \leq 1 - \delta + \frac{1}{m(1-\gamma)+1}.$$

See Section A.2 for the proof.

We discuss the relative merits of our algorithms. They both provide valid coverage as proved in Theorems 1 and 2. Multi-environment split conformal fits the model once, and is therefore computationally attractive. Moreover, we expect this method to be less conservative since it does not take a maximum or minimum over predictions coming from multiple models. In contrast, multi-environment split conformal uses fewer samples both to fit the initial model and to construct the residual quantiles. Thus, we expect it to suffer when the number of training environments is relatively small. We demonstrate these points further via real data experiments in Section 6.

2.3 Methods to achieve marginal coverage in hierarchical predictive inference

As we discuss earlier, Dunn et al. [9] and Lee et al. [16] provide predictive inference methods for the hierarchical (multi-environment) setting (1), focusing on guarantees that provide coverage for a single new observation (2), as in Definition 1.1. Focusing on the more recent paper [16], we review (with a correction to avoid an accidental infinite quantile [16, App. A.2.1]) their hierarchical jackknife+ and hierarchical conformal prediction algorithms. In the procedures, we let δ_z denote a point mass at z, and for a distribution P on $\mathbb{R}$, define the (usual) left quantile and (non-standard) right quantile mappings $Q_\alpha(P) := \inf\{t \mid P(Z \leq t) \geq \alpha\}$ and $Q'_\alpha(P) := \sup\{t \mid P(Z \leq t) < \alpha\}$.

```
Algorithm 3: Hierarchical Jackknife+ (Lee et al. [16])
Input: samples $\{X_j^i, Y_j^i\}_{j=1}^{n_i}$, $i \in [m]$, confidence level $\alpha$
For $i = 1, \ldots, m$, set
    $$\widehat{f}_{-i} = \mathsf{A}\left((X^1, Y^1), \ldots, (X^{i-1}, Y^{i-1}), (X^{i+1}, Y^{i+1}), \ldots, (X^m, Y^m)\right),$$
and construct residual quantiles
    $$R_j^i = |Y_j^i - \widehat{f}_{-i}(X_j^i)|, \quad j = 1, \ldots, n_i.$$
Return confidence interval mapping
    $$\widehat{C}_{m,\alpha}^{\text{hjk+}}(x) \coloneqq [\text{low}(x), \text{high}(x)],$$
where
    $$\text{low}(x) \coloneqq Q'_\alpha\left(\sum_{i=1}^{m}\sum_{j=1}^{n_k} \frac{1}{(m+1)n_i} \cdot \delta_{\widehat{f}_{-i}(x) - R_j^i} + \frac{1}{m+1} \cdot \delta_{-\infty}\right),$$
    $$\text{high}(x) \coloneqq Q_{1-\alpha}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_k} \frac{1}{(m+1)n_i} \cdot \delta_{\widehat{f}_{-i}(x) + R_j^i} + \frac{1}{m+1} \cdot \delta_{-\infty}\right).$$
```

```
Algorithm 4: Hierarchical Conformal Prediction
Input: samples $\{X_j^i, Y_j^i\}_{j=1}^{n_i}$, $i \in [m]$, confidence level $\alpha$, split ratio $\gamma$
Randomly partition $\{1, 2, \ldots, m\}$ into two sets $D_1$ and $D_2$ such that $\frac{|D_1|}{m} = \gamma$.
Set
    $$\widehat{f}_{D_1} = \mathsf{A}(\{X^j, Y^j\}_{j \in D_1}).$$
For $i \in D_2$, construct residual quantiles
    $$R_j^i = \left|Y_j^i - \widehat{f}_{D_1}\left(X_j^i\right)\right|, \quad j = 1, \ldots, n_i$$
Set
    $$T = Q_{1-\alpha}\left(\sum_{i=m\gamma+1}^{m}\sum_{j=1}^{n_k} \frac{1}{(|D_1|+1)\,n_i} \cdot \delta_{R_j^i} + \frac{1}{|D_1|+1} \cdot \delta_{+\infty}\right),$$
Return confidence interval mapping
    $$\widehat{C}_{m,\alpha}^{\text{hcp}}(x) := \left[\widehat{f}_{D_1}(x) - T, \widehat{f}_{D_1}(x) + T\right].$$
```

These algorithms construct prediction intervals for a single observation (X_1^{m+1}, Y_1^{m+1}) in the test environment $m+1$, providing guarantees of marginal coverage (2) as in Definition 1.1. In particular, Lee et al. [16] prove the following results.

Corollary 2.1 (Lee et al. [16], Theorems 1 and 5)**.** *The jackknife+ mapping $\widehat{C}_{m,\alpha}^{\text{hjk+}}$ that Algorithm 3 returns provides $1-2\alpha$ hierarchical coverage, and the conformal mapping $\widehat{C}_{m,\alpha}^{\text{hcp}}$ that Algorithm 4 returns provides $1-\alpha$ hierarchical coverage.*

These results are not completely comparable to (α, δ)-coverage guarantees (Definition 1.2). We do so somewhat heuristically in our experiments in Section 6.4, where for values of

$\alpha \in (0,1)$—the outer coverage guarantee, over environments—we may vary δ to compare performance of the methods. Previewing our results, it appears that both the hierarchical jackknife+ and split-conformal methods generate prediction sets with comparable size and coverage properties to the multi-environment methods in Algorithms 1 and 2.

3 General confidence sets and extensions

To this point, we have described our algorithms for real-valued predictions, where confidence intervals $C(x) = [a, b]$ are most practicable. Here, we generalize the algorithms beyond regression, where the target space may not be the real line and the prediction sets may be asymmetric. We first present the general formulation and abstract algorithms. Subsequently, we specialize our construction to demonstrate implementation in a few cases of interest: (i) when we represent general target spaces $\mathcal{Y}$ and confidence sets by labels y that suffer small loss under a prediction $f(x)$, i.e., $\{y \in \mathcal{Y} \mid \ell(y, f(x)) \le \tau\}$, and (ii) for quantile regression-type approaches, which allow asymmetric confidence sets in regression problems [21].

3.1 General nested confidence sets

We begin with our most general formulation. Here, we treat confidence sets themselves as the objects of interest (adopting the interpretation [11]), rather than any particular prediction method $\widehat{f}$, and assume that confidence sets are indexed by a threshold τ and nested in that

$$C_\tau(x) \subset C_{\tau+\delta}(x) \quad \text{for all } \delta \ge 0.$$

We assume now that the algorithm A returns a collection of confidence set mappings $\{\widehat{C}_\tau\}_{\tau\in\mathbb{R}}$, where each $\widehat{C}_\tau : \mathcal{X} \rightrightarrows \mathcal{Y}$ is a set-valued function. To see how this generalizes the initial Algorithm 1, note that we may write

$$\widehat{C}_\tau(x) = [f(x) - \tau, f(x) + \tau] \quad \text{or} \quad \widehat{C}_\tau(x) = [f_{\text{low}}(x) - \tau, f_{\text{high}}(x) + \tau].$$

Assuming A can perform this calculation, the immediate extension of Algorithm 1 follows.

```
Algorithm 5: Multi-environment Jackknife-minmax via nested confidence sets
Input: samples $\{X^i_j, Y^i_j\}_{j=1}^{n_i}$, $i \in [m]$, levels $\alpha, \delta$, predictive set algorithm $\mathsf{A}$
For $i = 1, \ldots, m$, set
    $\{\widehat{C}^{-i}_\tau\}_{\tau\in\mathbb{R}} = \mathsf{A}\left((X^1, Y^1), \ldots, (X^{i-1}, Y^{i-1}), (X^{i+1}, Y^{i+1}), \ldots, (X^m, Y^m)\right),$
    and construct residual quantiles
    $R^i_j = \inf\left\{\tau \mid Y^i_j \in \widehat{C}^{-i}_\tau(X^i_j)\right\}, \quad j = 1, \ldots, n_i, \quad \text{and} \quad S^i_{1-\alpha} = \widehat{q}^+_\alpha\left(R^i_1, R^i_2, \ldots, R^i_{n_i}\right).$
Return confidence interval mapping with $\widehat{\tau} = \widehat{q}^+_\delta(\{S^i_{1-\alpha}\}_{i=1}^m)$,
    $\widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}(x) := \bigcup_{i\in[m]} \widehat{C}^{-i}_{\widehat{\tau}}(x).$
```

Theorem 3. *The multi-environment confidence mapping $\widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}$ Algorithm 5 returns provides level (α, δ)-coverage (3).*

The proof of Theorem 3 mimics that of Theorem 1, and we present it in Section A.3.

Similarly, the extension of Algorithm 2 follows.

```
Algorithm 6: Multi-environment Split Conformal via nested confidence sets

Input: samples $\{X^i_j, Y^i_j\}_{j=1}^{n_i}$, $i \in [m]$, confidence levels $\alpha, \delta$, split ratio $\gamma$, predictive set algorithm $\mathsf{A}$

Randomly partition $\{1, 2, \dots, m\}$ into two sets $D_1$ and $D_2$ such that $\frac{|D_1|}{m} = \gamma$.

Set $\{\widehat{C}^{D_1}_\tau\}_{\tau \in \mathbb{R}} = \mathsf{A}(\{(X^i, Y^i)\}_{i \in D_1})$.

For $i \in D_2$, construct residual quantiles
    $R^i_j = \inf\left\{\tau \mid Y^i_j \in \widehat{C}^{D_1}_\tau\left(X^i_j\right)\right\}, \quad j = 1, \dots, n_i, \text{ and } S^i_{1-\alpha} = \widehat{q}^+_{n_i,\alpha}\left(R^i_1, R^i_2, \dots, R^i_{n_i}\right).$

Return confidence interval mapping with $\widehat{\tau} = \widehat{q}^+_\delta\left(\left\{S^i_{1-\alpha}\right\}_{i \in D_2}\right)$,
    $\widehat{C}^{\mathrm{split}}_{m,\alpha,\delta}(x) := \widehat{C}^{D_1}_{\widehat{\tau}}(x).$
```

Theorem 4. *The multi-environment confidence mapping $\widehat{C}^{\mathrm{split}}_{m,\alpha,\delta}$ Algorithm 6 returns provides level (α, δ)-coverage (3). If additionally the scores $S^i_{1-\alpha}$ are almost surely distinct, then*

$$\mathbb{P}\left[\sum_{j=1}^{n_{m+1}} 1\left\{Y^{m+1}_j \in \widehat{C}^{\mathrm{split}}_{m,\alpha,\delta}\left(X^{m+1}_j\right)\right\} \geq \lceil (1-\alpha)\left(n_{m+1}+1\right)\rceil\right] \leq 1 - \delta + \frac{1}{m(1-\gamma)+1}.$$

We omit the proof of Theorem 4, as it *mutatis mutandis* mimics that of Theorem 2.

3.2 Specializations and examples of the nested confidence set approach

We specialize the general nested prediction set Algorithms 5 and 6 to a few special cases where implementation is direct and natural.

3.2.1 General loss functions and targets

In extension to the preceding section, we consider the following scenario: we have targets $y \in \mathcal{Y}$, covariates $x \in \mathcal{X}$, and prediction functions $f \in \mathcal{F} \subset \mathcal{X} \to \mathbb{R}^k$. Then for a loss $\ell : \mathcal{Y} \times \mathbb{R}^k \to \mathbb{R}_+$, we consider predictive sets of the form

$$C_\tau(x) = \{y \in \mathcal{Y} \mid \ell(y, f(x)) \leq \tau\}$$

where, for now, $f \in \mathcal{F}$ and $\tau \in \mathbb{R}$ are left implicit. These are nested, allowing application of Theorem 3 and Algorithm 5. A slight specialization allows easier presentation: define the residual losses on environment i by

$$R^{(i)}_j := \ell\left(Y^{(i)}_j, \widehat{f}_{-i}(X^{(i)}_j)\right),$$

where as previously $\widehat{f}_{-i}$ is the leave-one-out predictor $\widehat{f}_{-i} = \mathsf{A}((X^k, Y^k)_{k \neq i})$. Setting $S^{(i)}_{1-\alpha} = \widehat{q}^{+}_{\alpha}(\{R^{(i)}_j\}_j)$, the nested union in Algorithm 5 is exactly

$$\widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}(x) := \left\{ y \in \mathcal{Y} \mid \min_{k \le m} \ell(y, \widehat{f}_{-k}(x)) \le \widehat{q}^{+}_{\delta} \left(\{S^{(i)}_{1-\alpha}\}_{i=1}^{m} \right) \right\}.$$

Corollary 3.1. *The loss-based set* $\widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}$ *provides* (α, δ)*-coverage* (3)*.*

3.2.2 Quantile regression

Romano et al. [21] highlight how moving beyond symmetric confidence sets to use quantile-based regressoin functions allows more accurate and tighter confidence bands even for $\mathbb{R}$-valued responses Y. Algorithm 5 and Theorem 3 let us adapt their technique to obtain quantile-type confidence sets in multi-environment settings. Imagine we have two algorithms fitting lower and upper predictors

$$\widehat{l}_{-i} = \mathsf{A}_{\text{low}} \left((X^k, Y^k)_{k \neq i} \right), \quad \widehat{u}_{-i} = \mathsf{A}_{\text{high}} \left((X^k, Y^k)_{k \neq i} \right),$$

where we leverage the idea that the methods target that Y lies in $[\widehat{l}_{-i}(x), \widehat{u}_{-i}(x)]$ with a prescribed probability $1 - \alpha$. To construct nested confidence sets from $\widehat{l}, \widehat{u}$, we set

$$\widehat{C}^{-i}_{\tau}(x) = \left[\widehat{l}_{-i}(x) - \tau, \widehat{u}_{-i}(x) + \tau \right].$$

Specializing the generic construction in Algorithm 5 to this case, set the residuals

$$R^i_j := \max \left\{ \widehat{l}_{-i}(X^i_j) - Y^i_j, Y^i_j - \widehat{u}_{-i}(X^i_j) \right\},$$

which by inspection satisfies

$$R^i_j = \inf \left\{ \tau \in \mathbb{R} \mid \widehat{l}_{-i}(X^i_j) - \tau \le Y^i_j \le \widehat{u}_{-i}(X^i_j) + \tau \right\}.$$

We then construct $S^i_{1-\alpha} = \widehat{q}^{+}_{\alpha}(\{R^i_j\}_{j=1}^{n_i})$ (as in Algorithm 5), and setting $\widehat{\tau} = \widehat{q}^{+}_{\delta} \left(\{S^k_{1-\alpha}\}_{k=1}^{m} \right)$, the multi-environment jackknife-minmax confidence set takes the form

$$\widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}(x) := \bigcup_{i=1}^{m} \left[\widehat{l}_{-i}(x) - \widehat{\tau}, \widehat{u}_{-i}(x) + \widehat{\tau} \right].$$

This set provides valid coverage by Theorem 3:

Corollary 3.2. *The lower/upper set* $\widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}$ *provides* (α, δ)*-coverage* (3)*.*

4 Resizing residuals to reduce interval lengths

Our experimental results in Section 6 make clear that naive application of multi-environment algorithms often produces wide prediction intervals. The bottom plots in Figures 6.1 and 6.3 show that the average size of the confidence sets are particularly large when the input parameter δ is small. This conservativeness is a natural consequence of the constructions of the multi-environment confidence sets, just as Romano et al. [21] note that naive symmetric and uniform prediction intervals (i.e., those of the form $\widehat{C}(x) = [f(x) - \tau, f(x) + \tau]$) can be overly conservative as they must typically cover even values X for which Y is highly non-symmetric or has high variance. In our context, this presents when an environment i has residual quantile $S^i_{1-\alpha}$ much larger than the rest—it is an outlier environment. For small δ, such outlier environments mostly determine the $1-\delta$ quantile of the training environments' scores $S^i_{1-\alpha}$, and consequently, these outlier training environments govern the size of the confidence set for test samples regardless of whether the test environment is outlying.

To mitigate this issue, we scale the residual quantiles by a resizing factor so that the adjusted quantiles (i) remain exchangeable, and (ii) are supported on a similar range for *all* environments. We compute the $1-\delta$ quantile of these adjusted residual quantiles. Finally, we multiply this $1-\delta$ quantile by the test environment's resizing factor to construct valid confidence sets. This way, the constructed confidence set length remains small if the test environment is not an outlier. Since the probability of any environment being an outlier is small, the size of the constructed confidence set tends to be small on average.

The question therefore turns to finding an accurate strategy for constructing these resizing factors. One natural approach is—if they are available—to incorporate environmental covariates. If environmental covariates $e \in \mathcal{E}$ are available for each environment, then we can estimate resizing factors using them, as an expanded covariate (X, e) remains i.i.d. conditional on the environment. (One could also incorporate these into the predictor $f : \mathcal{X} \times \mathcal{E} \to \mathcal{Y}$ via an expanded covariate (X, e).) Alternatively, we show that given access to a limited amount of labeled data from the test environment, one can indeed construct suitable resizing factors. We describe our strategy in the context of the multi-environment split conformal algorithm.

Suppose we observe L^{m+1} labeled random examples from the test environment. As before, we partition the training environments into two sets D_1 and D_2. Using data from environments in D_1, we construct a nested confidence set $\{\widehat{C}_\tau\}_{\tau\in\mathbb{R}}$. We randomly select L^{m+1} samples from each training environment and compute residuals for these samples using the $\{\widehat{C}_\tau\}_{\tau\in\mathbb{R}}$ confidence sets. We then pick some quantile α_0 close to 0, and compute the $1-\alpha_0$ quantile of these residuals as the resizing factor for each environment. As a result, the resizing factors for outlier environments tend to be large as desired. Empirically, we find that $\alpha_0 = 0.05$ works well for a reasonable range of input α. One can adapt this approach to any of multi-environment jackknife-minmax, hierarchical conformal prediction, or hierarchical jackknife+; Algorithm 7 shows how to do so for multi-environment split conformal prediction.

Algorithm 7: Resized Multi-environment Split Conformal

Input: samples $\{X_j^i, Y_j^i\}_{j=1}^{n_i}$, $i \in [m]$, labeled samples $\{X_j^{m+1}, Y_j^{m+1}\}_{j \in L^{m+1}}$, confidence levels α, δ, split ratio γ, predictive set algorithm A, resizing quantile α_0

Randomly partition $\{1, 2, \dots, m\}$ into two sets D_1 and D_2 such that $\frac{|D_1|}{m} = \gamma$.
Set $\{\widehat{C}_\tau^{D_1}\}_{\tau \in \mathbb{R}} = \mathsf{A}(\{(X^i, Y^i)\}_{i \in D_1})$.
For $i \in D_2$,

1. Compute residuals

$$R_j^i = \inf \left\{ \tau \mid Y_j^i \in \widehat{C}_\tau^{D_1}\left(X_j^i\right) \right\}, \quad j = 1, \dots, n_i.$$

2. Randomly select $|L^{m+1}|$ samples in environment i, and denote the set of selected samples as L^i. Compute the resizing factor

$$s^i := \widehat{q}^{\,+}_{|L^i|, \alpha_0}\left(\{R_j^i\}_{j \in L^i}\right),$$

and resized residual quantiles

$$S_{1-\alpha}^i = \widehat{q}^{\,+}_{n_i, \alpha}\left(\left\{R_j^i / s^i\right\}_{j \in [n_i] \setminus L^i}\right).$$

Compute the resizing factor for the test environment

$$s^{m+1} := \widehat{q}^{\,+}_{|L^{m+1}|, \alpha_0}\left(\left\{R_j^i\right\}_{j \in L^{m+1}}\right).$$

Return confidence interval mapping with $\widehat{\tau} = s^{m+1} \cdot \widehat{q}^{\,+}_\delta(\{S_{1-\alpha}^i\}_{i \in D_2})$,

$$\widehat{C}_{m,\alpha,\delta}^{\text{resized}}(x) := \widehat{C}_{\widehat{\tau}}^{D_1}(x).$$

As long as the resizing factors $s^i, i \in D_2 \cup \{m+1\}$ are exchangeable and independent of samples $(X^i, Y_i), i \in Y_j^i\}_{j=1}^{n_i}, i \in D_1$, the proof of Theorem 2 immediately extends to show the following result:

Theorem 5. *The multi-environment confidence mapping $\widehat{C}_{m,\alpha,\delta}^{\text{resized}}(x)$ that Algorithm 7 returns provides level (α, δ)-coverage (3), that is, the event*

$$E_m := \left\{ \sum_{j \in [n_{m+1}]} \backslash L^{m+1} 1\left\{ Y_j^{m+1} \in \widehat{C}_{m,\alpha,\delta}^{\text{resized}}(X_j^{m+1}) \right\} \geq \left\lceil (1 - \alpha)(n_{m+1} - |L^{m+1}| + 1) \right\rceil \right\}$$

satisfies $\mathbb{P}(E_m) \geq 1 - \delta$. If additionally the quantiles $S_{1-\alpha}^i$ are distinct with probability 1, then $\mathbb{P}(E_m) \leq 1 - \delta + \frac{1}{m(1-\gamma)+1}$.

5 General Consistency Results

In this section, we develop a theory of consistency for the nested confidence sets in Section 3. We define consistency via convergence to a particular limiting confidence set mapping. When this confidence set is in some sense optimal, then our methods are consistent. To that end, we shall assume there is a fixed collection $\{C_\tau\}$ of nested confidence sets, and we let E be a random environment drawn from the collection of possible environments $\mathcal{E}$, letting $\mathbb{P}(\cdot \mid E)$ denote the induced distribution conditional on E. For a fixed $\alpha, \delta \in (0,1)$, we define

$$\tau^\star(\delta, \alpha) := \inf \left\{ \tau \text{ s.t. } \mathbb{P}\left(\mathbb{P}\left(Y \in C_\tau(X) \mid E\right) \ge 1-\alpha\right) \ge 1-\delta \right\}.$$

Thus, for the given confidence mappings $C = \{C_\tau\}$, the value $\tau^\star(\delta, \alpha)$ yields the smallest confidence set providing (δ, α) coverage (Definition 1.2) at a population level. Then for Algorithms 5 and 6 to be consistent, we must have the resulting confidence sets $\widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}$ and $\widehat{C}^{\text{split}}_{m,\alpha,\delta}$ approach $C_{\tau^\star}$ in an appropriate sense.

5.1 Conditions and preliminary definitions

We require a few more definitions, providing examples as we state them. For a given nested confidence family $C = \{C_\tau\}$, we define the coverage threshold

$$\tau(x, y, C) := \inf \left\{ \tau \in \mathbb{R} \mid y \in C_\tau(x) \right\} \tag{4}$$

to be the smallest value τ such that $C_\tau(x)$ covers y.

Example 1 (Coverage thresholds)**:** The coverage threshold (4) has straightforwardly computable values for most of the "standard" cases of confidence sets we consider. For the interval-based confidence set $C_\tau(x) = [f(x) - \tau, f(x) + \tau]$, evidently $\tau(x, y, C) = |f(x) - y|$. For the lower/upper sets we use in quantile regression (recall Section 3.2.2), where $C_\tau(x) = [l(x) - \tau, u(x) + \tau]$, we have $\tau(x, y, C) = \max\{l(x) - y, y - u(x)\}$, so that $\tau(X_i^j, Y_i^j, \widehat{C}^{-i}) = R_j^i$ is simply the residual. For the generic loss-based set $C_\tau(x) = \{y \mid \ell(y, f(x)) \le \tau\}$ (recall Section 3.2.1), it is immediate that $\tau(x, y, C) = \ell(y, f(x))$, and so once again we have equality with the residuals $R_j^i = \tau(X_i^j, Y_i^j, \widehat{C}^{-i})$. $\diamond$

To discuss convergence of the estimators in general, we must address various modes of quantile convergence, which we do at a generic level via convergence in distribution. To that end, for an $\mathbb{R}$-valued random variable Z, let

$$\mathcal{L}(Z \mid P)$$

denote the induced probability law of Z under the probability distribution P. For example, given observed values of $(X_j^i, Y_j^i) \overset{\text{iid}}{\sim} P^i$, $j = 1, \ldots, n_i$, we denote the corresponding empirical distribution as $\widehat{P}^i$. Then $\mathcal{L}(\tau(X, Y, \widehat{C}^{-i}) \mid \widehat{P}^i)$ denotes the empirical distribution of the values $\tau(X_j^i, Y_j^i, \widehat{C}^{-i})$. We recall the *bounded Lipschitz metric* between distributions P and Q,

$$\|P - Q\|_{\text{BL}} := \sup_{\|f\|_\infty \le 1, \|f\|_{\text{Lip}} \le 1} \int f(dP - dQ),$$

noting that for any distribution P on $\mathbb{R}^d$, if $\widehat{P}_n$ denotes the empirical distribution of $Z_i \overset{\text{iid}}{\sim} P$, $i = 1, \ldots, n$, then $\|\widehat{P}_n - P\|_{\text{BL}} \overset{a.s.}{\to} 0$ (see, e.g., van der Vaart and Wellner [25, Chs. 1.10–1.12]).

An essentially standard lemma relates convergence in the bounded Lipschitz metric to quantiles; for completeness we include a proof in Appendix A.4.

Lemma 5.1. *Let Q be a distribution on $\mathbb{R}$ with α-quantile $\mathsf{Q}_\alpha(Q)$ such that if $Z \sim Q$, then*

$$Q(Z \le \mathsf{Q}_\alpha(Q) - u) < \alpha \quad \text{and} \quad Q(Z \le \mathsf{Q}_\alpha(Q) + u) > \alpha$$

for all $u > 0$. Then the quantile mapping Q_α is continuous at Q for the bounded Lipschitz metric: for all $\epsilon > 0$, there exists $\delta > 0$ such that if $\|P - Q\|_{\text{BL}} \le \delta$, then $|\mathsf{Q}_\alpha(P) - \mathsf{Q}_\alpha(Q)| \le \epsilon$.

With these definitions, we make a few assumptions on the convergence of the estimated families of nested confidence sets. After stating the assumptions, we will revisit the major examples of confidence sets we have considered—the symmetric sets in the basic multi-environment jackknife-minmax (Section 2), the loss-based sets in Section 3.2.1, and the quantile-type sets in Section 3.2.2—and show natural sufficient conditions for the assumptions to hold.

The first two relate to consistency and continuity of the nested confidence sets.

Assumption A1.a. *Fix environment $E = i$. As $n \to \infty$,*

$$\left\| \mathcal{L}(\tau(X, Y, C) \mid P^i) - \mathcal{L}(\tau(X, Y, \widehat{C}^{-i}) \mid \widehat{P}^i) \right\|_{\mathrm{BL}} \overset{a.s.}{\to} 0,$$

where $\widehat{C}^{-i} = \{\widehat{C}^{-i}_\tau\}_{\tau \in \mathbb{R}}$ is defined in Algorithm 5.

Assumption A1.b. *Fix any validation environment $E = i$. As $n \to \infty$,*

$$\left\| \mathcal{L}(\tau(X, Y, C) \mid P^i) - \mathcal{L}(\tau(X, Y, \widehat{C}^{D_1}) \mid \widehat{P}^i) \right\|_{\mathrm{BL}} \overset{a.s.}{\to} 0,$$

where $\widehat{C}^{D_1} = \{\widehat{C}^{D_1}_\tau\}_{\tau \in \mathbb{R}}$ is defined in Algorithm 6.

Assumption A2.a. *Let λ be a measure on $\mathcal{Y}$, fix $\tau^\star \in \mathbb{R}$, and let $\epsilon > 0$. Define the (random) subsets*

$$B^i_{n,\tau} := \left\{ x \in \mathcal{X} \ s.t.\ \lambda\left(\widehat{C}^{-i}_\tau(x) \triangle C_{\tau^\star}(x) \right) \geq \epsilon \right\},$$

indexed by $n \in \mathbb{N}, \tau \in \mathbb{R}$, and $i \leq m$, of $\mathcal{X}$. Let $\tau(n)$ be such that $\lim_{n\to\infty} \tau(n) = \tau^\star$. Then for suitably slowly growing $m = m(n) \to \infty$, the X-measure of $B_{n,\tau} := \cup_{i=1}^m B^i_{n,\tau}$ satisfies $\lim_{n\to\infty} P_X(B_{n,\tau(n)}) = 0$ with probability 1.

Assumption A2.b. *Let λ be a measure on $\mathcal{Y}$, fix $\tau^\star \in \mathbb{R}$, and let $\epsilon > 0$. Define the (random) subsets*

$$B^{\text{split}}_{n,\tau} := \left\{ x \in \mathcal{X} \ s.t.\ \lambda\left(\widehat{C}^{D_1}_\tau(X) \triangle C_{\tau^\star}(X) \right) \geq \epsilon \right\},$$

indexed by $n \in \mathbb{N}, \tau \in \mathbb{R}$, and $i \leq m$, of $\mathcal{X}$. Let $\tau(n)$ be such that $\lim_{n\to\infty} \tau(n) = \tau^\star$. Then the X-measure of $B^{\text{split}}_{n,\tau}$ satisfies $\lim_{n\to\infty} P_X(B^{\text{split}}_{n,\tau(n)}) = 0$ with probability 1.

These give a type of consistency of the confidence set mappings $\widehat{C}^{-i}$ and $\widehat{C}^{D_1}$ to C.

5.1.1 Examples realizing the assumptions

A few examples may make it clearer that we expect Assumptions A1.a and A2.a to hold. Similar arguments apply for Assumptions A1.b and A2.b, and thus are omitted. Throughout the examples, we make the standing assumption that if π is the (prior) distribution on environments and $P_X = \int P^e_X d\pi(e)$ is the marginal distribution over X, then

$$\sup_{e \in \mathcal{E}} D_{\chi^2}\left(P^e_X \| P_X\right) \leq \rho^2_{\chi^2} < \infty.$$

Note that in this case, for any function $g : \mathcal{X} \to \mathbb{R}$, we have

$$\left| \int g(dP^e - dP) \right| = \left| \int g\left(\frac{dP^e}{dP} - 1 \right) dP \right| \leq \sqrt{\mathbb{E}_P[g^2]} \sqrt{D_{\chi^2}\left(P^e \| P\right)} \leq \sqrt{\mathbb{E}_P[g^2]} \rho_{\chi^2},$$

and in particular, for any set $A \subset \mathcal{X}$, we have $|P^e(A) - P(A)| \le \sqrt{P(A)}\rho_{\chi^2}$.

We now give three examples: regression, quantile regression, and (multiclass) logistic regression, and for each we provide a simple sufficient condition for Assumptions A1.a and A2.a to hold. For all three examples, the sufficient condition assumes the leave-one-out predictions converge uniformly on compact sets. As the arguments are technical and do not particularly impact the main thread of the paper, we defer the formal proofs to appendices.

Example 2 (The regression case)**:** In the case that we perform regression as in Section 2, we assume the existence of a population function $f(x) = \mathbb{E}[Y \mid X = x]$, where the expectation is taken across environments, and $\widehat{f}_{-i} \to f$ for each i. We assume this convergence is nearly uniform on compact sets: for each $\epsilon > 0$, there exists a subset $\mathcal{X}_\epsilon$ such that

$$\sup_{x \in \mathcal{X}_\epsilon} |\widehat{f}_{-i}(x) - f(x)| \overset{a.s.}{\to} 0 \quad \text{and} \quad P(\mathcal{X}_\epsilon) \ge 1 - \epsilon. \tag{5}$$

Let λ be Lebesgue measure. Then the locally uniform convergence condition (5) implies Assumptions A1.a and A2.a for λ; see Appendix A.5 for the argument. $\diamond$

Uniform convergence on compact sets is not a particularly onerous condition for standard problems. For example, for a linear regression model, if the estimate for model coefficients converges almost surely to the model coefficients, then assumption (5) holds trivially.

Example 3 (Quantile regression)**:** In the case of quantile-type regression problems, we recall Section 3.2.2, and we assume the consistency conditions that

$$l(x) = \mathsf{Q}_{\alpha/2}(Y \mid X = x) \quad \text{and} \quad u(x) = \mathsf{Q}_{1-\alpha/2}(Y \mid X = x),$$

and that for each x, $Y \mid X = x$ has a positive density, so that l and u are unique. Then in analogue to condition (5), we assume that for each $\epsilon > 0$, there exists $\mathcal{X}_\epsilon$ such that

$$\sup_{x \in \mathcal{X}_\epsilon} \max \left\{ |\widehat{l}_{-i}(x) - l(x)|, |\widehat{u}_{-i}(x) - u(x)| \right\} \overset{a.s.}{\to} 0 \quad \text{and} \quad P_X(\mathcal{X}_\epsilon) \ge 1 - \epsilon. \tag{6}$$

As in Example 2, if λ is Lebesgue measure, then the convergence (6) implies Assumptions A1.a and A2.a for λ. See Appendix A.6 for a proof. $\diamond$

Example 4 (Classification and logistic regression)**:** We consider a k-class logistic regression problem, where we take the loss

$$\ell(y, v) = \log\left(\sum_{i=1}^k e^{v_i - v_y}\right) = \log\left(1 + \sum_{i \ne y} e^{v_i - v_y}\right).$$

We assume the predictors $f : \mathcal{X} \to \mathbb{R}^k$ are (Bayes) optimal in that they satisfy

$$\frac{\exp([f(x)]_y)}{\sum_{i=1}^k \exp([f(x)]_i)} = \mathbb{P}(Y = y \mid X = x).$$

As the loss ℓ is invariant to constant shifts we make the standing w.l.o.g. assumption that $f(x)$ and $\widehat{f}(x)$ are always mean zero (i.e. $\mathbf{1}^T f(x) = 0$) and assume the consistency condition that for each $\epsilon > 0$, there exists $\mathcal{X}_\epsilon \subset \mathcal{X}$ such that

$$\sup_{x \in \mathcal{X}_\epsilon} \left\| \widehat{f}_{-i}(x) - f(x) \right\| \overset{a.s.}{\to} 0 \quad \text{and} \quad P_X(\mathcal{X}_\epsilon) \ge 1 - \epsilon. \tag{7}$$

In this case, the uniqueness of quantile estimators requires a type of continuity condition that was unnecessary for the regression cases, and we make the additional continuity assumption that for each $c \in \mathbb{R}_+$ and $y \in [k]$, the set $\{x \in \mathcal{X} \mid \ell(y, f(x)) = c\}$ has measure zero. As a consequence, the sets

$$D_{c,\epsilon} := \{x \in \mathcal{X} \mid \text{there exists } y \text{ s.t. } |\ell(y, f(x)) - c| < \epsilon\}$$

satisfy $\lim_{\epsilon \to 0} P_X(D_{c,\epsilon}) = 0$ for each c, by continuity of measure.

An analogous argument to that in Examples 2 and 3 then shows that the conditions above suffice for Assumptions A1.a and A2.a to hold with counting measure. See Appendix A.7. ◇

5.2 Consistency of the multi-environment Jackknife-minmax

With the motivating examples in place, we now provide the main convergence theorem. We state one final assumption, which makes the environment-level quantiles sufficiently unique that identifiability is possible.

Assumption A3. *Define the quantile of the coverage threshold* (4) *on environment i by*

$$\mathsf{Q}_{1-\alpha}(P^i) := \inf \left\{ t \mid P^i(\tau(X, Y, C) \le t) \ge 1 - \alpha \right\}.$$

For a given δ, there exists a (unique) $q(\delta)$ such that for any $u > 0$,

$$\mathbb{P}\left(\mathsf{Q}_{1-\alpha}(P^E) \le q(\delta) - u\right) < 1 - \delta \quad \textit{and} \quad \mathbb{P}\left(\mathsf{Q}_{1-\alpha}(P^E) \le q(\delta) + u\right) > 1 - \delta,$$

where the probability is taken over E drawn randomly from $\mathcal{E}$.

Given this assumption, we can show that the X-measure of sets where the multi-environment jackknife-minmax and "true" confidence set $C_{\tau^\star(\delta,\alpha)}$ make different predictions converges to zero.

Theorem 6. *Let λ be a measure on $\mathcal{Y}$ such that Assumptions A1.a, A2.a, and A3 hold. Let $\epsilon > 0$ and $m = m(n)$ be a sufficiently slowly growing sequence. Then the P_X measure of*

$$\widehat{B}(\epsilon) := \left\{ x \in \mathcal{X} \quad \textit{such that} \quad \lambda\left(\widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}(x) \triangle C_{\tau^\star(\delta,\alpha)}(x)\right) \ge \epsilon \right\}$$

satisfies $P_X(\widehat{B}(\epsilon)) \overset{a.s.}{\to} 0$.

Proof The key step in the argument is to recognize that Assumptions A1.a and A3 give consistency of the estimated threshold $\widehat{\tau}$:

Lemma 5.2. *Let Assumption A1.a hold and A3 hold for the choice δ. Then for all suitably slowly growing sequences $m = m(n) \to \infty$, the global estimated threshold $\widehat{\tau} := \widehat{q}^{+}_{\delta}(\{S^i_{1-\alpha}\}_{i=1}^m)$ in Algorithm 5 converges almost surely: $\widehat{\tau} \overset{a.s.}{\to} q(\delta)$.*

We defer this proof to Section A.8, continuing with the main thread of the theorem.

Let $m = m(n)$ be a sequence tending to ∞ but such that the conclusions of Lemma 5.2 hold. Let $\widehat{\tau} = \widehat{q}^{+}_{\delta}(\{S^i_{1-\alpha}\}_{i=1}^m)$ be the random δ-quantile in Alg. 5. Lemma 5.2 guarantees that $\widehat{\tau} \overset{a.s.}{\to} q(\delta)$. Following a slight variation of the notation of Assumption A2.a, define the sets $B^i_n = \{x \mid \lambda(\widehat{C}^{-i}_{\widehat{\tau}}(x) \triangle C_{\tau^\star(\delta,\alpha)}(x)) \ge \epsilon\}$. Then

$$\left\{ x \mid \lambda\left(\widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}(x) \triangle C_{\tau^\star(\delta,\alpha)}(x)\right) \ge \epsilon \right\} \subset \bigcup_{i=1}^m B^i_n$$

and Assumption A2.a guarantees that the latter set has P_X-measure tending to zero. □

5.3 Consistency of multi-environment split conformal prediction

Similarly, we can show that the X-measure of sets where the split conformal and "true" confidence set $C_{\tau^\star(\delta,\alpha)}$ make different predictions converges to zero.

Theorem 7. *Let λ be a measure on $\mathcal{Y}$ such that Assumptions A1.b, A2.b, and A3 hold. Let $\epsilon > 0$ and $m = m(n)$ be a sufficiently slowly growing sequence. Then the P_X measure of*

$$\widehat{B}(\epsilon) := \left\{ x \in \mathcal{X} \quad \textit{such that} \quad \lambda\left(\widehat{C}^{\text{split}}_{m,\alpha,\delta}(x) \triangle C_{\tau^\star(\delta,\alpha)}(x)\right) \geq \epsilon \right\}$$

satisfies $P_X(\widehat{B}(\epsilon)) \overset{a.s.}{\to} 0$.

Proof As in Lemma 5.2, we let $m = m(n)$ be a sequence tending to ∞ such that the global estimated threshold $\widehat{\tau} = \widehat{q}^{+}_{\delta}(\{S^i_{1-\alpha}\}_{i\in D_2})$ in Algorithm 6 converges almost surely to $q(\delta)$. Use Assumption A2.b to complete the proof. □

6 Real Data Examples

6.1 Neurochemical Sensing

In this section, we apply our algorithms to the prediction of neurotransmitter concentration levels, with a specific focus on dopamine [14]. Estimating dopamine levels in awake, functioning humans at a relatively high frequency is a notoriously challenging task. Yet dopamine governs critical human behavior, thus understanding stimuli that maintain healthy dopamine levels is of crucial importance.

With the advent of modern technology, scientists now have access to extensive lab generated multi-environment data that can aid in improving human dopamine level predictions [19, 3]. To this end, scientists expose electrodes to various known dopamine concentrations and collect measurements of currents that pass through the electrodes at different voltage potentials (say p different potentials). From each electrode, the scientist obtains a matrix, where each row records the different current levels in the electrode, when one changes its potential over a set of p values while exposing it to a specific dopamine concentration level. For each electrode exposed to a certain concentration level, the scientist collects multiple p-dimensional measurements corresponding to different time points. By changing the concentration over several levels (say ℓ levels) and collecting multiple observations at each level (say t observations), the scientist obtains $n = \ell \times t$ observations in total, resulting in an $n \times p$ covariate matrix from each electrode. The t measurements corresponding to each level might exhibit weak correlations, but since state-of-the-art scientific work in the area [19] treats these to be independent, we stick to this convention. The outcome corresponding to each row is the dopamine level that generated that row's current values. Due to variations in electrode construction as well as the experimental setup under which measurements are obtained, the data from different electrodes follow different distributions.

To map this application to our setting, we may consider each electrode to be one of our environments. We use data from multiple electrodes to train our algorithms, hoping that such multi-environment learning would create robust prediction models that generalize better when applied in a different context, e.g. while predicting dopamine levels on the human brain. Formally, the training data comprises 15 environments corresponding to 15 electrodes. Each includes roughly 20,000 observations [19, 18]. The covariates are current measured in nanoamps (nA) collected at 1000 discrete voltage potentials. For each observation, the outcome is a measurement of dopamine concentration in nanomolars (nM). The outcomes lie in $[0, 2000]$, so we intersect our predictions with this range before producing intervals.

In each experiment, we select 5 environments at random for training and use the rest for testing. We train Algorithms 1 and 2 and examine their coverage on the test data under the (α, δ)-coverage notion in Definition 1.2. We use ridge regression for the base model $\widehat{f}$ and leave-one-out cross validation for choosing the ridge parameter, repeating the experiment 100 times and plotting the average coverage and the average set length (defined below).

For the k-th experiment ($k = 1, 2, \ldots, 100$), we let $\{e_{k,i}\}_{1 \le i \le 5}$ and $\{e_{k,i}\}_{6 \le i \le 15}$ denote the train and test environments, respectively. For $k \in [100], i \in [15]$, we use $n_{k,i}$ to denote the sample size in environment $e_{k,i}$ and $\{X_j^{k,i}, Y_j^{k,i}\}_{1 \le j \le n_{k,i}}$ the observations. We define the variable

$$A_j^{k,i} := 1\Big\{Y_j^{k,i} \in \widehat{C}\left(X_j^{k,i}\right)\Big\}$$

to indicate whether the outcome corresponding to the j-th sample in the i-th environment is covered by the constructed confidence set during the k-th experiment.

We say a test environment is covered if the fraction of covered samples in the environment is at least $1 - \alpha$. Then by Theorems 3 and 4, we expect that at least $1 - \delta$ fraction of the test environments are covered. We define "empirical $1 - \delta$" as the fraction of test environments covered across our experiments, "empirical set length" as the average length of constructed confidence sets averaged over the test environments, and "empirical $1 - \alpha$" as the average fraction of covered samples over the covered test environments:

$$\text{“Empirical } 1-\delta\text{”} := \frac{1}{1000} \sum_{k=1}^{100} \sum_{i=6}^{15} 1\Bigg\{ \sum_{j=1}^{n_{k,i}} A_j^{k,i} \ge \lceil (1-\alpha)\left(n_{k,i}+1\right) \rceil \Bigg\},$$

$$\text{“Empirical } 1-\alpha\text{”} := \frac{\sum_{k=1}^{100} \sum_{i=6}^{15} 1\Big\{\sum_{j=1}^{n_{k,i}} A_j^{k,i} \ge \lceil (1-\alpha)\left(n_{k,i}+1\right) \rceil\Big\} \left(\frac{1}{n_{k,i}} \sum_{j=1}^{n_{k,i}} A_j^{k,i}\right)}{\sum_{k=1}^{100} \sum_{i=6}^{15} 1\Big\{\sum_{j=1}^{n_{k,i}} A_j^{k,i} \ge \lceil (1-\alpha)\left(n_{k,i}+1\right) \rceil\Big\}},$$

$$\text{“Empirical Set Length”} := \frac{1}{1000} \sum_{k=1}^{100} \sum_{i=6}^{15} \frac{1}{n_{k,i}} \sum_{j=1}^{n_{k,i}} \left|\widehat{C}\left(X_j^{k,i}\right)\right|,$$

where $|\widehat{C}(X_j^{k,i})|$ denotes the length of $\widehat{C}(X_j^{k,i})$.

6.1.1 Influence of Input δ

To examine the influence of the input δ on the performance of our algorithms, we set $\alpha = 0.05$, and the split ratio to be 0.5 for Algorithm 2. We vary the values of δ and display the results in Figure 6.1. The plots show that both multi-environment split conformal and jackknife-minmax produce valid coverage. But, multi-environment split conformal tends to generate more conservative prediction intervals than jackknife-minmax. Moreover, we see that the relationship between the empirical $1 - \alpha$ and the input $1 - \delta$ is non-monotone. This occurs since an increase in the input $1 - \delta$ tends to increase the set length for both algorithms, which in turn may increase the empirical $1 - \alpha$. On the other hand, a higher $(1 - \delta)$ may be achieved by including more environments with low coverage per environment. This may lead to a decreased empirical $(1 - \alpha)$. These opposing factors lead to the non-monotonicity.

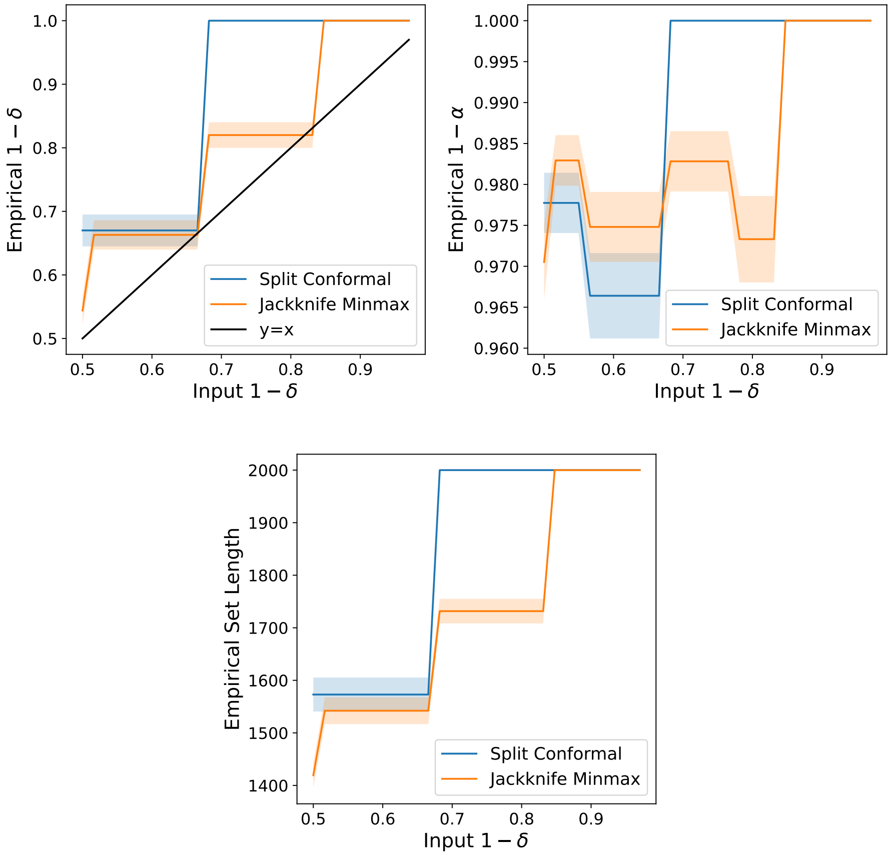

Figure 6.1. Influence of input δ on the performance of conformal algorithms applied to the neurochemical sensing data. For these experiments, α is set to be 0.05. The plots show the empirical $1-\delta$, empirical $1-\alpha$, and empirical set length for both the split conformal and jackknife-minmax algorithms with various input δ.

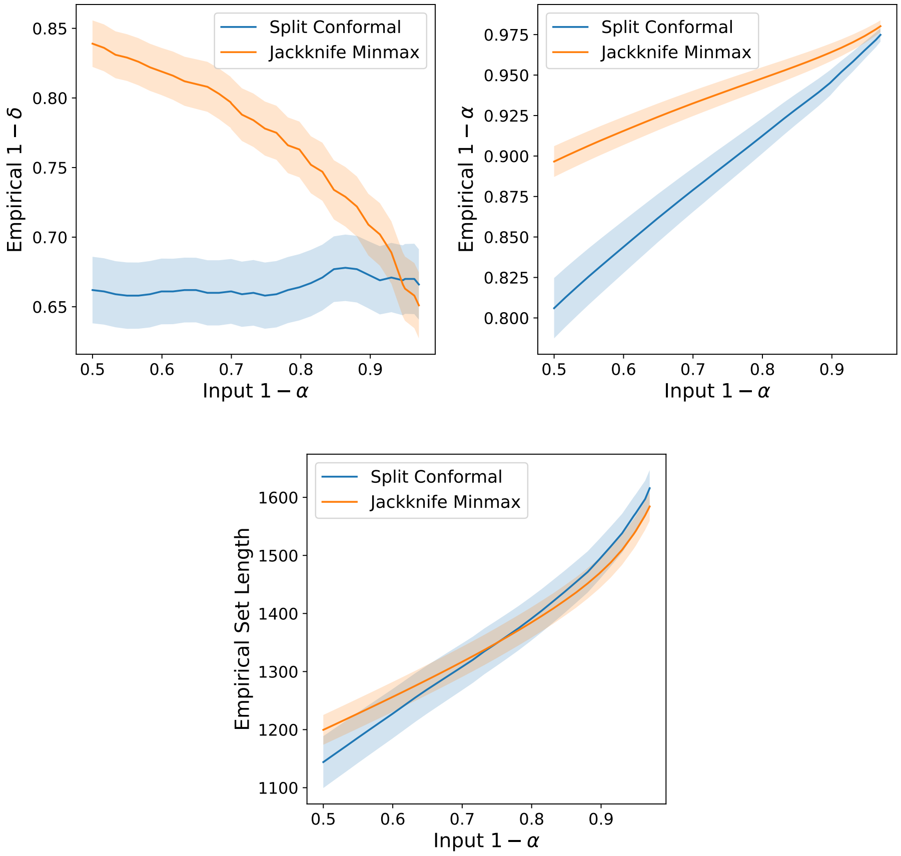

Figure 6.2. Influence of input α on the performance of conformal algorithms the neurochemical sensing data. For these experiments, δ is set to be 0.33. The plots show the empirical $1-\delta$, empirical $1-\alpha$, and empirical set length for both the split conformal and jackknife-minmax algorithms with various input α.

6.1.2 Influence of Input α

To examine the influence of the input α on the performance of our conformal algorithms, we set $\delta = 0.33$, and the split ratio to be 0.5 for Algorithm 2. We chose $\delta = 0.33$ as opposed to a smaller value since otherwise multi-environment conformal always outputs $[0, 2000]$ as the prediction interval regardless of the input α. We vary the α values and display the results in Figure 6.2. We observe that for both algorithms, the empirical $1-\alpha$ tends to increase as the input $1-\alpha$ increases. However, the relationship between the empirical $1-\delta$ and the input $1-\alpha$ is less clear. Two factors influence this relationship. As $1-\alpha$ increases, the set length of conformal intervals will increase so that each sample is more likely to be covered. Nonetheless, as the input $1-\alpha$ increases, more samples in each test environment need to be covered, and the fraction of environments satisfing the condition may decrease.

6.2 Species Classification

We next apply our algorithms in the context of species classification. To monitor wildlife biodiversity, ecologists use camera traps—heat or motion-activated cameras placed in the wild—which exhibit variation in illumination, color, camera angle, background, vegetation, and relative animal frequencies. We thus consider each camera trap an environment. Ecologists seek to use existing camera trap shots to train machine learning models that classify wildlife species accurately in new camera trap deployments [15].

The covariates of this species classification data are 2D images, and the targets are species of animals present in the images [6]. We pre-process the data by removing environments with at most 100 observations, removing labels that appear in less than 5 percent of the remaining environments. After the pre-processing, we obtain 219 environments and 57 labels. On average, each environment consists of 874 images. We then randomly choose 50 environments for training and keep the remaining 169 for testing. For the base model, we use a ResNet-50 model pretrained on ImageNet using a learning rate of $3 \cdot 10^{-5}$ and no ℓ_2-regularization [12]. Since the pretrained model takes in images of size 448×448, we rescale the inputs to the same size. We repeat the experiment 20 times, and then plot the average coverage and the average set length.

6.2.1 Influence of Input δ

We examine the performance of our algorithms as the input δ varies (Figure 6.3) similar to Section 6.1.1 earlier. We observe that the performance is now flipped, the multi-environment jackknife-minmax is now more conservative. With an increased number of training environments, the multi-environment split conformal outperforms jackknife-minmax also in terms of empirical set length.

6.2.2 Influence of Input α

We examine the performance of our algorithms as the input α varies (Figure 6.4) similar to Section 6.1.2. We observe that the empirical $1 - \alpha$ and empirical set length both increase as the input $1 - \alpha$ increases. Interestingly, the conformal sets output by jackknife-minmax are not much larger than those output by split conformal under this setting.

6.3 Resizing Residuals to Reduce Average Set Size

We observe that for both datasets, when the input δ is small, the conformal intervals have large set length on average. We next investigate the effects of our resizing technique (Section 4) in reducing the average set length.

6.3.1 Resized Multi-environment Split Conformal

We apply the resized multi-environment split conformal (Algorithm 7) to both the neurochemical sensing and the species classification datasets. We demonstrate that the resized split conformal algorithm is able to reduce the average set length of conformal intervals. In the following two examples, we set the resizing quantile α_0 to be 0.05.

Each environment in the neurochemical sensing data consists of around 20,000 samples. We use 30 randomly selected labeled data in each environment to construct the resizing factors. The experimental settings are the same as in Section 6.1, except that we randomly sample

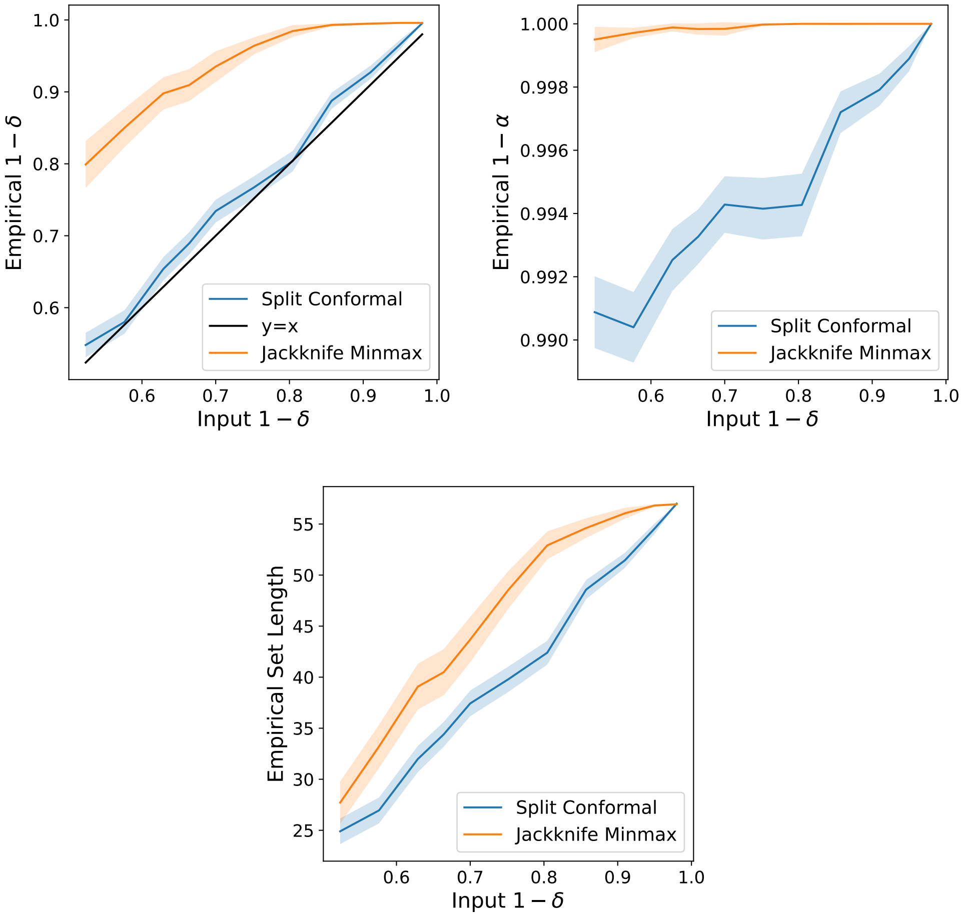

Figure 6.3. Influence of input δ on the performance of conformal algorithms applied to the species classification data. For these experiments, α is set to be 0.05. The plots show the empirical $1-\delta$, empirical $1-\alpha$, and empirical set length for both the split conformal and jackknife-minmax algorithms with various input δ.

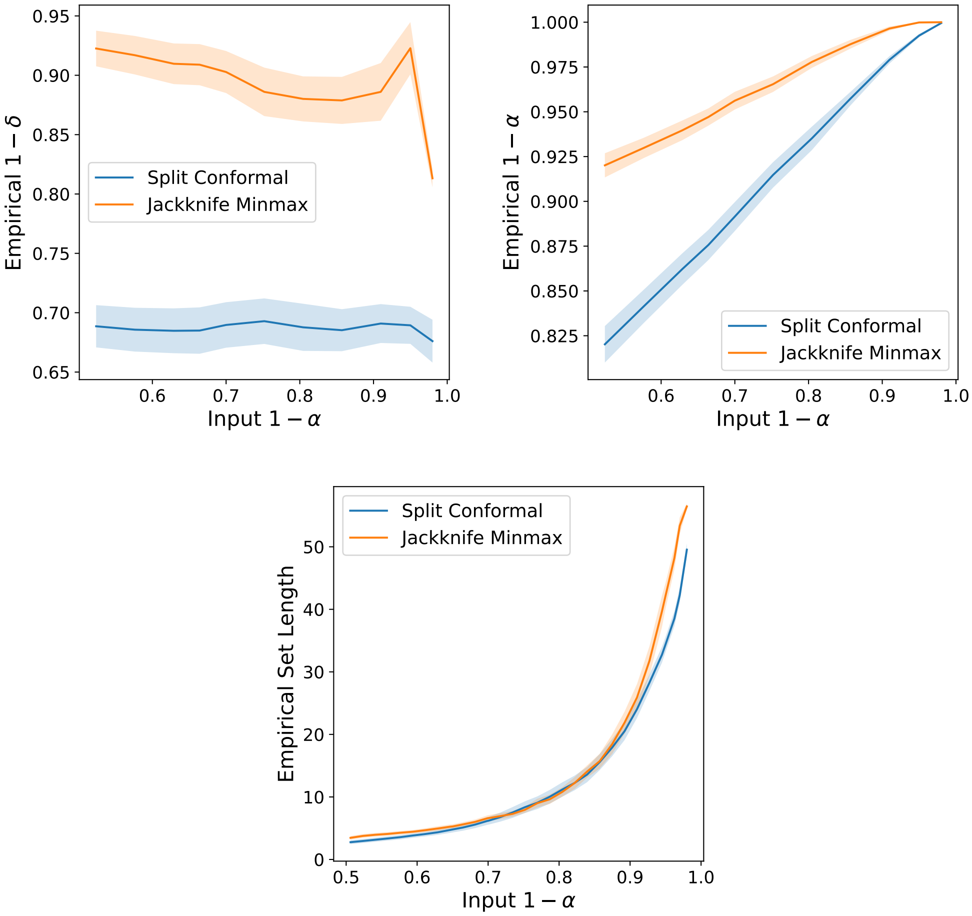

Figure 6.4. Influence of input α on the performance of conformal algorithms the species classification data. For these experiments, δ is set to be 0.33. The plots show the empirical $1-\delta$, empirical $1-\alpha$, and empirical set length for both the split conformal and jackknife-minmax algorithms with various input α.

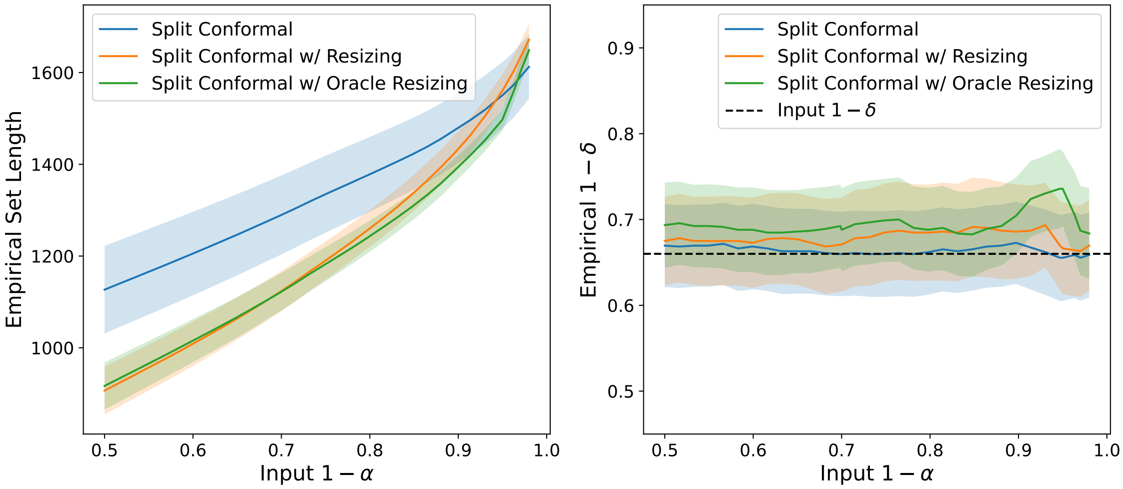

Figure 6.5. Performance of the split conformal algorithm with and without resizing applied to the neurochemical sensing data. For these experiments, δ is set to be 0.33. The left plot shows the relation between empirical set length and input $1-\alpha$, while the right plot shows the relation between empirical $1-\delta$ and input $1-\alpha$.

30 labeled data in each environment to construct the resizing factors. The coverage results presented in Figure 6.5 are with respect to the unlabeled data in each test environment.

The left plot shows the relation between the empirical set length and the input $1-\alpha$. The blue and the orange curves correspond to the results of split conformal with and without resizing, respectively. The green curve corresponds to the oracle case where we know the $1-\alpha_0$ residual quantile of all the unlabeled data in each test environment. In the oracle case, we use the $1-\alpha_0$ residual quantile as the resizing factor. We observe that when the input $1-\alpha$ is close to 1, the empirical set length is large for all three methods. This is because the $1-\alpha$ residual quantile is large in each test environment, so the resizing method cannot reduce the average set length by much. For moderate $1-\alpha$, on the other hand, the resizing methods can reduce the average set length. The right plot shows the relation between empirical $1-\delta$ and input $1-\alpha$. The results demonstrate that the resizing methods provide valid coverage, corroborating the statement of Theorem 5.

For the species classification data, we use 20 randomly selected labeled samples in each environment to construct the resizing factors. The experimental settings are the same as in 6.2, except that 1) we fix the input δ to be 0.05, and vary the input α, and 2) we randomly sample 20 labeled data in each environment to construct the resizing factors. We display the results in Figure 6.6. Again, we observe that the resizing method reduces the average set length without breaking our coverage guarantees.

6.3.2 Resized HCP

To illustrate the utility of our resizing technique beyond our algorithms, we study its effects on Lee et al. HCP algorithm [16]. Since Lee et al. design HCP to construct conformal intervals for a single sample in each test environment, it may not be practical to use additional labeled samples from the test environments to construct resizing factors. For illustration purposes, we consider the oracle setting where we know the $1-\alpha_0$ residual quantile of all samples in each test environment. We use this $1-\alpha_0$ residual quantile as the resizing factor. For both the species classification data and the neurochemical sensing data, we vary the value of the

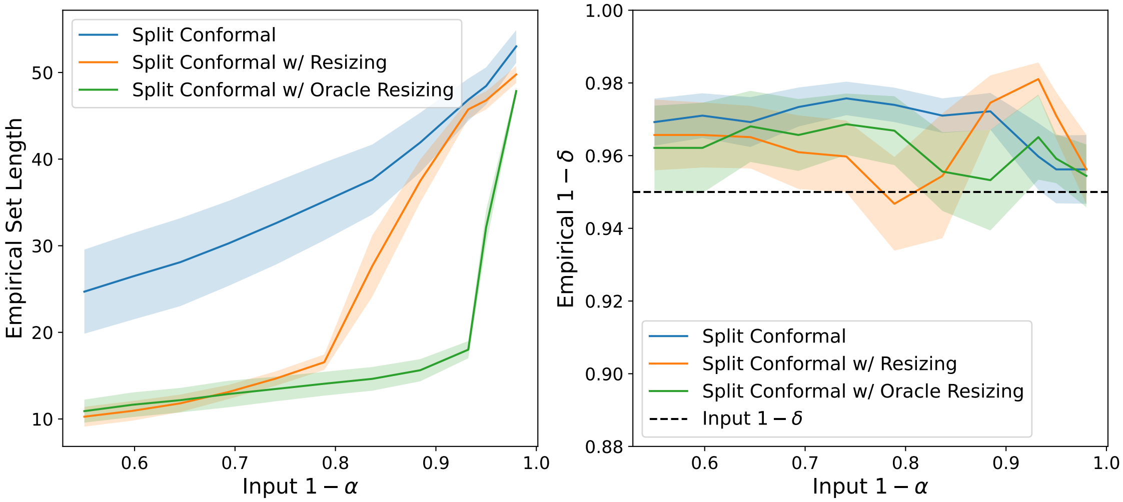

Figure 6.6. Performance of the split conformal algorithm with and without resizing applied to the species classification data. For these experiments, δ is set to be 0.05. The left plot shows the relation between empirical set length and input $1-\alpha$, while the right plot shows the relation between empirical $1-\delta$ and input $1-\alpha$.

input α and compare the performance of HCP and resized HCP algorithms. We display the results in Figure 6.7 and 6.8. We find that the oracle resizing method reduces the average size of the HCP prediction intervals without breaking the coverage guarantees.

6.4 Comparison with Hierarchical Conformal Prediction and Jackknife+

To conclude, we provide a comparison of our algorithms with HCP. Setting up this comparison is non-trivial since HCP provides intervals with a different form of coverage guarantee. Nonetheless, since they work under similar hierarchical models, we believe a comparison to be instructive. To set this up, for each fixed value of α, we find the largest δ such that the fraction of overall test samples covered by multi-environment split conformal exceeds that of HCP. We then compare the performance of multi-environment split conformal with parameters α, δ and HCP with parameter α. We apply the two methods on the neurochemical sensing and species classification data. We display the results in Figures 6.9 and 6.10, respectively.

Due to the way of selecting δ, multi-environment split conformal and jackknife-minmax produce slightly larger prediction sets than HCP and hierachical jackknife+, respectively. Moreover, we observe that for all the conformal algorithms considered, coverage (i.e. empirical $1-\delta$ and empirical $1-\alpha$) generally increase as the input $1-\alpha$ increases. However, as shown in Figure 6.10, this relation does not always hold. As the input $1-\alpha$ becomes larger, the average set size of the produced prediction sets also becomes larger, which tend to increase the empirical $1-\delta$. On the other hand, as the input $1-\alpha$ becomes larger, fewer environments will have at least $1-\alpha$ fraction of samples covered, which tend to decrease the empirical $1-\delta$. Thus the relation between empirical $1-\delta$ and input $1-\alpha$ is not necessarily monotone. Similarly, as the definition of empirical $1-\alpha$ involves the parameter δ, the relation between empirical $1-\alpha$ and input $1-\alpha$ is not necessarily monotone either.

In sum, with an appropriate choice of δ, we observe that the multi-environment split conformal and jackknife-minmax produce prediction sets with similar size and coverage properties as the HCP and hierarchical jackknife+, respectively.

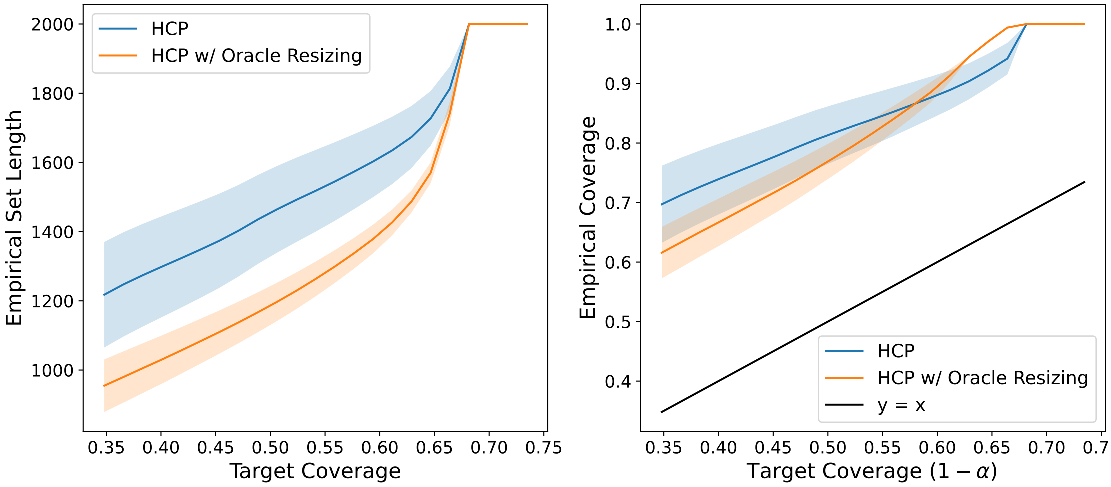

Figure 6.7. Performance of HCP and resized HCP applied to the neurochemical sensing data. Left plot shows the average set size over all test samples, and right plot shows the fraction of all test samples covered by their conformal sets.

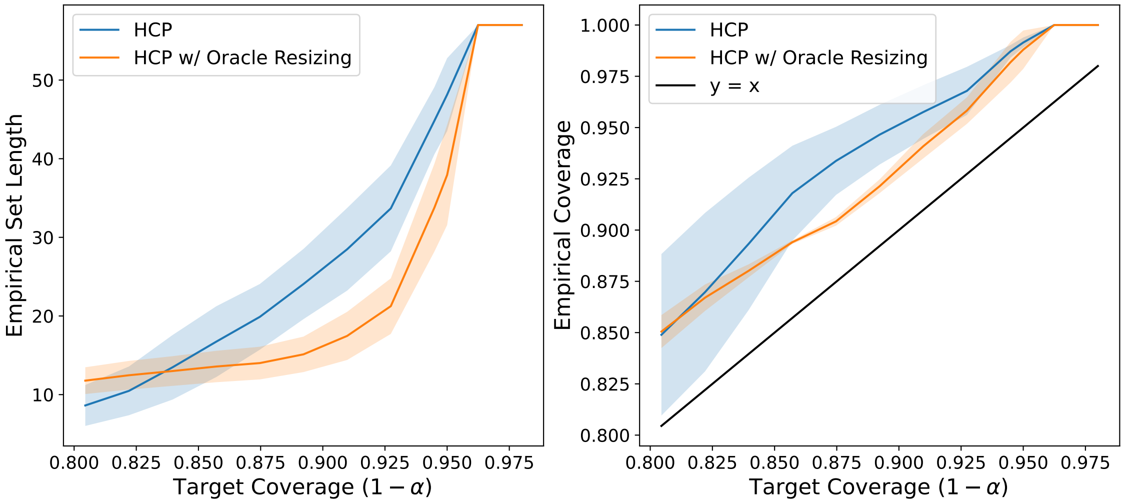

Figure 6.8. Performance of HCP and resized HCP applied to the species classification data. Left plot shows the average set size over all test samples, and right plot shows the fraction of all test samples covered by their conformal sets.

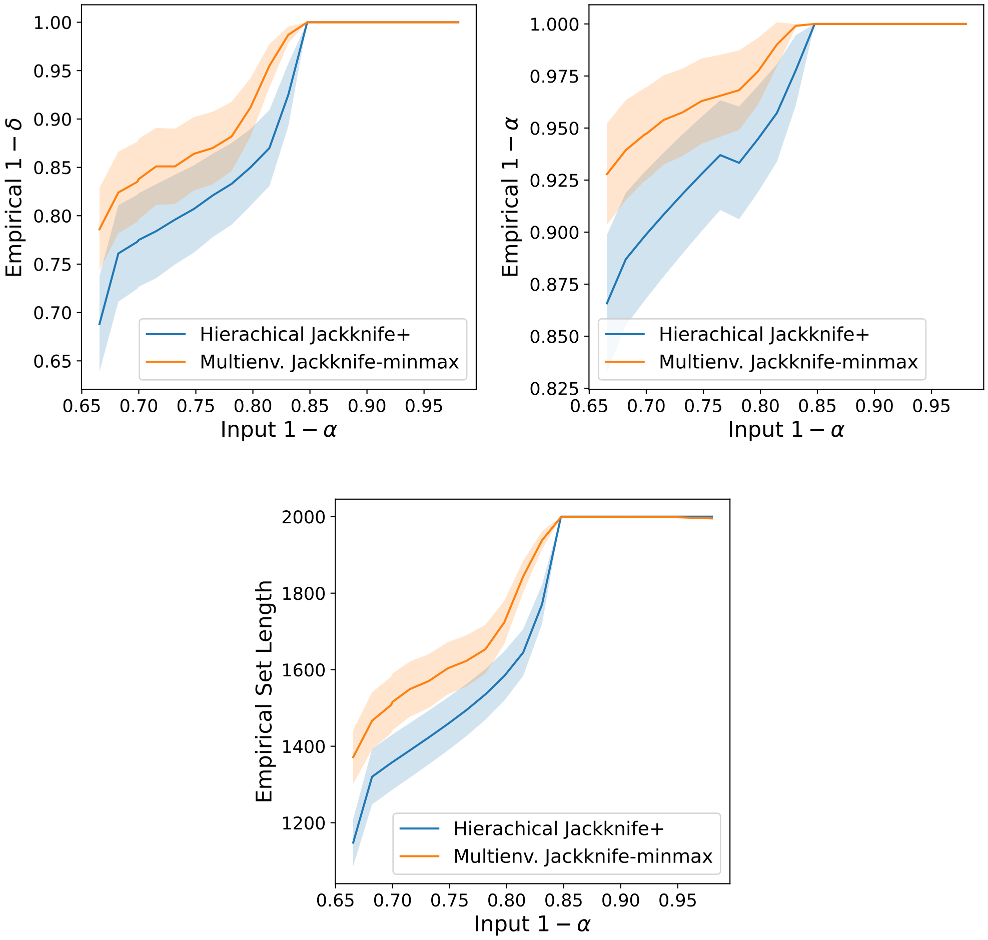

Figure 6.9. Performance of multi-environment jackknife-minmax and hierarchical jackknife+ applied to the neurochemical sensing data. Multi-environment jackknife-minmax takes in the parameters α, δ, and hierarchical jackknife+ takes in the parameter α. For each value of α, we find the largest δ such that the fraction of test samples covered by multi-environment split conformal exceeds that of hierarchical jackknife+.

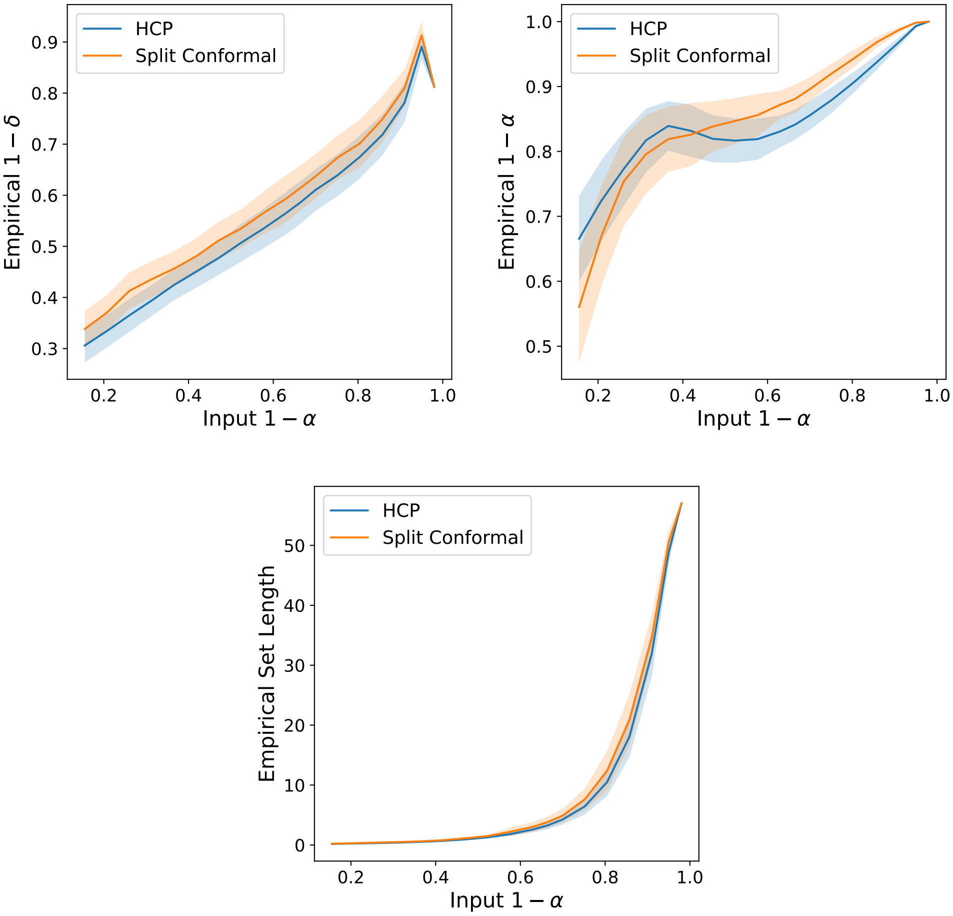

Figure 6.10. Performance of multi-environment split conformal and HCP applied to the species classification data. Multi-environment split conformal takes in the parameters α, δ, and HCP takes in the parameter α. For each value of α, we find the largest δ such that the fraction of test samples covered by multi-environment split conformal exceeds that of HCP.

7 Discussion and conclusions

The challenge of maintaining predictive validity when distributions change remains one of the core challenges in statistics and machine learning. At the most basic level, any claim of a study's external validity is a claim that statistical conclusions remain valid in environments distinct—however slightly—from the study's population [13]. A growing literature in machine learning also highlights the challenges of prediction across environments. In some cases, those environments are obvious, arising from distinct experimental conditions or measurement devices [15] or from changing populations, such as identifying pathologies from lung scans across hospitals [27]. In others, new environments arise even when constructing new evaluation datasets that replicate original data collection as exactly as possible [20, 23].

The approaches this paper outlines to predictive inference across environments and in other hierarchical data collection scenarios should therefore see wide application. One might argue that, given that applications of learning algorithms always involve some distribution shift, however mild, we should always employ some type of corrective measure to attempt to maintain validity. Section 4 highlights a key insight, which we believe is one of the main takeaways of this work: measuring variance and the scale of uncertainty across environments is essential for practicable confidence sets and predictions. Most types of predictive inference repose on some type of exchangability [26, 24, 8, 5]—excepting a recent line of work moves toward coverage guarantees that hold asymptotically irrespective of the data [10, 1]—as does this work and others on maintaining validity across populations [16, 9]. The optimality and adaptivity of predictive inference procedures, such as the techniques we develop in Section 5, also rely on some type of independence and exchangability. Future research to identify more nuanced ways in which data remains exchangeable could thus have substantial impact, allowing us to enhance the versatility and utility of conformal prediction, the jackknife, and other predictive inferential approaches.

Acknowledgments

This work began when SG was a student at Stanford. JCD and SG were partially supported by the National Science Foundation grant RI-2006777 and Office of Naval Research grant N00014-22-1-2669. PS and KJ were partially funded by the National Science Foundation grant DMS-2113426. PS was also partially funded by the Eric and Wendy Schmidt Fund for Strategic Innovation.

A Technical proofs

A.1 Proof of Theorem 1

We take as inspiration the proof of Barber et al. [4, Theorem 3]. For the sake of the proof only, to demonstrate the appropriate exchangeability, we assume we have access to both the features and responses of the test environment $\{(X_j^{m+1}, Y_j^{m+1})\}_{j=1}^{n_{m+1}}$. Then we let $\widetilde{f}_{-(i,k)}$ define the predictive function fitted on all environments (training and test) except that environments i and k are removed, i.e.,

$$\widetilde{f}_{-(i,k)} = \mathsf{A}\left(\{X^l, Y^l\}_{l\neq i, l\neq k}\right).$$

With this construction, we have $\widetilde{f}_{-(i,k)} = \widetilde{f}_{-(k,i)}$ for $i \neq k$ and

$$\widetilde{f}_{-(i,m+1)} = \widehat{f}_{-i} \quad \text{for } i \in [m],$$

the key identity that allows us to exploit block exchangeability.

We define a matrix of $(1-\alpha)$-quantile of residuals, $R \in \mathbb{R}^{(m+1)\times(m+1)}$, with entries

$$R_{ik} = \begin{cases} \infty & \text{if } i = k \\ \widehat{q}^{+}_{n_i,\alpha}\left(\left\{|Y^i_j - \tilde{f}_{-(i,k)}(X^i_j)|\right\}_{j=1}^{n_i}\right) & \text{if } i \neq k, \end{cases} \tag{8}$$

and so $R_{i,m+1} = S^i_{1-\alpha}$ in Algorithm 1.

Now [cf. 4, Proof of Theorem 3], define the comparison matrix $A \in \{0,1\}^{(m+1)\times(m+1)}$ with entries

$$A_{ik} = 1\left\{\min_{k'} R_{ik'} > R_{ki}\right\},$$

so that the smallest residual $(1-\alpha)$ quantile when omitting environment i is larger than the $(1-\alpha)$ quantile of residuals in environment k when not including i or k. Define the set $\mathcal{S}(A) \subseteq \{1,\ldots,m+1\}$ of strange environments

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

where $A_{i,\bullet} = \sum_{k=1}^{m+1} A_{ik}$ is the ith row sum of A, to be those where environment i typically has "too small" residuals.

We identify three steps that together yield the proof.

Step 1. Observe that $|\mathcal{S}(A)| \leq \delta(m+1)$.

Step 2. Using that the environments are exchangeable, the probability that the test environment $m+1$ is strange (i.e., $m+1 \in \mathcal{S}(A)$) satisfies

$$\mathbb{P}(m+1 \in \mathcal{S}(A)) \leq \delta. \tag{9}$$

Step 3. Prove the desired coverage guarantee of the theorem by showing that if coverage in environment $m+1$ fails, then it is strange, i.e., $m+1 \in \mathcal{S}(A)$.

The result in Step 1 is an immediate consequence of Barber et al. [4, Thm. 3, Step 1]. Step 2 similarly follows immediately [4, Thm. 3, Step 2]. It remains to prove step 3.

Proof of Step 3: Suppose that coverage fails, that is,

$$\sum_{j=1}^{n_{m+1}} 1\left\{Y^{m+1}_j \in \widehat{C}^{\text{jk-minmax}}_{m,\alpha,\delta}(X^{m+1}_j)\right\} < \lceil (1-\alpha)(n_{m+1}+1)\rceil. \tag{10}$$

We will show that on the event (10), the environment $m+1$ is strange, so that by Step 2 the failure has probability at most δ.

Before we complete the proof, we provide two technical lemmas that we use. These lemmas allow us to transition between coverage guarantees "columnwise," in the sense of within an environment, and "rowwise" in the sense of across environments.

Lemma A.1. *Let $B \in \mathbb{R}^{n\times m}$ and $c \in \mathbb{R}$. Then*

$$\widehat{q}^{+}_{n,\alpha}\left(1\left\{c < \min_k B_{jk}\right\}_{j=1}^{n}\right) = 1 \quad \textit{implies} \quad c < \min_k \widehat{q}^{+}_{n,\alpha}\left(\{B_{jk}\}_{j=1}^{n}\right).$$

That is, if fewer than $\lceil(1-\alpha)(n+1)\rceil$ indicators $1\{c < \min_k B_{jk}\}$ are 0, then c is less than each ($k = 1,2,\ldots,m$) of the $(1-\alpha)$ quantiles $\widehat{q}^{+}_{n,\alpha}\{B_{jk}\}_{j=1}^{n}$.

Proof We have $\widehat{q}^{+}_{n,\alpha}(1\{c < \min_k B_{jk}\}_{j=1}^{n}) = 1$ if and only if at least $n - \lceil(1-\alpha)(n+1)\rceil$ values in $1\{c < \min_k B_{jk}\}$ are 1. In this case, there are at least $n - \lceil(1-\alpha)(n+1)\rceil$ rows $J_{\text{dom}} \subset [n]$ in B satisfying $c < \min_k B_{jk}$ for $j \in J_{\text{dom}}$. Then for each column k, the indices $j \in J_{\text{dom}}$ satisfy $B_{jk} > c$, and as $|J_{\text{dom}}| \geq n - \lceil(1-\alpha)(n+1)\rceil$, the $(1-\alpha)$ quantile satisfies

$$\widehat{q}^{+}_{n,\alpha}(\{B_{jk}\}_{j=1}^{n}) \geq \min_{j \in J_{\text{dom}}} B_{jk} > c,$$

giving the lemma. □

Lemma A.2. *Let $B \in \{0,1\}^{n \times m}$. Assume that there exist rows $J \subset [n]$ with $|J| \geq \lfloor \alpha(n+1) \rfloor$ and columns $K \subset [m]$ such that $|K| \geq \lceil (1-\delta)m \rceil$ such that $b_{jk} = 1$ for $j \in J$ and $k \in K$. Then*

$$\sum_{k=1}^{m} \widehat{q}^{+}_{n,\alpha}\left(\{B_{jk}\}_{j=1}^{n}\right) \geq (1-\delta)m.$$

Proof Fix any column $k \in K$. Then at least $\lfloor \alpha(n+1) \rfloor$ elements of $\{B_{jk}\}_{j=1}^{n}$ are 1 (those in J), so that $\widehat{q}^{+}_{n,\alpha}(\{B_{jk}\}_{j=1}^{n}) = 1$ for these $k \in K$. Thus

$$\sum_{k=1}^{m} \widehat{q}^{+}_{n,\alpha}\left(\{B_{jk}\}_{j=1}^{n}\right) \geq \sum_{k \in K} \widehat{q}^{+}_{n,\alpha}\left(\{B_{jk}\}_{j=1}^{n}\right) = |K| \geq (1-\delta)m$$

as desired. □

We return to the main thread. On the event (10), there exists a set J_{bad} of indices where coverage fails and $|J_{\text{bad}}| \geq \lfloor \alpha(n_{m+1}+1) \rfloor$, that is, such that

$$Y_j^{m+1} > \max_{i \in [m]} \widehat{f}_{-i}(X_j^{m+1}) + \widehat{q}^{+}_{m,\delta}(\{S^k_{1-\alpha}\}) \text{ or } Y_j^{m+1} < \min_{i \in [m]} \widehat{f}_{-i}(X_j^{m+1}) - \widehat{q}^{+}_{m,\delta}(\{S^k_{1-\alpha}\})$$

for each $j \in J_{\text{bad}}$. We can also be a bit more precise about the indices of $\{S^k_{1-\alpha}\}_{k=1}^{m}$: by the definitions of the quantiles $\widehat{q}^{+}_{m,\delta}$, there exists an index set $K_{\text{bad}} \subset [m]$ such that $|K_{\text{bad}}| \geq \lceil (1-\delta)(m+1) \rceil$ and for each $k \in K_{\text{bad}}$ and $j \in J_{\text{bad}}$,

$$Y_j^{m+1} > \max_{i \in [m]} \widehat{f}_{-i}(X_j^{m+1}) + S^k_{1-\alpha} \quad \text{or} \quad Y_j^{m+1} < \min_{i \in [m]} \widehat{f}_{-i}(X_j^{m+1}) - S^k_{1-\alpha}$$

by taking K_{bad} to be the order statistics of $S^k_{1-\alpha}$.

We now show that on event (10), the environment $m+1$ is strange, that is, $A_{m+1,\bullet}$ is large. We have

$$\begin{aligned}
\sum_{k=1}^{m+1} A_{m+1,k} &= \sum_{k=1}^{m+1} 1\left\{R_{k,m+1} < \min_{k'} R_{m+1,k'}\right\} \\
&= \sum_{k=1}^{m+1} 1\left\{R_{k,m+1} < \min_{k'} \widehat{q}^{+}_{n_{m+1},\alpha}\left(\{|Y_j^{m+1} - \widehat{f}_{-k'}(X_j^{m+1})|\}_j\right)\right\} \\
&\geq \sum_{k=1}^{m+1} \widehat{q}^{+}_{n_{m+1},\alpha}\left(\left[1\{R_{k,m+1} < \min_{k'} |Y_j^{m+1} - \widehat{f}_{-k'}(X_j^{m+1})|\}\right]_{j=1}^{n_{m+1}}\right),
\end{aligned}$$

where we use Lemma A.1 with the choices $c = R_{k,m+1}$ and B as the residual matrix with entries $B_{jk} = |Y_j^{m+1} - \widehat{f}_{-k}(X_j^{m+1})|$.

Finally, recall that by construction of the residual quantile matrix (8) we have $R_{k,m+1} = S_{1-\alpha}^k$. Then $R_{k,m+1} < \min_{k'} |Y_j^{m+1} - \widehat{f}_{-k'}(X_j^{m+1})|$ if and only if $S_{1-\alpha}^k < \min_{k'} |Y_j^{m+1} - \widehat{f}_{-k'}(X_j^{m+1})|$, which in turn occurs if and only if

$$Y_j^{m+1} < \min_{k'} \widehat{f}_{-k'}(X_j^{m+1}) - S_{1-\alpha}^k \quad \text{or} \quad Y_j^{m+1} > \max_{k'} \widehat{f}_{-k'}(X_j^{m+1}) + S_{1-\alpha}^k.$$

Revisiting the sum above, then, we have

$$\begin{aligned} A_{m+1,\bullet} &\geq \sum_{k=1}^{m+1} \widehat{q}^{+}_{n_{m+1},\alpha}\left(\left[1\left\{Y_j^{m+1} \notin \left[\min_{k'} \hat{f}_{-k'}(X_j^{m+1}) - S_{1-\alpha}^k, \max_{k'} \hat{f}_{-k'}(X_j^{m+1}) + S_{1-\alpha}^k\right]\right\}\right]_{j=1}^{n_{m+1}}\right) \\ &\geq (1-\delta)(m+1), \end{aligned}$$

where in the last line we used Lemma A.2 with the choice $B_{jk} = 1\{Y_j^{m+1} \notin [\min_{k'} \hat{f}_{-k'}(X_j^{m+1}) - S_{1-\alpha}^k, \max_{k'} \hat{f}_{-k'}(X_j^{m+1}) + S_{1-\alpha}^k]\}$, recognizing that $B_{jk} = 1$ for all indices $j \in J_{\text{bad}}$ and $k \in K_{\text{bad}}$. In particular, we have shown that $m+1 \in \mathcal{S}(A)$. As $\mathbb{P}(m+1 \in \mathcal{S}(A)) \leq \delta$ by step 2 (recall Eq. (9)), we have the theorem.

A.2 Proof of Theorem 2

With loss of generality, let

$$D_1 = \{1, 2, \ldots, m\gamma\}, \quad D_2 = \{m\gamma+1, m\gamma+2, \ldots, m\}.$$

We show that the rank of $\{S_{1-\alpha}^{m+1}\}$ among $\{S_{1-\alpha}^{m\gamma+1}, S_{1-\alpha}^{m\gamma+2}, \ldots, S_{1-\alpha}^{m+1}\}$ is uniformly distributed over $\{1, 2, \ldots, m(1-\gamma)+1\}$. For simplicity, we assume there are no ties. The proof can be modified to address the case where ties exists, and are broken randomly. For any $i \in \{m\gamma+1, m\gamma+2, \ldots, m+1\}$, define $\text{Rank}(S_{1-\alpha}^i)$ to be the rank of $S_{1-\alpha}^i$ among $\{S_{1-\alpha}^j\}_{m\gamma+1, m\gamma+2, \ldots, m+1}$. Moreover, define

$$Z^i := \left\{X_j^i, Y_j^i\right\}_{j=1}^{n_i}, \; 1 \leq i \leq m+1.$$

For any permutation π on $\{1, 2, \ldots, m(1-\gamma)+1\}$, let C_π be the set such that

$$\{(Z^1, Z^2, \ldots, Z^{m+1}) \in C_\pi\} = \{\text{Rank}(S_{1-\alpha}^{m\gamma+j}) = \pi(j), \; 1 \leq j \leq m(1-\gamma)+1\}.$$

By exchangeability, for any permutations π on $\{1, 2, \ldots, m(1-\gamma)+1\}$, we have

$$\begin{aligned} &\mathbb{P}\left(\left\{\text{Rank}(S_{1-\alpha}^{m\gamma+j}) = \pi(j), \; 1 \leq j \leq m(1-\gamma)+1\right\}\right) \\ &= \mathbb{P}\left(\left\{\left(Z^1, Z^2, \ldots, Z^{m\gamma}, Z^{m\gamma+1}, Z^{m\gamma+2}, \ldots, Z^{m+1}\right) \in C_\pi\right\}\right) \\ &= \mathbb{P}\left(\left\{\left(Z^1, Z^2, \ldots, Z^{m\gamma}, Z^{m\gamma+\pi(1)}, Z^{m\gamma+\pi(2)}, \ldots, Z^{m\gamma+\pi(m(1-\gamma)+1)}\right) \in C_\pi\right\}\right) \\ &\leq \mathbb{P}\left(\left\{\text{Rank}(S_{1-\alpha}^{m\gamma+j}) = j, \; 1 \leq j \leq m(1-\gamma)+1\right\}\right). \end{aligned}$$

By symmetry, we also have

$$\begin{aligned} &\mathbb{P}\left(\left\{\text{Rank}(S_{1-\alpha}^{m\gamma+j}) = j, \; 1 \leq j \leq m(1-\gamma)+1\right\}\right) \\ &\leq \mathbb{P}\left(\left\{\text{Rank}(S_{1-\alpha}^{m\gamma+j}) = \pi(j), \; 1 \leq j \leq m(1-\gamma)+1\right\}\right). \end{aligned}$$

As a result, for any permutation π on $\{1, 2, \ldots, m(1-\gamma)+1\}$,

$$\mathbb{P}\left(\left\{\text{Rank}\left(S_{1-\alpha}^{m\gamma+j}\right) = \pi(j),\ 1 \leq j \leq m(1-\gamma)+1\right\}\right) = \frac{1}{(m(1-\gamma)+1)!},$$

which implies that the rank of $\{S_{1-\alpha}^{m+1}\}$ among $\{S_{1-\alpha}^{m\gamma+1}, S_{1-\alpha}^{m\gamma+2}, \ldots, S_{1-\alpha}^{m+1}\}$ is uniformly distributed over $\{1, 2, \ldots, m(1-\gamma)+1\}$.

Next, we show that with probability at most δ,

$$\sum_{j=1}^{n_{m+1}} 1\left\{Y_j^{m+1} \in \widehat{C}_{m,\alpha,\delta}^{\text{split}}\left(X_j^{m+1}\right)\right\} < \lceil (1-\alpha)\left(n_{m+1}+1\right) \rceil. \tag{11}$$

Since $S_{1-\alpha}^{m+1}$ is the $\lceil (1-\alpha)\left(n_{m+1}+1\right) \rceil$-th largest residuals among all residuals in environment $m+1$, if

$$S_{1-\alpha}^{m+1} \leq \widehat{q}_{m,\delta}^{+}\left(\left\{S_{1-\alpha}^{i}\right\}_{i=m\gamma+1}^{m}\right),$$

then at least $\lceil (1-\alpha)\left(n_{m+1}+1\right) \rceil$ samples in environment $m+1$ will have residuals less than or equal to $\widehat{q}_{m,\delta}^{+}\left(\left\{S_{1-\alpha}^{i}\right\}_{i=m\gamma+1}^{m}\right)$. For these samples, their corresponding outcomes will be covered by the predictive intervals, which contradicts inequality (11). As a result, we know inequality (11) implies

$$S_{1-\alpha}^{m+1} > \widehat{q}_{m,\delta}^{+}\left(\left\{S_{1-\alpha}^{i}\right\}_{i=m\gamma+1}^{m}\right).$$

Therefore

$$\begin{aligned}
&\mathbb{P}\left[\sum_{j=1}^{n_{m+1}} 1\left\{Y_j^{m+1} \in \widehat{C}_{m,\alpha,\delta}^{\text{split}}\left(X_j^{m+1}\right)\right\} < \lceil (1-\alpha)\left(n_{m+1}+1\right) \rceil\right] \\
&\leq \mathbb{P}\left[S_{1-\alpha}^{m+1} > \widehat{q}_{m,\delta}^{+}\left(\left\{S_{1-\alpha}^{i}\right\}_{i=m\gamma+1}^{m}\right)\right] \leq \delta,
\end{aligned}$$

where the last step is due the fact that the rank of $\{S_{1-\alpha}^{m+1}\}$ among $\{S_{1-\alpha}^{m\gamma+1}, S_{1-\alpha}^{m\gamma+2}, \ldots, S_{1-\alpha}^{m+1}\}$ is uniformly distributed over $\{1, 2, \ldots, m(1-\gamma)+1\}$.

On the other hand, the predictive sets $\left\{\widehat{C}_{m,\alpha,\delta}^{\text{split}}\left(X_j^{m+1}\right)\right\}_{1 \leq j \leq n_{m+1}}$ do not provide valid coverage (i.e. inequality (11) holds) if the corresponding $S_{1-\alpha}^{m+1}$ is among the $m(1-\gamma) - \lceil (m(1-\gamma)+1)(1-\delta) \rceil$ largest in $\left\{S_{1-\alpha}^{m(1-\gamma)+1}, S_{1-\alpha}^{m(1-\gamma)+2}, \ldots, S_{1-\alpha}^{m}\right\}$. By symmetry, we have

$$\begin{aligned}
&\mathbb{P}\left[\sum_{j=1}^{n_{m+1}} 1\left\{Y_j^{m+1} \in \widehat{C}_{m,\alpha,\delta}^{\text{split}}\left(X_j^{m+1}\right)\right\} < \lceil (1-\alpha)\left(n_{m+1}+1\right) \rceil\right] \\
&\geq \frac{m(1-\gamma) - \lceil \left(m(1-\gamma)+1\right)(1-\delta) \rceil}{m(1-\gamma)+1} \\
&\geq \delta - \frac{1}{m(1-\gamma)+1}.
\end{aligned}$$

This completes the proof.

A.3 Proof of Theorem 3

The proof is quite similar to that of Theorem 1, with appropriate redefinitions of residual matrices and the A matrix. We begin with the analogues of $\widetilde{f}_{-(i,k)}$ and the R matrix (8). Define the collections of confidence set mappings

$$\left\{\widetilde{C}_\tau^{-(i,k)}\right\}_{\tau\in\mathbb{R}} = \mathsf{A}\left(\{X^l, Y^l\}_{l\neq i, l\neq k}\right),$$

so that

$$\widetilde{C}_\tau^{-(i,m+1)} = \widehat{C}_\tau^{-i} \text{ for } i \in [m].$$

We can then define the $(1-\alpha)$-quantile residual matrix $R \in \mathbb{R}^{(m+1)\times(m+1)}$ with entries

$$R_{ik} = \begin{cases} +\infty & \text{if } i = k \\ \widehat{q}_\alpha^+\left(\left[\inf\left\{\tau \mid Y_j^i \in \widetilde{C}_\tau^{-(i,k)}(X_j^i)\right\}\right]_{j=1}^{n_i}\right) & \text{if } i \neq k, \end{cases} \tag{12}$$

so that again $R_{i,m+1} = S_{1-\alpha}^i$ in Alg. 5, while $R_{m+1,i} = \widehat{q}_{n,\alpha}^+([\inf\{\tau \mid Y_j^{m+1} \in \widehat{C}_\tau^{-i}(X_j^{m+1})]_{j=1}^{n_{m+1}})$ gives quantiles for coverage on the new environment $m+1$. We define the matrix A identically as in the proof of Theorem 1, $A_{ik} = 1\{\min_{k'} R_{ik'} > R_{ki}\}$ and the set of strange environments $\mathcal{S}(A) = \{i \in [m+1] \mid A_{i,\bullet} \geq (1-\delta)(m+1)\}$ as before.

We again have that $|\mathcal{S}(A)| \leq \delta(m+1)$ and that $\mathbb{P}(m+1 \in \mathcal{S}(A)) \leq \delta$, as in Eq. (9). We show that if coverage in environment $m+1$, fails, then environment $m+1$ is strange. To that end, suppose that coverage fails, that is,

$$\sum_{j=1}^{n_{m+1}} 1\left\{Y_j^{m+1} \in \widehat{C}_{m,\alpha,\delta}^{\text{jk-minmax}}(X_j^{m+1})\right\} < \lceil(1-\alpha)(n_{m+1}+1)\rceil. \tag{13}$$

Recall that for the threshold $\widehat{\tau} = \widehat{q}_\delta^+(\{S_{1-\alpha}^i\}_{i=1}^m)$, Algorithm 5 sets $\widehat{C}_{m,\alpha,\delta}^{\text{jk-minmax}}(x) = \cup_{i=1}^m \widehat{C}_{\widehat{\tau}}^{-i}(x)$. Then on the event (13), there necessarily exists a set J_{bad}, $|J_{\text{bad}}| \geq \lfloor\alpha(n_{m+1}+1)\rfloor$, such that coverage fails for examples X_j^{m+1} whose indices $j \in J_{\text{bad}}$:

$$Y_j^{m+1} \notin \bigcup_{i=1}^m \widehat{C}_{\widehat{\tau}}^{-i}(X_j^{m+1}) \text{ for } j \in J_{\text{bad}}.$$

By definition of $\widehat{\tau}$ as the quantile $\widehat{q}_\delta^+(\{S_{1-\alpha}^i\})$, then, we also see that there exists a set $K_{\text{bad}} \subset [m]$ with cardinality $|K_{\text{bad}}| \geq \lceil(1-\delta)(m+1)\rceil$ and for which if $k \in K_{\text{bad}}$ and $j \in J_{\text{bad}}$, we have

$$Y_j^{m+1} \notin \bigcup_{i=1}^m \widehat{C}_{S_{1-\alpha}^k}^{-i}(X_j^{m+1}).$$

With these equivalences of failing to cover, we replicate the chain of inequalities in the proof of Theorem 1. Assuming event (13) occurs, we have

$$\begin{aligned} A_{m+1,\bullet} &= \sum_{k=1}^{m+1} 1\left\{R_{k,m+1} < \min_{k'} R_{m+1,k'}\right\} \\ &= \sum_{k=1}^{m+1} 1\left\{R_{k,m+1} < \min_{k'} \widehat{q}_\alpha^+\left(\left[\inf\left\{\tau \mid Y_j^i \in \widehat{C}_\tau^{-k'}(X_j^{m+1})\right\}\right]_{j=1}^{n_{m+1}}\right)\right\} \\ &\geq \sum_{k=1}^{m+1} \widehat{q}_\alpha^+\left(\left[1\left\{R_{k,m+1} < \min_{k'} \inf\left\{\tau \mid Y_j^i \in \widehat{C}_\tau^{-k'}(X_j^{m+1})\right\}\right\}\right]_{j=1}^{n_{m+1}}\right) \end{aligned}$$

where the inequality follows from Lemma A.1. But of course, by the construction (12) of the residual matrix, we have $R_{k,m+1} = S^k_{1-\alpha}$, and $S^k_{1-\alpha} < \min_{k'} \inf\{\tau \mid Y^i_j \in \widehat{C}^{-k'}_\tau(X^{m+1}_j)\}$ if and only if

$$Y_j^{m+1} \not\in \widehat{C}^{-k'}_{S^k_{1-\alpha}}(X_j^{m+1}) \;\; \text{for any } k'$$

by the assumed nesting property of the confidence sets $\widehat{C}_\tau$. We therefore obtain

$$A_{m+1,\bullet} \ge \sum_{k=1}^{m+1} \widehat{q}^+_\alpha \left(\left[1\left\{ Y_j^{m+1} \not\in \cup_{i=1}^m \widehat{C}^{-i}_{S^k_{1-\alpha}}(X_j^{m+1}) \right\} \right]_{j=1}^{n_{m+1}} \right) \ge (1-\delta)(m+1),$$

where the final inequality uses Lemma A.2 and that the index sets J_{bad} and K_{bad} have cardinalities $|K_{\text{bad}}| \ge \lceil (1-\delta)(m+1) \rceil$ and $|J_{\text{bad}}| \ge \lfloor \alpha(n_{m+1}+1) \rfloor$.

On the event (13), the environment $m+1$ is thus strange, and so applying the probability bound (9) gives Theorem 3.

A.4 Proof of Lemma 5.1

Let $t = \mathsf{Q}_\alpha(Q)$ and $u > 0$ be otherwise arbitrary, and let $v > 0$ be such that $Q(z \le t - u/2) \le \alpha - v$ and $Q(Z \le t + u/2) \ge \alpha + v$. We show that for small enough $\epsilon > 0$, if $\|P - Q\|_{\text{BL}} \le \epsilon$, then

$$P(Z \le t - u) \le \alpha - \frac{v}{2} \;\; \text{and} \;\; P(Z \le t + u) \ge \alpha + \frac{v}{2}.$$

As $\mathsf{Q}_\alpha(P) = \inf\{t' \mid P(Z \le t') \ge \alpha\}$, we then immediately see that $t - u \le \mathsf{Q}_\alpha(P) \le t + u$, and as u is otherwise arbitrary, this proves the lemma.

To see the first claim, let $0 < \delta \le u/2$, and define the $1/\delta$-Lipschitz continuous and bounded function

$$f_\delta(z) \coloneqq \begin{cases} 1 & \text{if } z \le t \\ 1 - z/\delta & \text{if } t \le z \le t + \delta \\ 0 & \text{if } t + \delta \le z, \end{cases}$$

which approximates the threshold $1\{z \le t\}$. Then we have

$$P(Z \le t - u) \le P f_\delta(Z + u) \overset{(\star)}{\le} Q f_\delta(Z + u) + \frac{\epsilon}{\delta} \le Q(Z + u \le t + \delta) + \frac{\epsilon}{\delta},$$

where inequality $(\star)$ follows because $\|P - Q\|_{\text{BL}} \le \epsilon$. As $\delta \le u/2$, we have

$$Q(Z \le t + \delta - u) \le Q(Z \le t - u/2) \le \alpha - v,$$

and so we have

$$P(Z \le t - u) \le \alpha - v + \frac{\epsilon}{\delta}.$$

Any $\epsilon < v\delta/2$ thus guarantees $P(Z \le t - u) \le \alpha - v/2$. A completely similar argument gives $P(Z \le t + u) \ge \alpha + v/2$ for small enough ϵ.

A.5 Proof of Example 2

To see how Assumption A1.a follows, note that

$$\left| \tau(x, y, \widehat{C}^{-i}) - \tau(x, y, C) \right| = \left| |\widehat{f}_{-i}(x) - y| - |f(x) - y| \right| \le \left| \widehat{f}_{-i}(x) - f(x) \right|,$$

so

$$\sup_{y} \sup_{x \in \mathcal{X}_\epsilon} \left| \tau(x, y, \widehat{C}^{-i}) - \tau(x, y, C) \right| \overset{a.s.}{\to} 0$$

for any $\epsilon > 0$. Let $\widehat{\tau} = \tau(\cdot, \cdot, \widehat{C}^{-i})$ for shorthand and $\tau = \tau(\cdot, \cdot, C)$. Then we claim that

$$\left\| \mathcal{L}(\widehat{\tau} \mid \widehat{P}^i) - \mathcal{L}(\tau \mid P^i) \right\|_{\rm BL} \le \left\| \mathcal{L}(\widehat{\tau} \mid \widehat{P}^i) - \mathcal{L}(\tau \mid \widehat{P}^i) \right\|_{\rm BL} + \left\| \mathcal{L}(\tau \mid \widehat{P}^i) - \mathcal{L}(\tau \mid P^i) \right\|_{\rm BL}. \quad (14)$$

We consider the two terms in turn. For the first, let $\eta > 0$ and consider the event that $|\tau(x, y, \widehat{C}^{-i}) - \tau(x, y, C)| \le \eta$ for $x \in \mathcal{X}_\epsilon$, which occurs eventually (with probability 1). Note that for any function h with $\|h\|_\infty \le 1$ and $\|h\|_{\rm Lip} \le 1$, we have

$$\begin{aligned} \int [h(\widehat{\tau}(x, y)) - h(\tau(x, y))] \, d\widehat{P}^i(x, y) &= \int_{\mathcal{X}_\epsilon} [h(\widehat{\tau}(x, y)) - h(\tau(x, y))] \, d\widehat{P}^i(x, y) + \int_{\mathcal{X}_\epsilon^c} [h(\widehat{\tau}) - h(\tau)] \, d\widehat{P}^i \\ &\le \widehat{P}^i(\mathcal{X}_\epsilon) \sup_{x \in \mathcal{X}_\epsilon, y} |\widehat{\tau}(x, y) - \tau(x, y)| + 2\widehat{P}^i(\mathcal{X}_\epsilon^c) \\ &\le \eta + 2\widehat{P}^i(\mathcal{X}_\epsilon^c). \end{aligned}$$

The final term converges a.s. to $P^i(\mathcal{X}_\epsilon^c) \le P_X(\mathcal{X}_\epsilon^c) + \sqrt{P_X(\mathcal{X}_\epsilon^c)} \rho_{\chi^2} \le \epsilon + \rho_{\chi^2} \sqrt{\epsilon}$. For the second term in (14), τ is fixed and so standard bounded Lipschitz convergence [25] guarantees its a.s. convergence to 0. Then with probability 1, for any $\epsilon > 0$ and $\eta > 0$, we have

$$\limsup_n \left\| \mathcal{L}(\widehat{\tau} \mid \widehat{P}^i) - \mathcal{L}(\tau \mid P^i) \right\|_{\rm BL} \le \eta + 2(\epsilon + \rho_{\chi^2} \sqrt{\epsilon}),$$

which gives Assumption A1.a.

For Assumption A2.a, let λ be Lebesgue measure, and recognize that for any τ_0, τ, we have

$$\widehat{C}_{\tau_0}^{-i}(x) \triangle C_\tau(x) = [\widehat{f}_{-i}(x) \pm \tau_0] \triangle [f(x) \pm \tau],$$

so $\lambda(\widehat{C}_{\tau_0}^{-i}(x) \triangle C_\tau(x)) \le 2|f(x) - \widehat{f}_{-i}(x)| + 2|\tau - \tau_0|$. For any $\epsilon > 0$, if $|\tau^\star - \tau| \le \epsilon/4$, the sets $B_{n,\tau}$ in Assumption A2.a satisfy

$$B_{n,\tau}^i = \left\{ x \mid \lambda(\widehat{C}_\tau^{-i}(x) \triangle C_{\tau^\star}(x)) \ge \epsilon \right\} \subset \left\{ x \in \mathcal{X} \mid |\widehat{f}_{-i}(x) - f(x)| \ge \epsilon/2 \right\}.$$

The conditions (5) guarantee that for any large enough n such that $\tau = \tau(n)$ satisfies $|\tau(n) - \tau^\star| \le \epsilon/4$ and any $m \in \mathbb{N}$,

$$P_X\left(\bigcup_{i=1}^m B_{n,\tau}^i \right) \le P_X\left(\bigcup_{i=1}^m \left\{ x \mid |\widehat{f}_{-i}(x) - f(x)| \ge \epsilon/2 \right\} \right) \to 0$$

as $n \to \infty$. This then must occur for any slowly enough growing sequence $m(n)$.

A.6 Proof of Example 3

The argument to justify Assumption A1.a is similar to that in Example 2 (see Section A.5): as $C_\tau(x) = [l(x) - \tau, u(x) + \tau]$ and $\tau(x, y, C) = \max\{l(x) - y, y - u(x)\}$, we have

$$\begin{aligned} |\tau(x, y, C) - \tau(x, y, \widehat{C}^{-i})| &= \left| \max\{l(x) - y, y - u(x)\} - \max\{\widehat{l}_{-i}(x) - y, y - \widehat{u}_{-i}(x)\} \right| \\ &\le \left| |l(x) - y| - |\widehat{l}_{-i}(x) - y| \right| + \left| |y - u(x)| - |y - \widehat{u}_{-i}(x)| \right| \\ &\le \left| \widehat{l}_{-i}(x) - l(x) \right| + \left| \widehat{u}_{-i}(x) - u(x) \right|. \end{aligned}$$

In particular, as in Example 2, we have $\sup_y \sup_{x \in \mathcal{X}_\epsilon} |\tau(x,y,C) - \tau(x,y,\widehat{C}^{-i})| \overset{a.s.}{\to} 0$, and so the bounded Lipschitz convergence argument there applies and Assumption A1.a follows.

To obtain the conditions in Assumption A2.a, recognize that for Lebesgue measure λ an application of the triangle inequality gives

$$\lambda\left(\widehat{C}^{-i}_{\tau_0}(x) \triangle C_\tau(x)\right) \le 2\left(\left|\widehat{l}_{-i}(x) - l(x)\right| + \left|\widehat{u}_{-i}(x) - u(x)\right| + |\tau - \tau_0|\right)$$

for any τ, τ_0. The remainder of the argument is, *mutatis mutandis*, identical to that in Example 2.

A.7 Proof of Example 4

We present an analogous argument to that we use in Example 2, Section A.5 to show how Assumptions A1.a and A2.a follow from the convergence (7). In this case, the loss ℓ is Lipschitz continuous, and so as

$$C_\tau(x) = \{y \in [k] \mid \ell(y, f(x)) \le \tau\}$$

and $\tau(x,y,C) = \ell(y, f(x))$, we have

$$\max_{y \in [k]} \sup_{x \in \mathcal{X}_\epsilon} \left|\tau(x,y,\widehat{C}^{-i}) - \tau(x,y,C)\right| \overset{a.s.}{\to} 0$$

under the convergence (7). Then exactly as in the proof of Example 2 in Section A.5, we have

$$\left\|\mathcal{L}(\tau(X,Y,\widehat{C}^{-i}) \mid \widehat{P}^i) - \mathcal{L}(\tau(X,Y,C) \mid P^i)\right\|_{\mathrm{BL}} \overset{a.s.}{\to} 0,$$

implying Assumption A1.a holds.

Consider Assumption A2.a. Let $\tau^\star = \tau^\star(\delta, \alpha)$ for shorthand and $x \not\in D_{\tau^\star,\epsilon}$, so there is no y such that $|\ell(y, f(x)) - \tau^\star| < \epsilon$. Then

$$C_{\tau^\star}(x) = \{y \mid \ell(y, f(x)) \le \tau^\star\} = \{y \mid \ell(y, f(x)) \le \tau^\star - \epsilon\} = \{y \mid \ell(y, f(x)) \le \tau^\star + \epsilon\}$$

by definition of $D_{\tau^\star,\epsilon}$. The Lipschitz continuity of $v \mapsto \ell(y, v)$ implies there exists $\eta > 0$ such that if $\|\widehat{f}^{-i}(x) - f(x)\| \le \eta$, we have $|\ell(y, \widehat{f}^{-i}(x)) - \ell(y, f(x))| \le \epsilon/4$, and so if $|\tau - \tau^\star| < \epsilon/4$ and $\|\widehat{f}^{-i}(x) - f(x)\| \le \eta$, then $\ell(y, f(x)) \le \tau^\star - \epsilon$ implies $\ell(y, \widehat{f}^{-i}(x)) \le \tau^\star - \epsilon/4 \le \tau - \epsilon/2$, and similarly, $\ell(y, f(x)) > \tau^\star + \epsilon$ implies that $\ell(y, \widehat{f}^{-i}(x)) > \tau^\star + 3\epsilon/4 \ge \tau + \epsilon/2$. That is,

$$C_{\tau^\star}(x) = \widehat{C}^{-i}_\tau(x).$$

Recalling the notation of Assumption A2.a, the sets $B_{n,\tau}$ then satisfy

$$(B^i_{n,\tau})^c = \left\{x \mid \widehat{C}^{-i}_\tau(x) = C_{\tau^\star}(x)\right\} \supset \left\{x \in \mathcal{X} \mid \|\widehat{f}_{-i}(x) - f(x)\| \le \eta, |\tau - \tau^\star| \le \epsilon/4, x \not\in D_{\tau^\star,\epsilon}\right\}.$$

As by assumption the sets $\mathcal{X}_\epsilon$ on which $\widehat{f}^{-i}$ uniformly converges have $P_X(\mathcal{X}_\epsilon) \le \epsilon$, the convergence (7) yields that if $\tau = \tau(n) \to \tau^\star$, then for any fixed $m \in \mathbb{N}$,

$$\lim_{n \to \infty} P_X\left(\bigcup_{i=1}^m B^i_{n,\tau}\right) \to 0.$$

Once again, this must occur for any sequence $m(n)$ growing slowly enough to ∞.

A.8 Proof of Lemma 5.2

We show the argument in a few steps, first showing that $S^i_{1-\alpha}$ and $\mathsf{Q}_{1-\alpha}(P^i)$ are quite close, then using Assumption A3 to show that the $1-\delta$ quantile of $\mathsf{Q}_{1-\alpha}(P^i)$ converges.

First, we leverage Lemma 5.1. Recalling that in Algorithm 5, the residuals $R^i_j = \tau(X^i_j, Y^i_j, \widehat{C}^{-i}) = \inf\{\tau \mid Y^i_j \in \widehat{C}^{-i}_\tau(X^i_j)\}$, we see that the empirical distribution $\widehat{P}^i_R = \frac{1}{n_i}\sum_{j=1}^{n_i} 1_{R^i_j}$ satisfies

$$\left\|\widehat{P}^i_R - \mathcal{L}(\tau(X, Y, C) \mid P^i)\right\|_{\mathrm{BL}} \overset{a.s.}{\to} 0$$

by Assumption A1.a. In particular, the assumption that $\tau(X, Y, C)$ has a density under $(X, Y) \sim P^i$ in a neighborhood of

$$\mathsf{Q}_{1-\alpha}(P^i) \coloneqq \inf\{\tau \mid P^i(\tau(X, Y, C) \le \tau) \ge 1-\alpha\}$$

then guarantees, via Lemma 5.1 and the continuous mapping theorem, that as $n_i \to \infty$ the quantile $S^i_{1-\alpha} = \widehat{q}^+_\alpha(\{R^i_j\}_{j=1}^{n_i})$ satisfies

$$S^i_{1-\alpha} - \mathsf{Q}_{1-\alpha}(P^i) \overset{a.s.}{\to} 0.$$

As an immediate consequence, we obtain

$$\max_{i \le m} |\mathsf{Q}_{1-\alpha}(P^i) - S^i_{1-\alpha}| \overset{a.s.}{\to} 0$$

for any fixed m, and hence a sequence $m(n)$ growing slowly enough as $m(n) \to \infty$ as $n \to \infty$. From this convergence, an application of the triangle inequality gives that if $\mathcal{L}(\mathsf{Q}_{1-\alpha}(P^E))$ denotes the induced probability law over $\mathsf{Q}_{1-\alpha}(P^E)$ by sampling $E \in \mathcal{E}$ and $\mathcal{L}(\{S^i_{1-\alpha}\}_{i=1}^m)$ denotes the empirical law of the S^i, then

$$\left\|\mathcal{L}(\mathsf{Q}_{1-\alpha}(P^E)) - \mathcal{L}(\{S^i_{1-\alpha}\}_{i=1}^{m(n)})\right\|_{\mathrm{BL}} \overset{a.s.}{\to} 0$$

as $n \to \infty$. Combining Assumption A3 and Lemma 5.1 yields that

$$\widehat{q}^+_\delta\left(\{S^i_{1-\alpha}\}_{i=1}^m\right) \overset{a.s.}{\to} q(\delta),$$

where $q(\delta)$ in the assumption is the unique $1-\delta$ quantile of $\mathsf{Q}_{1-\alpha}(P^E)$ over random E.

B Performance of Multi-environment Jackknife+ Quantile

One may consider a slightly different jackknife algorithm compared to Algorithm 1:

```
Algorithm 8: Multi-environment Jackknife+ Quantile: the regression case
Input: samples $\{X^i_j, Y^i_j\}_{j=1}^{n_i}$, $i = 1, \ldots, m$, confidence levels $\alpha, \delta$
For $i = 1, \ldots, m$, set
    $\widehat{f}_{-i} = \mathsf{A}\left((X^1, Y^1), \ldots, (X^{i-1}, Y^{i-1}), (X^{i+1}, Y^{i+1}), \ldots, (X^m, Y^m)\right),$
and construct residual quantiles
    $R^i_j = |Y^i_j - \widehat{f}_{-i}(X^i_j)|, \quad j = 1, \ldots, n_i, \quad \text{and} \quad S^i_{1-\alpha} = \widehat{q}^+_{n_i,\alpha}\left(R^i_1, R^i_2, \ldots, R^i_{n_i}\right)$
Return confidence interval mapping
    $\widehat{C}_{m,\alpha,\delta}(x) \coloneqq \left[\widehat{q}^-_{m,\delta}\left(f_{-i}(x) - \{S^i_{1-\alpha}\}_{i=1}^m\right), \widehat{q}^+_{m,\delta}\left(f_{-i}(x) + \{S^i_{1-\alpha}\}_{i=1}^m\right)\right].$
```

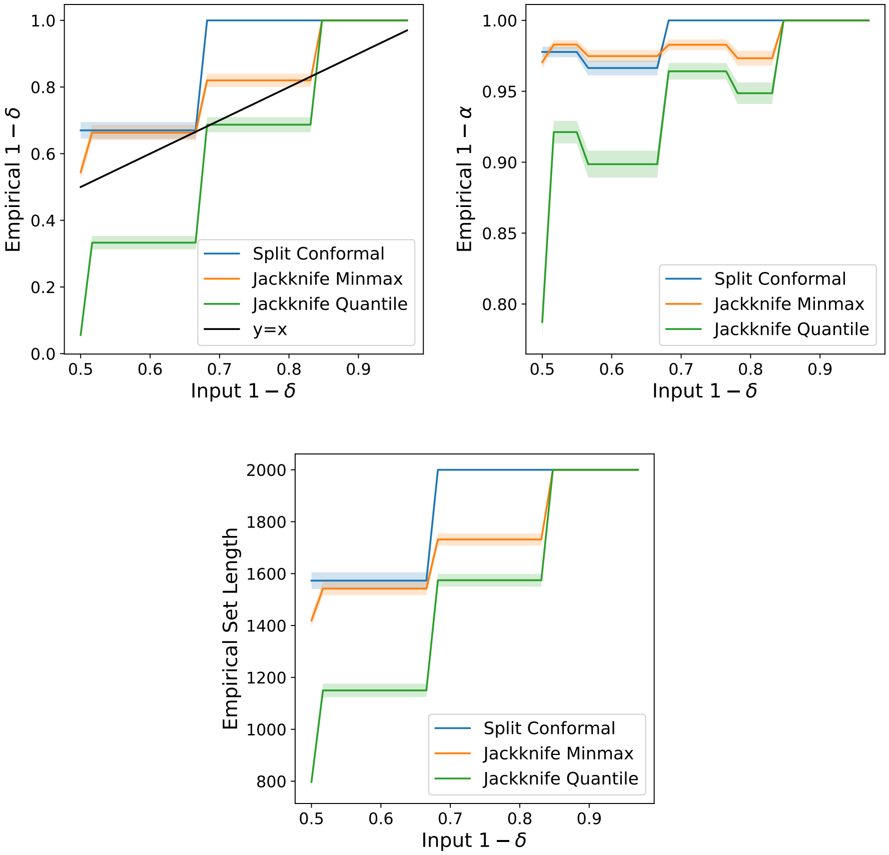

Figure B.11. Influence of input δ on the performance of Algorithms 1, 2, and 8 applied to the neurochemical sensing data. For these experiments, α is set to be 0.05. The plots shows the empirical $1 - \delta$, empirical $1 - \alpha$, and empirical set length for both the split conformal and jackknife-minmax algorithms with various input δ.

We apply Algorithms 1, 2, and 8 to the neurochemical sensing data introduced in Section 6.1. We set $\alpha = 0.05$, and the split ratio to be 0.5 for Algorithm 2. During each experiment, we vary the values of δ, and record the empirical $1 - \alpha$, $1 - \delta$, and set length. We repeat the experiment 100 times, and display the results in Figure B. We observe that although Algorithm 8 tends to output less conservative confidence intervals than the other two, it does not provide valid coverage except when the input $1 - \delta$ is large.

References

[1] A. Angelopoulos, E. Candès, and R. J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems 36*, 2023.

[2] A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foun-*

dations and Trends in Machine Learning, 16(4), 2023.

[3] D. Bang, K. T. Kishida, T. Lohrenz, J. P. White, A. W. Laxton, S. B. Tatter, S. M. Fleming, and P. R. Montague. Sub-second dopamine and serotonin signaling in human striatum during perceptual decision-making. *Neuron*, 108(5):999–1010.e6, 2020. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2020.09.015. URL https://www.sciencedirect.com/science/article/pii/S0896627320307157.

[4] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486 – 507, 2021. doi: 10.1214/20-AOS1965. URL https://doi.org/10.1214/20-AOS1965.

[5] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021.

[6] S. Beery, E. Cole, and A. Gjoka. The iWildCam 2020 competition dataset. *arXiv:2004.10340 [cs.CV]*, 2020.

[7] M. Cauchois, S. Gupta, and J. Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021.

[8] V. Chernozhukov, K. Wuthrich, and Y. Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the Thirty First Annual Conference on Computational Learning Theory*, 2018.

[9] R. Dunn, L. Wasserman, and A. Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, 118(544):2491–2502, 2023.

[10] I. Gibbs and E. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems 34*, 2021.

[11] C. Gupta, A. K. Kuchibhotla, and A. Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.108496. URL https://www.sciencedirect.com/science/article/pii/S0031320321006725.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[13] G. Imbens and D. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[14] K. T. Kishida, I. Saez, T. Lohrenz, M. R. Witcher, A. W. Laxton, S. B. Tatter, J. P. White, T. L. Ellis, P. E. M. Phillips, and P. R. Montague. Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences*, 113(1):200–205, 2016. doi: 10.1073/pnas.1513619112. URL https://www.pnas.org/doi/abs/10.1073/pnas.1513619112.

[15] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. WILDS: A benchmark of in-the-wild distribution shifts. 2021.

[16] Y. Lee, R. F. Barber, and R. Willett. Distribution-free inference with hierarchical data. *arXiv preprint arXiv:2306.06342*, 2023. URL https://arxiv.org/abs/2306.06342.

[17] J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society, Series B*, 76(1):71–96, 2014.

[18] G. Loewinger, P. Patil, K. T. Kishida, and G. Parmigiani. Multi-study learning for real-time neurochemical sensing in humans using the "study strap ensemble". *bioRxiv*, 2021. doi: 10.1101/856385. URL https://www.biorxiv.org/content/early/2021/01/14/856385.

[19] G. Loewinger, P. Patil, K. T. Kishida, and G. Parmigiani. Hierarchical resampling for bagging in multistudy prediction with applications to human neurochemical sensing. *The Annals of Applied Statistics*, 16(4):2145 – 2165, 2022.

[20] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[21] Y. Romano, E. Patterson, and E. J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32*, 2019.

[22] Y. Romano, M. Sesia, and E. J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems 33*, 2020.

[23] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems 33*, 2020.

[24] R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32*, 2019.

[25] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

[26] V. Vovk, A. Grammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[27] J. R. Zech, M. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, November 2018.