Improving Coverage in Combined Prediction Sets with Weighted p-values

Gina Wong **Drew Prinster** **Suchi Saria** **Rama Chellappa** **Anqi Liu**
Johns Hopkins University

Abstract

Conformal prediction quantifies the uncertainty of machine learning models by augmenting point predictions with valid prediction sets. For complex scenarios involving multiple trials, models, or data sources, conformal prediction sets can be aggregated to create a prediction set that captures the overall uncertainty, often improving precision. However, aggregating multiple prediction sets with individual $1-\alpha$ coverage inevitably weakens the overall guarantee, typically resulting in $1-2\alpha$ worst-case coverage. In this work, we propose a framework for the *weighted aggregation of prediction sets*, where weights are assigned to each prediction set based on their contribution. Our framework offers flexible control over how the sets are aggregated, achieving tighter coverage bounds that interpolate between the $1-2\alpha$ guarantee of the combined models and the $1-\alpha$ guarantee of an individual model depending on the distribution of weights. Importantly, our framework generalizes to data-dependent weights, as we derive a procedure for weighted aggregation that maintains finite-sample validity even when the weights depend on the data. This extension makes our framework broadly applicable to settings where weights are learned, such as mixture-of-experts (MoE), and we demonstrate through experiments in the MoE setting that our methods achieve adaptive coverage.

1 INTRODUCTION

In recent years, machine learning models have achieved remarkable accuracy across a wide range of predictive tasks (Min et al., 2023; Liang et al., 2024a). Understanding the uncertainty associated with each prediction is essential to decision-making in real-world scenarios, but the black box nature of many of these models hinders their deployment in safety-critical applications such as medical diagnosis (Chua et al., 2023; Grote and Berens, 2023), industrial control systems (Kumar et al., 2023; Lawrence et al., 2024), and extreme weather forecasting (Eyring et al., 2024; Lai et al., 2024). Conformal prediction (Vovk et al., 2005) emerged as a popular wrapper method around machine learning models because it provides a statistically valid quantification of uncertainty. In particular, it transforms point predictions to prediction sets with finite-sample coverage guarantees, as long as the test data is exchangeable with the training data. Complex scenarios with multiple predictions—for example, when there are multiple trials, models, or data sources—naturally produce multiple conformal prediction sets (Figure 1).

A number of methods have been proposed to aggregate prediction sets. For conformal prediction, where the popular split conformal variant (Papadopoulos et al., 2002) introduces a one-time random split, there are multiple methods of aggregating predictions to reduce the randomness over multiple splits: popular examples include cross-conformal (Vovk, 2015), CV+ (Barber et al., 2021), and jackknife+ (Barber et al., 2021). These aggregation methods are all *symmetric*, in that individual prediction sets contribute equally to the aggregate set. Although comparatively less studied, a more general approach is to aggregate sets *asymmetrically*—that is, to weight sets based on their prior importance to the overall result (Gasparin and Ramdas, 2024a). Work in both symmetric and asymmetric aggregation establishes that aggregating individual prediction sets with $1-\alpha$ coverage guarantees results in an overall coverage guarantee of $1-2\alpha$ (Vovk and Wang, 2020).

Our work is based on the observation that if the overall coverage guarantee for asymmetric aggregation reflects the contributions of the individual sets, then we can achieve a tighter guarantee than a constant $1-2\alpha$. Consider, as an extreme, the case where almost all the weight is on one set; then, because the coverage

Correspondence to: gwong15@jhu.edu.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

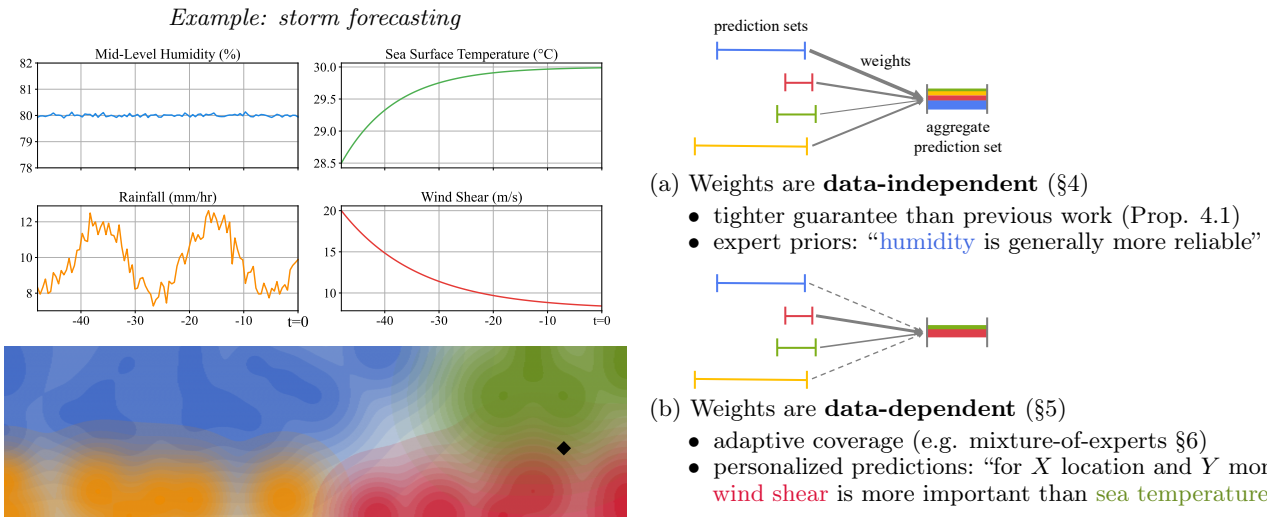

Figure 1: **Left:** Storm forecasting example with different models tracking humidity, sea temperature, rainfall, and wind shear over time. Below, an abstract representation of how models vary in predictive strength across the input space (colored regions). At the given test point (black), the red and green models dominate, so their prediction sets matter most. **Right:** Model prediction sets are combined with weighted aggregation. (a) *Data-independent* weights reflect expert priors (e.g. we up-weight the blue model as it is the most generally reliable, in the sense that it is the most predictive across the largest portion of the input space). (b) *Data-dependent* weights adapt to context, yielding forecasts better aligned with current conditions (e.g. at the test point, red and green dominate).

of that set dominates aggregation, the overall coverage guarantee should be closer to the $1-\alpha$ guarantee of the dominant individual set. To this end, we propose a method for asymmetric aggregation with *data-independent weights*, where the overall coverage guarantee is based on the distribution of weights across the prediction sets. This approach leverages the results of Vovk and Wang (2020) on averaging p-values to achieve tighter guarantees when there exists strong asymmetry in the importance of the prediction sets—i.e. when a single prediction set is significantly more important than the rest—with $1-2\alpha$ as the worst-case guarantee when the contributions of the prediction sets are more equal.

Data-independent weights can encode expert priors on the relative importance of the prediction sets, allowing weights to reflect prior information such as the reliability of the predictors used to generate the sets (Figure 1a). Already, these prior weights generalize standard symmetric aggregation. Yet in practice, many applications demand even more flexibility, and require weights that adapt directly to the data: for example, in the popular mixture-of-experts setting (Jacobs et al., 1991), expert weights are learned from inputs. To deal with this general case, we additionally propose a method for aggregating prediction sets with *data-dependent weights* via a linear transformation on the weighted average of p-values associated with each set. Our method allows us to construct a valid aggregate prediction set while also preserving the proportions of the weights assigned to the individual sets. Since data-dependent weights adjust the contributions of the individual sets based on the observed input, our method also achieves a form of conditional/data-adaptive coverage (Figure 1b). We apply our method to the mixture-of-experts setting, and we demonstrate its effectiveness in experiments with real and synthetic data.

To summarize, our main contributions are as follows:

- We propose a framework for the asymmetric aggregation of prediction sets based on weighted p-values (§4). With *data-independent weights*, our framework improves the coverage guarantee beyond the standard $1-2\alpha$, depending on the distribution of weights (Prop. 4.1).
- We derive a general method for transforming a random variable to a p-value variable, which enables the construction of a prediction set with finite-sample guarantees (§5). This method allows us to extend our framework to *data-dependent weights*, and we demonstrate that incorporating data-dependent weights provides adaptive coverage (Prop. 5.1, Prop. 5.2).
- We apply our method to the mixture-of-experts setting, and we demonstrate that using a weighted aggregation of experts improves local validity (§6).

2 BACKGROUND

We begin by reviewing preliminaries leading up to the basics of conformal prediction, with a focus on the widely-used split conformal method. To provide context for our method based on p-values, we establish the

connection between the more typical quantile presentation of conformal prediction and its original p-value presentation (Vovk et al., 2005).

Quantiles Given a random variable $X \in \mathcal{X}$, the left α-quantile of X, which we refer to as the *α-quantile* of X for simplicity, is defined as

$$q_\alpha(X) = \inf\{x \in \mathcal{X} : \mathbb{P}(X \leq x) \geq \alpha\}.$$

We denote the distribution function of X as F. The α-quantile is then the inverse of F; that is, $q_\alpha(X) = F^{-1}(\alpha)$.

For a set of n elements $Z = \{z_1, \ldots, z_n\}$, the left *α-empirical quantile* is given by

$$\widehat{Q}^-_\alpha(Z) = \text{the } \lfloor (n+1)\alpha \rfloor \text{-th smallest value of } Z,$$

and the right α-empirical quantile, or the $(1-\alpha)$-empirical quantile, by

$$\widehat{Q}^+_\alpha(Z) = \text{the } \lceil (n+1)(1-\alpha) \rceil \text{-th smallest value of } Z,$$

where the $n+1$ term serves as a finite sample correction.

p-values A *p-variable* (Vovk and Wang, 2020) is a random variable P' such that

$$\mathbb{P}\{P' \leq \alpha\} \leq \alpha \quad \forall \alpha \in (0,1). \tag{1}$$

The values taken by a p-variable are called *p-values*. In hypothesis testing, p-values represent the probability of observing results at least as extreme as the ones obtained, assuming the null hypothesis is true. When the inequality becomes equality, the p-variable is *precise* (rather than *conservative*) and follows a $\mathcal{U}(0,1)$ distribution.

Conformal prediction Suppose we have training data $\{(X_1, Y_1), \ldots (X_n, Y_n)\}$ and a test point (X_{n+1}, Y_{n+1}) with unknown label Y_{n+1}. Let $\mathcal{P}$ denote the joint distribution on $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. Then, assuming that the training and test data are *exchangeable*, or that the distribution of the training and test data is permutation-invariant, conformal prediction can be used to construct a valid prediction set for X_{n+1} with no further assumptions on $\mathcal{P}$.

Split conformal prediction, a popular variant, starts by splitting the training data indices into disjoint "proper" training and calibration subsets so that $\{1, \ldots, n\} = S_{\text{train}} \cup S_{\text{cal}}$. We fit a predictor to the subset S_{train}, so that $\widehat{\mu} = \mathcal{A}(\{(X_i, Y_i)\}_{i \in S_{\text{train}}})$, where $\mathcal{A}$ is a model fitting algorithm. Then, given a nonconformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we can compute the scores $R_i = s(X_i, Y_i)$, or the values of the score function that characterize how nonconformal a label Y_i is from its predicted value $\widehat{\mu}(X_i)$. We compute the scores on the subset S_{cal}. At a given significance level α, this procedure allows us to define the prediction set

$$\widehat{C}^{\text{split}}_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : \\ s(X_{n+1}, y) \leq \widehat{Q}^+_\alpha(\{s(X_i, Y_i)\}_{i \in S_{\text{cal}}})\Big\}.$$

By construction, the split conformal prediction set results in the (marginal) coverage guarantee

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\text{split}}_\alpha(X_{n+1})\} \geq 1 - \alpha \quad \forall \alpha \in (0,1).$$

In other words, the prediction set $\widehat{C}^{\text{split}}_\alpha(X_{n+1})$ is *valid*, or guaranteed to contain the true label Y_{n+1} with probability at least $1 - \alpha$.

As an alternative perspective, the quantile can be replaced with the p-value function

$$\widehat{p}(x, y) = \frac{1 + \sum_{i \in S_{\text{cal}}} \mathbb{1}\{s(x, y) < s(X_i, Y_i)\}}{|S_{\text{cal}}| + 1},$$

which returns the proportion of calibration points that are less conformal than some test point (x, y), with a finite sample correction. The equivalent prediction set constructed from the p-value function is

$$\widehat{C}^{\text{split}}_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : \widehat{p}(X_{n+1}, y) > \alpha\}.$$

3 RELATED WORK

We categorize existing work on combining conformal results into two approaches: methods for selecting or combining prediction sets, and methods for combining p-values.

Combining prediction sets Combining conformal prediction sets was first introduced by Lei et al. (2018), who propose combining K split conformal prediction sets by taking their intersection. Taking the intersection of these sets reduces the size of the resulting prediction set, but the authors demonstrate that, under general conditions, the intersection is wider than the individual sets with probability tending to 1 with an asymptotic number of samples. Another method proposed by Yang and Kuchibhotla (2021) selects the predictor (from a set of K predictors) that returns the best prediction set. They present two methods: one to select the predictor that gives the most efficient prediction set, but with only approximate validity, and one to select the predictor with the most valid prediction set (i.e. with minimal coverage slack), but with a width only close to the minimum. Liang et al. (2024b) point out that the approximate validity of the first method is due to selection bias, and introduce an alternative approach that both selects the most efficient predictor and uses it to construct a valid prediction set.

A recent line of work explores combining conformal prediction sets by majority vote with a $1-2\alpha$ coverage guarantee (Cherubin, 2019; Solari and Djordjilović, 2021; Gasparin and Ramdas, 2024a,b). In particular, Gasparin and Ramdas (2024a) characterize the width of the majority vote set and introduce many extensions, including a weighted majority vote method that incorporates prior information in the weights. Our work on combining conformal results by weighted p-values is closely related to this method. However, our formulation allows us to leverage the results of Vovk and Wang (2020) to improve the coverage guarantee beyond the standard $1-2\alpha$, and can also be generalized to data-dependent weights to enable adaptive coverage. (For a more detailed comparison, see §C.3.)

Combining p-values A substantial body of research has been devoted to developing methods for combining p-values. Here, we focus specifically on approaches that accommodate arbitrary dependence, with a particular emphasis on their applications to conformal prediction. (For a more comprehensive review, see Balasubramanian et al. (2015) and DiCiccio et al. (2020).)

An early example of combining p-values with arbitrary dependence is the Bonferroni method, where the minimum of a set of p-values is scaled by the number of p-values. This method has multiple extensions (Rüger, 1978; Hommel, 1983), and was first applied to conformal prediction by Lei et al. (2018).

Rüschendorf (1982) finds that twice the average of p-values is a p-value, a result that was later extended to a more general notion of average by Vovk and Wang (2020). Stutz et al. (2023) use the result of Rüschendorf (1982) to get a $1-2\alpha$ guarantee for the average of p-values, but also propose a novel transformation to get a p-value average with improved coverage. (For a discussion of how to extend their method to apply to our framework, see §C.4.)

Liu and Xie (2020) propose the Cauchy combination test, where the test statistic is a weighted sum of Cauchy-transformed p-values. Wu et al. (2022) apply the Cauchy combination test to conformal prediction, and show that their method yields asymptotically exact prediction sets under a weak stability assumption and a joint distribution assumption.

4 COMBINING PREDICTION SETS BY WEIGHTED p-VALUE

In this section, we extend the method of combining p-values, as proposed by Vovk and Wang (2020), to the setting of conformal prediction. Suppose we have K predictors, denoted by $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$, where each predictor $\widehat{\mu}_k$ is associated with a nonconformity score function s_k and calibration set S_k. For each predictor, we define a p-value function $\widehat{p}_k(x,y)$ that measures how well a candidate label y conforms to the predicted outcome for a given x:

$$\widehat{p}_k(x,y) = \frac{1 + \sum_{i \in S_k} \mathbb{1}\{s_k(x,y) < s_k(X_i,Y_i)\}}{|S_k|+1}. \quad (2)$$

The corresponding prediction set for X_{n+1} is

$$\widehat{C}^{\text{split}}_{\alpha,k}(X_{n+1}) = \{y \in \mathcal{Y} : \widehat{p}_k(X_{n+1}, y) > \alpha\},$$

with marginal coverage guarantee

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{\text{split}}_{\alpha,k}(X_{n+1})\right\} \geq 1-\alpha \quad \forall \alpha \in (0,1). \quad (3)$$

We assign a weight v_k to each p-value function $\widehat{p}_k$, yielding the weighted average p-value function

$$\bar{p}(x,y) = \sum_{k=1}^{K} v_k \widehat{p}_k(x,y) \quad \text{where} \sum_{k=1}^{K} v_k = 1, \quad (4)$$

with prediction set

$$\widehat{C}^{\text{avg}}_{\alpha}(X_{n+1}) = \{y \in \mathcal{Y} : \bar{p}(X_{n+1}, y) > \alpha\}. \quad (5)$$

We now provide a coverage guarantee for this aggregated prediction set.

Proposition 4.1. *Let $\widehat{C}^{split}_{\alpha,1}(X_{n+1}), \ldots, \widehat{C}^{split}_{\alpha,K}(X_{n+1})$ be K prediction sets defined by p-value functions $\widehat{p}_1, \ldots, \widehat{p}_K$ (2) on X_{n+1}, where $1-\alpha$ coverage (3) holds for each set $k \in [K]$. Then, the prediction set $\widehat{C}^{avg}_{\alpha}(X_{n+1})$ (5) from thresholding the weighted average p-value function (4) gives the coverage guarantee*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}^{avg}_{\alpha}(X_{n+1})\right\} \geq 1 - \min\left\{\frac{1}{v}, 2\right\}\alpha \quad \forall \alpha \in (0,1)$$

where $v = \max\{v_1, v_2, \ldots, v_K\}$ is the largest weight assigned to any of the p-values.

Weighted aggregation provides more flexible coverage guarantees In a sense, this result provides a spectrum of coverage guarantees based on the weight distribution among the models. When one predictor dominates (i.e., $v > 1/2$), the guarantee improves beyond the standard $1-2\alpha$ of typical aggregation methods, allowing the overall guarantee to approach that of the most influential predictor. In the extreme case, when all the weight is assigned to a single model ($v = 1$), we recover the standard split conformal guarantee of $1-\alpha$ coverage for that model. Thus, we can interpolate between the standard $1-2\alpha$ guarantee of combined models and the $1-\alpha$ guarantee of individual models, with the weights controlling the trade-off. This improves the coverage guarantee for asymmetric aggregation, and also opens up the utility of asymmetric generalizations of the many established symmetric

aggregation methods (e.g. set-weighted versions of cross-conformal, CV+, etc.)

Since the work of Vovk and Wang holds for arbitrarily dependent p-values, this method is robust across a wide range of scenarios. In practice, independent weights allow users to incorporate prior knowledge about the relative quality of different predictors (Vovk and Wang, 2020; Gasparin and Ramdas, 2024a). The weighting can reflect, for example, expert insights on which of K models should be prioritized as being more reliable.

From p-values to prediction sets The work of Vovk and Wang is central to our result, allowing us to generalize and improve upon existing conformal guarantees. Still, despite its broad applicability, their work remains underexplored in conformal literature, and the weight-dependent coverage result has not yet been applied to conformal prediction sets.[1] We attribute this oversight to several factors.

Vovk and Wang frame their work in terms of merging functions and p-values, without reference to prediction sets or conformal prediction. As a result, subsequent research has similarly focused on p-values and related test statistics, with limited connection to conformal literature. Meanwhile, conformal literature typically uses a quantile-based construction of prediction sets, rather than a p-value construction, and work in conformal prediction set aggregation typically operates directly on the sets, rather than working with the associated p-value functions. These trends may contribute to the pattern where p-value results are not always propagated to the wider conformal prediction community. Our result shows that the p-value presentation offers unique benefits to prediction set aggregation, and we hope that our work may encourage renewed interest in the connection between p-values and conformal prediction.

5 DATA-DEPENDENT WEIGHTS

In many practical settings, model weights are determined by the observed data rather than fixed in advance. Such data-dependent weights naturally adapt the influence of each model to the characteristics of the input, making them central to applications like ensemble learning and mixture-of-experts. This adaptivity, however, creates a technical challenge: once the weights depend on the data, they also depend on the associated p-variables, and the theory of Vovk and Wang (2020) no longer applies. Nevertheless, we can build on their approach to develop a method that allows us to recover a valid coverage guarantee, even when the weights are data-dependent.

The key idea is to directly use the definition of a p-variable (1). The sum of weighted p-variables, where the weights are dependent on the p-variables, is not necessarily a p-variable. However, the weighted sum can *become* a p-variable by a linear transformation that both satisfies the definition of a p-variable and preserves the proportions of the weights.

Let the p-variables of the predictors be $P_1, \ldots, P_K \in \mathcal{U}$, where $\mathcal{U}$ is the set of all uniformly distributed random variables. The weights are given by a random vector $\mathbf{W} = (W_1, \ldots, W_K)$ in the $(K-1)$-dimensional simplex $\Delta_{K-1} := \{\mathbf{w} = (w_1, \ldots, w_K) \in [0,1]^K : w_1 + \cdots + w_K = 1\}$ depending on the data $\{(X_i, Y_i)\}_{i\in[n]} \cup \{X_{n+1}\}$. We define the weighted average function

$$p_{\text{all}}(x, y; \mathbf{w}) = \sum_k w_k \widehat{p}_k(x, y), \tag{6}$$

giving associated random variable $P_{\text{all}} := p_{\text{all}}(X, Y; \mathbf{W})$ with distribution $F_{P_{\text{all}}}$. Then, we propose to learn the scalar

$$\begin{aligned} m^* &= \inf\{m \in \mathbb{R}^+ : \mathbb{P}\{mP_{\text{all}} \le \alpha\} \le \alpha \quad \forall \alpha \in (0,1)\} \\ &= \inf\{m \in \mathbb{R}^+ : F_{P_{\text{all}}}(\alpha/m) \le \alpha \quad \forall \alpha \in (0,1)\} \\ &= \sup_{\delta \in (0,\infty)} \frac{F_{P_{\text{all}}}(\delta)}{\delta}. \end{aligned} \tag{7}$$

(See §A.4 for a derivation of the supremum expression.) The scaling defined in (7) transforms the weighted average P_{all} into a valid p-variable, recovering a coverage guarantee. We formalize this in the following result.

Proposition 5.1 (Infinite-sample guarantee)**.** *Let $\widehat{C}^{split}_{\alpha,1}(X_{n+1}), \ldots, \widehat{C}^{split}_{\alpha,K}(X_{n+1})$ be K prediction sets defined by p-value functions $\widehat{p}_1, \ldots, \widehat{p}_K$ (2) on X_{n+1} corresponding to p-variables $P_1, \ldots, P_K$. Suppose $1-\alpha$ coverage (3) holds for each set $k \in [K]$. Let $(W_1, \ldots, W_K)$ be a random vector in Δ_{K-1} depending on $\{(X_i, Y_i)\}_{i\in[n]} \cup \{X_{n+1}\}$, and let p_{all} be the weighted average (6), with random variable P_{all}.*

The prediction set from thresholding p_{all} is $\widehat{C}^{unscaled}_{\alpha}(X_{n+1}) = \{y : p_{all}(X_{n+1}, y) > \alpha\}$, and it satisfies the coverage guarantee

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{unscaled}_{\alpha}(X_{n+1})\} \ge 1 - m^*\alpha \quad \forall \alpha \in (0,1).$$

The prediction set from thresholding $m^* p_{all}$ is $\widehat{C}^{scaled}_{\alpha}(X_{n+1}) = \{y : m^* p_{all}(X_{n+1}, y) > \alpha\}$, and it satisfies the coverage guarantee

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{scaled}_{\alpha}(X_{n+1})\} \ge 1 - \alpha \quad \forall \alpha \in (0,1).$$

Here, m^* is the correction factor defined in (7).

[1] To the best of our knowledge, the works of Gasparin and Ramdas (Gasparin and Ramdas, 2024a,b) are the only ones so far to use asymmetric set aggregation, and they derive the standard $1-2\alpha$ guarantee of symmetric aggregation methods.

Algorithm 1 Constructing a weighted aggregate of prediction sets with valid coverage

Input: Data with indices $\{1, \ldots, n\} = S_{\text{cal}} \cup S_{\text{merge}}$, K predictors from potentially overlapping datasets, K weights which may depend on data (e.g. learned router weights), test example X_{n+1}.

Output: Prediction set around the test example $\widehat{C}_\alpha^{\text{scaled}}(X_{n+1})$, with coverage of at least $1 - (\alpha + \epsilon + \delta)$.

Step 1. Derive the p-value function p_{all}.

1.1. Using calibration set S_{cal}, derive p-value functions $\widehat{p}_k$ for each of the predictors (2).

1.2. Compute the aggregated p-value function $p_{\text{all}} = \sum_{k=1}^{K} w_k \, \widehat{p}_k$, using the (potentially data-dependent) weights w_k.

Step 2. Compute the correction factor $\widehat{m}^*$.

2.1. Using the points of merging set S_{merge} with the function p_{all}, get samples of $\widehat{F}_{P_{\text{all}}}$.

2.2. Derive the correction factor $\widehat{m}^*$ from the samples of $\widehat{F}_{P_{\text{all}}}$ (8).

Step 3. For test example X_{n+1}, construct the prediction set

$$\widehat{C}_\alpha^{\text{scaled}}(X_{n+1}) = \{y : \widehat{m}^* p_{\text{all}}(X_{n+1}, y) > \alpha\}.$$

Computing m^* with a merging set By definition, m^* is the minimal scaling factor that makes P_{all} a valid p-variable, and its value is determined by the CDF $F_{P_{\text{all}}}$. Because $F_{P_{\text{all}}}$ is a population quantity that is not accessible in practice, we cannot evaluate m^* directly. Instead, we define a computable proxy $\widehat{m}^*$ using an empirical CDF constructed from a designated *merging set* S_{merge}[2].

Typically, the empirical CDF for P_{all} computed from S_{merge} is given by

$$\widehat{F}_{P_{\text{all}}}(\alpha) = \frac{1}{|S_{\text{merge}}|} \sum_{i \in S_{\text{merge}}} \mathbb{1}\{p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}) \leq \alpha\}.$$

The corresponding empirical correction factor is

$$\widehat{m}^* = \max_{i \in S_{\text{merge}}} \frac{\widehat{F}_{P_{\text{all}}}(p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}))}{p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)})}, \tag{8}$$

following (7). (For more detail on why (8) is equivalent to (7) for an empirical CDF, see §A.5.) The construction of $\widehat{m}^*$ yields the following finite-sample guarantee:

[2]So named because S_{merge} is used to learn the correction that makes the merging function (Vovk and Wang, 2020) for the weighted average empirically precise.

Proposition 5.2 (Finite-sample guarantee)**.** *Under the same assumptions as Proposition 5.1, fix a user-chosen failure probability $\delta \in (0, 1)$ and set*

$$\varepsilon = \sqrt{\frac{\log(2/\delta)}{2|S_{merge}|}}.$$

The prediction set from thresholding p_{all} is $\widehat{C}_\alpha^{unscaled}(X_{n+1}) = \{y : p_{all}(X_{n+1}, y) > \alpha\}$, and it satisfies the coverage guarantee

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_\alpha^{unscaled}(X_{n+1})\} \geq 1 - \big(\alpha \mathbb{E}[\widehat{m}^*] + \varepsilon + \delta\big) \quad \forall \alpha \in (0, 1).$$

The prediction set from thresholding $\widehat{m}^* p_{all}$ is $\widehat{C}_\alpha^{scaled}(X_{n+1}) = \{y : \widehat{m}^* p_{all}(X_{n+1}, y) > \alpha\}$, and it satisfies the coverage guarantee

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_\alpha^{scaled}(X_{n+1})\} \geq 1 - (\alpha + \varepsilon + \delta) \quad \forall \alpha \in (0, 1).$$

Here, $\widehat{m}^*$ is the empirical correction factor defined in (8).

The two formulations in Propositions 5.1 and 5.2 demonstrate the complementary roles of m^*. In the unscaled sets, m^* scales the coverage lower bound in parallel with Proposition 4.1, while in the scaled sets it rescales the p-variable to yield a standard $1 - \alpha$ style guarantee. For consistency with other conformal methods, we use the scaled formulation in our experiments.

An important practical question is how the quality of the correction $\widehat{m}^*$ depends on the merging set S_{merge}. In §B.1.1, we study this dependence by varying $|S_{\text{merge}}|$, and we show that even a modest number of samples ($<$ 200) is sufficient for accurate coverage. We summarize our full procedure in Algorithm 1.

Data-dependent weights give a form of conditional coverage Data-dependent weights allow the influence of each model to be adjusted based on how well it performs for a specific data point. This approach enables the construction of aggregated prediction sets that are tailored to the characteristics of the given data, which can be viewed as a form of *locally conditional coverage*. To be clear, true X-conditional coverage—where the coverage guarantee is conditioned on the current input—is impossible without additional distributional assumptions (Lei and Wasserman, 2014; Vovk, 2012). However, we demonstrate in §6 that data-dependent weights allow us to create aggregated prediction sets with data-adaptive coverage, which can greatly improve conditional validity in practice.

Achieving tighter guarantees Our scaling correction factor m^* provides a coverage guarantee that holds for all significance levels $\alpha \in (0, 1)$. However, guarantees on coverage for all α can lead to overly conservative prediction sets, which may be unnecessarily

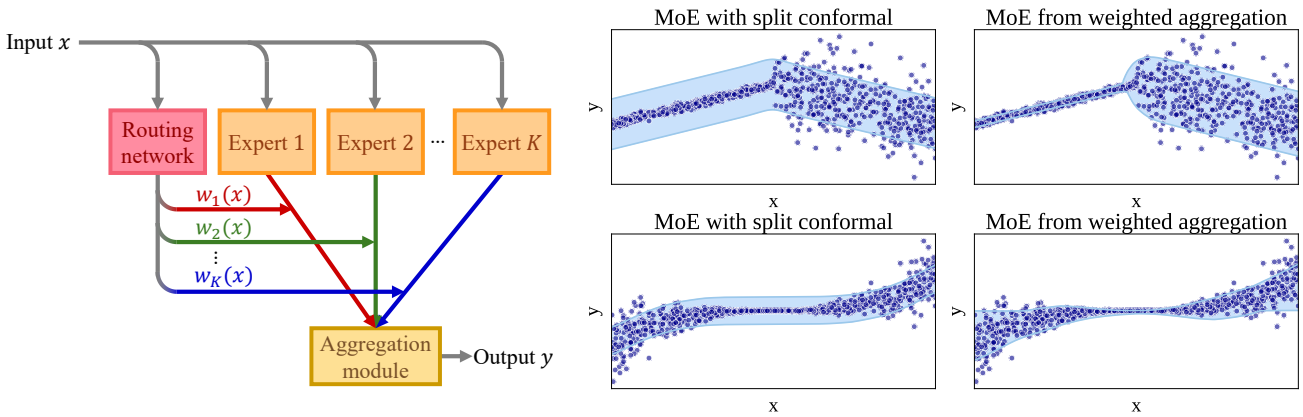

Figure 2: **Left:** Network diagram for MoE. For traditional MoE, the aggregation module takes a weighted sum of the outputs from each expert. To learn weight-dependent prediction sets, we instead propose to combine the prediction sets of each expert by weighted p-value. **Right, top row:** Comparison of split conformal prediction sets with those learned from weighted aggregation. Weighted aggregation allows overall coverage to follow the coverage of the dominant expert, rather than remain purely marginal. **Right, bottom row:** Another comparison of split conformal with weighted aggregation, with the latter showing local coverage with smooth transitions.

restrictive in practice when we do not need guarantees for every possible significance level. Thus, we propose two alternatives: for a specific significance level α', we can learn a correction factor

$$m^{\dagger} = \inf\{m \in \mathbb{R}^+ : \mathbb{P}\{mP_{\text{all}} \leq \alpha\} \leq \alpha \quad \forall \alpha \in (0, \alpha']\}, \tag{9}$$

or the even stricter

$$m^{\ddagger} = \inf\{m \in \mathbb{R}^+ : \mathbb{P}\{mP_{\text{all}} \leq \alpha'\} \leq \alpha'\}. \tag{10}$$

(Note that (9) is not without precedent: for the $1 - 2\alpha$ coverage guarantees of aggregation methods like cross-conformal and jackknife+, $\alpha' = 0.5$ is the highest significance level of interest.)

6 APPLICATION: MIXTURE-OF-EXPERTS

Mixture-of-experts (MoE) is a machine learning framework designed to combine the predictions of multiple specialized models, called experts (Jacobs et al., 1991). Each expert in an MoE focuses on unique aspects of the problem by learning different representations of the input data. A central component of this framework is the routing network, which determines how to combine the experts' outputs. Specifically, for an input x, the routing network of a traditional (soft) MoE assigns weights $W_k(x)$ to each expert output $f_k(x)$, producing the final prediction as a weighted sum

$$f(x) = \sum_k W_k(x) f_k(x). \tag{11}$$

By learning how to route different inputs to the most appropriate experts, the routing network implicitly conditions the model's final prediction on the combination of experts that fits the given input. In this way, the routing network can be viewed as learning a form of conditional relationship between the input features and the expertise of each model.

The routing network's ability to learn data-dependent weights for each expert makes MoE a natural setting for applying our method of aggregating prediction sets by weighted p-values. We collect the routing weights into the simplex-valued vector $W(x) := (W_1(x), \ldots, W_K(x))$; then, for a new input X_{n+1}, the routing network outputs the weight vector $\mathbf{W}^{(n+1)} := W(X_{n+1})$, whose kth component we denote $W_k^{(n+1)}$. Let $\widehat{p}_k$ denote the p-value function of the kth expert. The weighted aggregate p-value function for label y is

$$p_{\text{all}}(X_{n+1}, y; \mathbf{W}^{(n+1)}) = \sum_k W_k^{(n+1)} \widehat{p}_k(X_{n+1}, y).$$

At significance level α, we form the MoE prediction set by thresholding the corrected p-variable $\widehat{m}^* p_{\text{all}}$, as established in Proposition 5.2:

$$\widehat{C}_{\alpha}^{\text{MoE}}(X_{n+1}) = \{y \in \mathcal{Y} : \widehat{m}^* p_{\text{all}}(X_{n+1}, y; \mathbf{W}^{(n+1)}) > \alpha\}.$$

Intuitively, the routing weights adapt the contribution of each expert's p-value to the input so that experts with higher predictive relevance for X_{n+1} have greater influence. This pairs naturally with weighted p-value aggregation for two reasons: (1) averaging p-values retains graded evidence from each expert, so that the weight-aggregated p-value faithfully reflects expert disagreement; (2) although data-dependent weights generally break p-value validity, the p-value formulation admits a scaling correction (7) that restores finite-sample validity, even though the routing weights are learned from the data. We expand on both these points in §C.1,

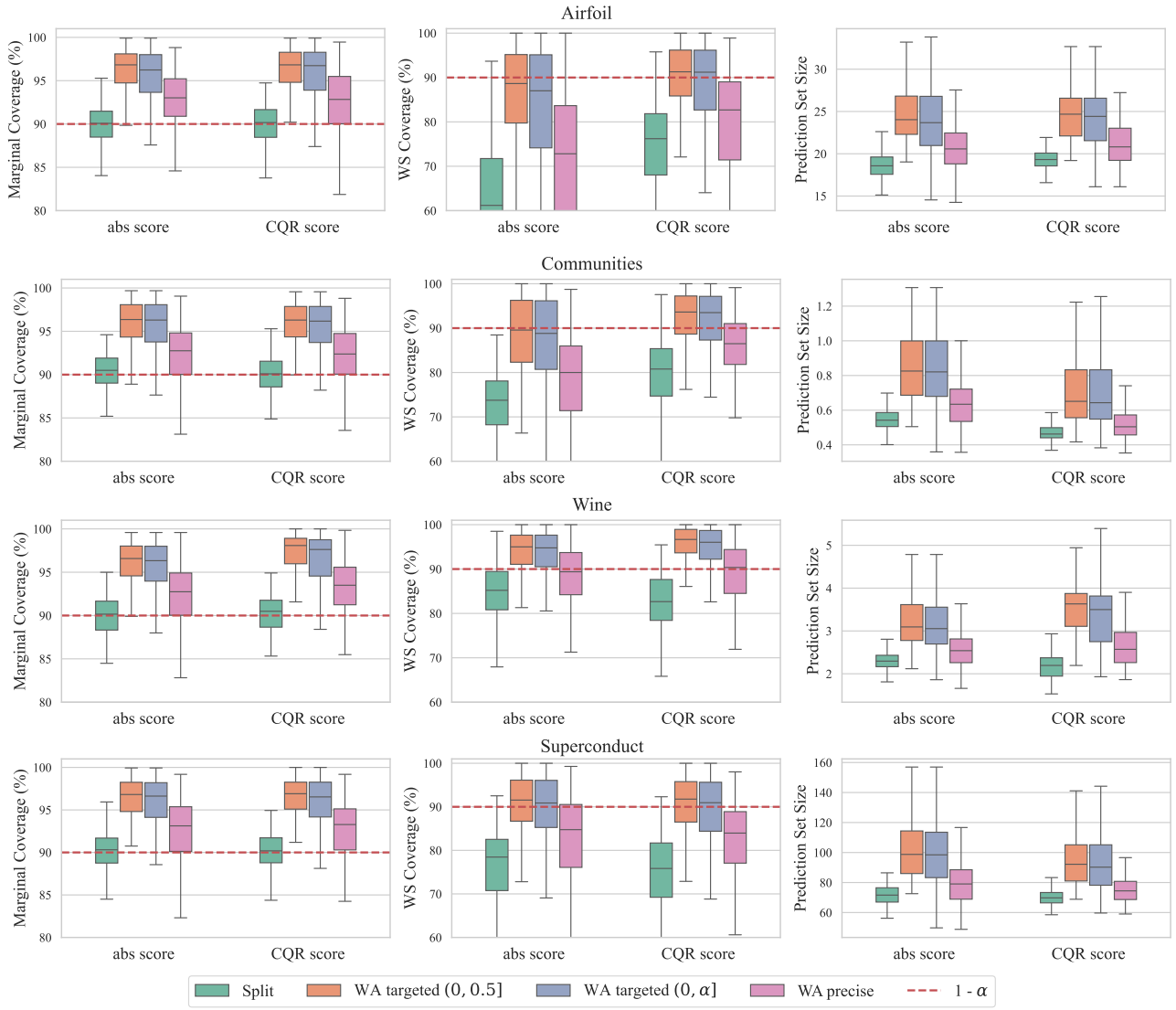

Figure 3: Local validity experiments comparing split conformal to weighted aggregation using absolute residual scores and CQR scores. Each row corresponds to a dataset, with plots for marginal coverage, WS coverage, and prediction set size from left to right. The weighted aggregation methods consistently improves WS coverage across datasets, in most cases meeting target coverage on the WS slab. Split conformal is less conservative marginally, but undercovers on the WS slab; using CQR scores is not enough to cover the gap.

and we illustrate the resulting gains in local validity in Figure 2 and in the following experiments.

Baselines We refer to our proposed method of aggregating expert p-value functions with learned weights as *weighted aggregation*. As our main baseline, we compare against split conformal with the full MoE predictor, and we refer to this simply as *split conformal* for brevity.

To evaluate local validity, we compare against conformal quantile regression (CQR) (Romano et al., 2019), a widely used locally adaptive method. A strength of our framework is that it complements, rather than competes with, other adaptive methods, allowing weighted aggregation to be layered on top of CQR. Accordingly, we also evaluate a hybrid method that combines weighted aggregation with CQR. We assess performance using marginal coverage, worst-slice (WS) coverage (Romano et al., 2020), and prediction set size.

We consider two practical variants of weighted aggregation:

- *WA targeted* $(0, \alpha']$: coverage is guaranteed for all $\alpha \in (0, \alpha')$ using the $m^\dagger$ correction (9).
- *WA precise* α': coverage is guaranteed for $\alpha = \alpha'$ only, using the $m^\ddagger$ correction (10).

Experiment overview For clarity of exposition, we present our main experimental results on the local validity of our method for real-world regression tasks

in Figure 3, and defer additional ablation studies and further analyses on local validity to the Appendix.

Figure 3 summarizes our main regression results, with CQR as our main adaptive baseline. We also evaluate our method against additional data-adaptive baselines in §B.2.3 and §C.2, and we present results on classification tasks in §B.2.1.

Our primary metric for evaluating local validity is WS coverage. To complement this, we also examine coverage disparities across demographic groups on the Communities dataset (Redmond and Baveja, 2002) in §B.2.2

In Appendix B, we analyze how various factors impact coverage and prediction set size in a series of ablation studies on synthetic data. We investigate: the effects of merging set size $|S_{\text{merge}}|$ (§B.1.1), shared feature information (§B.1.2), and the different weighted aggregation variants (§B.1.3).

The following experiments are conducted with an MoE where the routing network and experts are linear models for simplicity. All results are averaged over 200 trials.

Main regression results In Figure 3, we compare split conformal to weighted aggregation using both absolute residual scores and CQR scores. From the marginal coverage plots (left column), we see that split conformal consistently achieves coverage closest to the nominal $1 - \alpha$ level, while the weighted aggregation methods tend to overcover. However, the WS coverage plots (middle column) expose a well-known limitation of split conformal: it does not maintain coverage across the data, systematically undercovering on the WS slab. CQR scores improve split conformal WS coverage on some datasets (Airfoil and Communities), but they are still insufficient to close the coverage gap.

Weighted aggregation significantly improves WS coverage, coming close to or meeting the nominal level on all datasets. In addition, the gap between marginal and WS coverage is notably smaller for weighted aggregation than for split conformal, indicating that weighted aggregation provides more uniform coverage, even over the challenging regions of the data. These results suggest that weighted aggregation is preferable in applications where local validity is a priority.

The prediction set size plots (right column) illustrate the standard trade-off between coverage and efficiency, where higher-coverage methods like WA targeted tend to produce larger prediction sets than lower-coverage methods like split conformal. According to these results, we suggest WA targeted if maintaining coverage in challenging regions is the primary concern. If both efficiency and local validity are important, WA precise provides a reasonable middle ground—offering improved WS coverage over split conformal, while also maintaining more compact prediction set size.

7 CONCLUSION

The asymmetric (weighted) aggregation of prediction sets is a flexible generalization of standard symmetric aggregation. Intuitively, coverage in this setting should vary with the distribution of weights. The results of Vovk and Wang (2020) allow us to formalize this intuition for conformal prediction, so that for data-independent weights based on expert priors, we obtain improved guarantees when the weights are sufficiently asymmetric. We extend this framework to data-dependent weights (e.g. weights learned from data), enabling adaptive coverage that reflects the observed input. Experiments on WS coverage and demographic subgroups confirm the practical benefits of this extension, showing that weighted aggregation yields more reliable coverage in challenging settings.

Acknowledgments

We thank Josh Gleason and Uthsav Chitra for their thoughtful feedback, and Josh Gleason in particular for many insightful discussions throughout this work. GW and RC are partially supported by a ONR MURI grant N00014-20-1-2781. DP is partially supported by an Amazon AI2AI fellowship. DP, SS, and AL are partially supported by the Gordon and Betty Moore Foundation grant #12128. AL is also supported by an Amazon Research Award.

Reproducibility

Code is available at: `https://github.com/ginawong/weighted_pvalues`. The code contains all the scripts necessary to reproduce the figures in the paper.

References

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024a.

Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G Lee, Vikram Deshpande, Joseph Schwab,

Michael H Lev, Ramon G Gonzalez, Michael S Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7(6):711–718, 2023.

Thomas Grote and Philipp Berens. Uncertainty, evidence, and the integration of machine learning into medical practice. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, 48(1):84–97, 2023.

Sachin Kumar, T Gopi, N Harikeerthana, Munish Kumar Gupta, Vidit Gaur, Grzegorz M Krolczyk, and ChuanSong Wu. Machine learning techniques in additive manufacturing: a state of the art review on design, processes and production control. *Journal of Intelligent Manufacturing*, 34(1):21–55, 2023.

Nathan P Lawrence, Seshu Kumar Damarla, Jong Woo Kim, Aditya Tulsyan, Faraz Amjad, Kai Wang, Benoit Chachuat, Jong Min Lee, Biao Huang, and R Bhushan Gopaluni. Machine learning for industrial sensing and control: A survey and practical perspective. *Control Engineering Practice*, 145:105841, 2024.

Veronika Eyring, William D Collins, Pierre Gentine, Elizabeth A Barnes, Marcelo Barreiro, Tom Beucler, Marc Bocquet, Christopher S Bretherton, Hannah M Christensen, Katherine Dagon, et al. Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, pages 1–13, 2024.

Ching-Yao Lai, Pedram Hassanzadeh, Aditi Sheshadri, Maike Sonnewald, Raffaele Ferrari, and Venkatramani Balaji. Machine learning for climate physics and simulations. *arXiv preprint arXiv:2404.13227*, 2024.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

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

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *Annals of Statistics*, 2021.

Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. *arXiv preprint arXiv:2401.09379*, 2024a.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, pages 1–13, 2021.

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after efficiency-oriented model selection. *arXiv preprint arXiv:2408.07066*, 2024b.

Giovanni Cherubin. Majority vote ensembles of conformal predictors. *Machine Learning*, 108(3):475–488, 2019.

Aldo Solari and Vera Djordjilović. Multi split conformal prediction. *Statistics & Probability Letters*, 184: 109395, 2021.

Matteo Gasparin and Aaditya Ramdas. Conformal online model aggregation. *arXiv preprint arXiv:2403.15527*, 2024b.

Vineeth N Balasubramanian, Shayok Chakraborty, and Sethuraman Panchanathan. Conformal predictions for information fusion: A comparative study of p-value combination methods. *Annals of Mathematics and Artificial Intelligence*, 74(1):45–65, 2015.

Cyrus J DiCiccio, Thomas J DiCiccio, and Joseph P Romano. Exact tests via multiple data splitting. *Statistics & Probability Letters*, 166:108865, 2020.

Bernhard Rüger. Das maximale signifikanzniveau des tests:"lehne h o ab, wenn k unter n gegebenen tests zur ablehnung führen". *Metrika*, 25:171–178, 1978.

Gerhard Hommel. Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal*, 25(5):423–430, 1983.

Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.

David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia Strachan, Ali Taylan Cemgil, and Arnaud Doucet. Conformal prediction under ambiguous ground truth. *arXiv preprint arXiv:2307.09302*, 2023.

Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

Xiaoyang Wu, Yuyang Huo, and Changliang Zou. Multi-split conformal prediction via cauchy aggregation. *Stat*, 12(1):e522, 2022.

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

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

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33: 3581–3591, 2020.

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

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Rui Luo and Zhixin Zhou. Conformity score averaging for classification. In *Forty-second International Conference on Machine Learning*, 2025.

Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

Meichen Liu, Lei Ding, Dengdeng Yu, Wulong Liu, Linglong Kong, and Bei Jiang. Conformalized fairness via quantile regression. *Advances in Neural Information Processing Systems*, 35:11561–11572, 2022.

Vilde Jensen, Filippo Maria Bianchi, and Stian Normann Anfinsen. Ensemble conformalized quantile regression for probabilistic time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9014–9025, 2022.

Raphael Rossellini, Rina Foygel Barber, and Rebecca Willett. Integrating uncertainty awareness into conformalized quantile regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1548. PMLR, 2024.

Leying Guan. Conformal prediction with localization. *arXiv preprint arXiv:1908.08558*, 2019.

Xiaonan Nie, Xupeng Miao, Shijie Cao, Lingxiao Ma, Qibin Liu, Jilong Xue, Youshan Miao, Yi Liu, Zhi Yang, and Bin Cui. Evomoe: An evolutional mixture-of-experts training framework via dense-to-sparse gate. *arXiv preprint arXiv:2112.14397*, 2021.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Alexandre Decan. portion: Python data structure and operations for intervals, 2018. URL `https://github.com/AlexandreDecan/portion`.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The uci machine learning repository. *https://archive.ics.uci.edu*, 2010.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **[Yes]**
   We establish our theoretical contributions in §4 and §5, explicitly stating our setting and assumptions. We also discuss how to implement data-dependent weights in practice in Algorithm 1.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **[Yes]**
   Please see §D for a discussion on the computational complexity of our method.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **[Yes]**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **[Yes]**
   All theoretical results include a full list of assumptions at the start of the result.
   (b) Complete proofs of all theoretical results. **[Yes]**
   Please see §A for complete proofs to all theoretical results, with any additional theoretical details that may be of interest.
   (c) Clear explanations of any assumptions. **[Yes]**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **[Yes]**
   The code to reproduce all experiments and figures is in the supplemental material.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**]
We tried to include all important training details in §6 in "Baselines" and "Experiment overview". For further details, please see §E.

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Not Applicable**]
All experiments can be run on a consumer-grade CPU.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator if your work uses existing assets. [**Yes**]

   (b) The license information of the assets, if applicable. [**Yes**]
   License information is included in the source code, in the supplementary.

   (c) New assets either in the supplemental material or as a URL, if applicable. [**Yes**]
   The code to reproduce all experiments and figures is in the supplemental material.

   (d) Information about consent from data providers/curators. [**Not Applicable**]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

A Proofs and additional theoretical details

A.1 Proof of Proposition 4.1

Following the notation of Vovk and Wang (2020), we define the merging function

$$M_{1,\mathbf{v}}(p_1, \ldots, p_K) = (v_1 p_1 + \cdots + v_K p_K) \quad \text{where} \sum_{k=1}^{K} v_k = 1.$$

Recall that we use p-variables to represent our p-value functions applied to data; that is, $P_k = \widehat{p}_k(X, Y)$. Then $\bar{P} = M_{1,\mathbf{v}}(P_1, \ldots, P_K)$. By Proposition 9 of Vovk and Wang (2020), $A_{\mathbf{v}} M_{1,\mathbf{v}}$ is a precise merging function, where $A_{\mathbf{v}} = \min\left\{\frac{1}{v}, 2\right\}$ and $v = \max\{v_1, ..., v_K\}$. Thus, $A_{\mathbf{v}}\bar{P}$ is a p-variable, and $\mathbb{P}\left\{A_{\mathbf{v}}\bar{P} \leq \alpha\right\} \leq \alpha$, or

$$\mathbb{P}\left\{\bar{P} \leq \alpha\right\} \leq A_{\mathbf{v}}\alpha = \min\left\{\frac{1}{v}, 2\right\}\alpha.$$

Under exchangeability, the prediction set

$$\widehat{C}_{\alpha}^{\text{avg}}(X_{n+1}) = \{y \in \mathcal{Y} : \bar{p}(X_{n+1}, y) > \alpha\}$$

constructed from $\bar{p}$ has the coverage guarantee

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

A.2 Proof of Proposition 5.1

We define m^* to be the smallest positive value such that $m^* P_{\text{all}}$ is a p-variable, i.e.

$$\mathbb{P}\left\{m^* P_{\text{all}} \leq \alpha\right\} \leq \alpha, \tag{12}$$

or, equivalently,

$$\mathbb{P}\left\{m^* P_{\text{all}} > \alpha\right\} \geq 1 - \alpha.$$

Recall that $P_{\text{all}} := p_{\text{all}}(X, Y, \mathbf{W})$. Then under exchangeability, the prediction set

$$\widehat{C}_{\alpha}^{\text{scaled}}(X_{n+1}) = \left\{y \in \mathcal{Y} : m^* p_{\text{all}}(X_{n+1}, y; \mathbf{W}^{(n+1)}) > \alpha\right\}$$

has coverage

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

Alternatively, we can rearrange (12) to be

$$\mathbb{P}\left\{P_{\text{all}} \leq \frac{\alpha}{m^*}\right\} \leq \alpha;$$

if we define $\alpha' = \alpha/m^*$, this becomes

$$\mathbb{P}\left\{P_{\text{all}} \leq \alpha'\right\} \leq m^*\alpha'$$

or

$$\mathbb{P}\left\{P_{\text{all}} > \alpha'\right\} \geq 1 - m^*\alpha'.$$

Under exchangeability, the prediction set

$$\widehat{C}_{\alpha}^{\text{unscaled}}(X_{n+1}) = \left\{y \in \mathcal{Y} : p_{\text{all}}(X_{n+1}, y; \mathbf{W}^{(n+1)}) > \alpha\right\}$$

satisfies

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

Note that other transformations of P_{all} can also yield valid p-variables while preserving the relative proportion of weights. For example, if P_{all} is shifted rather than scaled, then the condition

$$\mathbb{P}\left\{m' + P_{\text{all}} \leq \alpha\right\} \leq \alpha$$

also leads to coverage of at least $1 - \alpha$ for the scaled set, and at least $1 - (m' + \alpha)$ for the unscaled set. We focus on the correction factor m^* because it aligns with the framework of Vovk and Wang (2020) and works well in practice.

A.3 Proof of Proposition 5.2

Let $\mathcal{G}$ be the event that the worst-case distance between the true CDF $F_{P_{\text{all}}}$ and the empirical CDF $\widehat{F}_{P_{\text{all}}}$ is bounded by some maximum allowable deviation; that is,

$$\mathcal{G} = \left\{ \sup_{x \in \mathbb{R}} \left| \widehat{F}_{P_{\text{all}}}(x) - F_{P_{\text{all}}}(x) \right| \leq \varepsilon \right\} \quad \text{where } \varepsilon = \sqrt{\frac{\log(2/\delta)}{2|S_{\text{merge}}|}}$$

for some user-chosen failure probability $\delta \in (0, 1)$. By the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality,

$$\mathbb{P}(\mathcal{G}) \geq 1 - \delta.$$

Now, suppose we have a fixed merging set S_{merge} such that $\mathcal{G}$ holds. By definition of a CDF,

$$\mathbb{P}\left\{ \widehat{m}^* P_{\text{all}} \leq \alpha \mid S_{\text{merge}} \right\} = F_{P_{\text{all}}}\left(\frac{\alpha}{\widehat{m}^*} \right), \tag{13}$$

and under the same fixed S_{merge},

$$F_{P_{\text{all}}}(x) \leq \widehat{F}_{P_{\text{all}}}(x) + \varepsilon.$$

We take $x = \alpha/\widehat{m}^*$; this allows us to express the bound as

$$F_{P_{\text{all}}}\left(\frac{\alpha}{\widehat{m}^*} \right) \leq \widehat{F}_{P_{\text{all}}}\left(\frac{\alpha}{\widehat{m}^*} \right) + \varepsilon. \tag{14}$$

By construction of $\widehat{m}^*$ (8), we know that $\widehat{F}_{P_{\text{all}}}(x)/x \leq \widehat{m}^*$ for all x. Substituting again $x = \alpha/\widehat{m}^*$, we have

$$\widehat{F}_{P_{\text{all}}}\left(\frac{\alpha}{\widehat{m}^*} \right) \leq \widehat{m}^* \frac{\alpha}{\widehat{m}^*} = \alpha. \tag{15}$$

Combining (14) and (15) gives

$$F_{P_{\text{all}}}\left(\frac{\alpha}{\widehat{m}^*} \right) \leq \alpha + \varepsilon.$$

Applying this to (13) gives

$$\mathbb{P}\left\{ \widehat{m}^* P_{\text{all}} \leq \alpha \mid S_{\text{merge}} \right\} \leq \alpha + \varepsilon. \tag{16}$$

Under exchangeability, the prediction set

$$\widehat{C}_\alpha^{\text{scaled}}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \widehat{m}^* p_{\text{all}}(X_{n+1}, y; \mathbf{W}^{(n+1)}) > \alpha \right\}$$

has a conditional *miscoverage* guarantee of

$$\mathbb{P}\left\{ Y_{n+1} \notin \widehat{C}_\alpha^{\text{scaled}}(X_{n+1}) \mid S_{\text{merge}} \right\} \leq \alpha + \varepsilon.$$

Let us denote the miscoverage indicator as

$$\text{MC} = \mathbb{1}\left\{ Y_{n+1} \notin \widehat{C}_\alpha^{\text{scaled}}(X_{n+1}) \right\}.$$

The indicator random variable allows us to easily switch between probability and expectation in order to marginalize over S_{merge}:

$$\begin{aligned}
\mathbb{P}\{\text{MC} = 1\} &= \mathbb{E}[\text{MC}] \\
&= \mathbb{E}[\mathbb{E}[\text{MC} \mid S_{\text{merge}}]] \\
&= \mathbb{E}[\mathbb{P}\{\text{MC} = 1 \mid S_{\text{merge}}\}] \\
&\leq (\alpha + \varepsilon)\,\mathbb{P}(\mathcal{G}) + 1 \cdot \mathbb{P}(\mathcal{G}^c) \\
&\leq \alpha + \varepsilon + \delta.
\end{aligned}$$

This gives the coverage guarantee

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

To derive the guarantee for the *unscaled* prediction set, we apply a change of variable to (16) to get

$$\mathbb{P}\left\{P_{\text{all}} \leq \alpha' \mid S_{\text{merge}}\right\} \leq \widehat{m}^* \alpha' + \varepsilon.$$

The remainder of the proof follows the same conditioning and marginalization argument, with the substitution of $\widehat{m}^*\alpha$ for α carried through. To be explicit, the prediction set

$$\widehat{C}_\alpha^{\text{unscaled}}(X_{n+1}) = \left\{y \in \mathcal{Y} : p_{\text{all}}(X_{n+1}, y; \mathbf{W}^{(n+1)}) > \alpha\right\}$$

has a conditional miscoverage guarantee of

$$\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}_\alpha^{\text{unscaled}}(X_{n+1}) \mid S_{\text{merge}}\right\} \leq \widehat{m}^* \alpha + \varepsilon.$$

Then miscoverage indicator $\text{MC}' = \mathbb{1}\left\{Y_{n+1} \notin \widehat{C}_\alpha^{\text{unscaled}}(X_{n+1})\right\}$ can help us marginalize over S_{merge}:

$$\begin{aligned}\mathbb{P}\left\{\text{MC}' = 1\right\} &= \mathbb{E}\left[\mathbb{P}\left\{\text{MC}' = 1 \mid S_{\text{merge}}\right\}\right] \\ &\leq \left(\alpha \mathbb{E}[\widehat{m}^*] + \varepsilon\right) \mathbb{P}(\mathcal{G}) + 1 \cdot \mathbb{P}(\mathcal{G}^c) \\ &\leq \alpha \mathbb{E}[\widehat{m}^*] + \varepsilon + \delta.\end{aligned}$$

This gives the guarantee

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha^{\text{unscaled}}(X_{n+1})\right\} \geq 1 - \left(\alpha \mathbb{E}[\widehat{m}^*] + \varepsilon + \delta\right).$$

Note that the same argument applies directly if we replace the empirical CDF $\widehat{F}_{P_{\text{all}}}$ by the conservative version $\widehat{F}_{P_{\text{all}}}^{\text{cons}}$ (18) because the two differ by at most $(|S_{\text{merge}}| + 1)^{-1}$, a deterministic offset that can simply be absorbed into the DKW tolerance by replacing ε with $\varepsilon + (|S_{\text{merge}}| + 1)^{-1}$.

A.4 More detailed derivation of m^*

For some distribution function $F_{P_{\text{all}}}$, we define

$$m^* = \inf\left\{m > 0 : F_{P_{\text{all}}}(\alpha/m) \leq \alpha \text{ for all } \alpha \in (0,1)\right\}, \qquad c = \sup_{\delta > 0} \frac{F_{P_{\text{all}}}(\delta)}{\delta}.$$

We aim to prove that $m^* = c$ to establish the equivalence in (7). To aid our proof, we define the feasible set

$$S = \left\{m > 0 : F_{P_{\text{all}}}(\alpha/m) \leq \alpha \text{ for all } \alpha \in (0,1)\right\},$$

where $m^* = \inf(S)$.

We begin by showing that $m^* \leq c$. To this end, consider any $m > c$. By definition of c,

$$F_{P_{\text{all}}}(\delta) \leq c\,\delta < m\,\delta \quad \forall \delta > 0.$$

Pick $\delta = \alpha/m$ for any $\alpha \in (0,1)$ so that $\delta \in (0, 1/m)$. Substituting yields

$$F_{P_{\text{all}}}(\alpha/m) < m(\alpha/m) = \alpha.$$

This shows that every $m > c$ satisfies the feasibility condition, so every $m > c$ is in S; it follows that $S \supset (c, \infty)$. Then, as the infimum of S, $m^* \leq c$.

To show that $m^* \geq c$, consider any feasible $m > 0$, i.e. assume

$$F_{P_{\text{all}}}(\alpha/m) \leq \alpha \quad \forall \alpha \in (0,1).$$

For any $\delta > 0$, there are two cases.

1. If $\delta \geq 1/m$, since $F_{P_{\text{all}}}(\delta) \leq 1$,
$$F_{P_{\text{all}}}(\delta)/\delta \leq 1/\delta \leq m.$$

2. If $\delta \in (0, 1/m)$, let us select $\alpha = m\delta \in (0, 1)$. By feasibility of m,

$$F_{P_{\text{all}}}(\delta) = F_{P_{\text{all}}}(\alpha/m) \leq \alpha = m\delta,$$

or

$$F_{P_{\text{all}}}(\delta)/\delta \leq m.$$

The two cases above establish $m \geq c$ for every $m \in S$, or that $S \subset (c, \infty)$. Thus, c is a lower bound of S and $m^* \geq c$.

A.5 Computing $\widehat{m}^*$ from an empirical CDF

For random variable $P_{\text{all}} = p_{\text{all}}(X, Y; \mathbf{W})$, a merging correction factor λ is a positive scalar that ensures that a λP_{all} is a valid p-value. That is,

$$\lambda \in \{m > 0 : \mathbb{P}\{mP_{\text{all}} \leq \alpha\} \leq \alpha \ \forall \alpha \in (0, 1)\} = \{m > 0 : F_{P_{\text{all}}}(\alpha/m) \leq \alpha \ \forall \alpha \in (0, 1)\}.$$

Scaling by a merging correction factor gives us the guarantee

$$\mathbb{P}\{P_{\text{all}} > \alpha\} = 1 - \lambda\alpha.$$

To achieve the tightest guarantee (with application to all $\alpha \in (0, 1)$), we define the minimal merging correction factor m^* to be

$$m^* = \inf\{m > 0 : F_{P_{\text{all}}}(\alpha/m) \leq \alpha \ \forall \alpha \in (0, 1)\} = \sup_{\delta > 0} \frac{F_{P_{\text{all}}}(\delta)}{\delta}. \tag{17}$$

(Equation (17) is a restatement of (7) from the main body; we replicate it here for easy reference.)

In practice, we compute the minimal merging factor $\widehat{m}^*$ by first constructing the empirical CDF of P_{all} over the merging set

$$\widehat{F}_{P_{\text{all}}}(\alpha) = \frac{\sum_{i \in S_{\text{merge}}} \mathbb{1}\left\{p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}) \leq \alpha\right\}}{|S_{\text{merge}}|},$$

or its conservative version $\widehat{F}^{\text{cons}}_{P_{\text{all}}}$ (18). For the sake of conciseness, let

$$\widehat{F}_i := \widehat{F}_{P_{\text{all}}}\left(p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)})\right).$$

Since $\widehat{F}_{P_{\text{all}}}$ is a right-continuous step function that only jumps at the observed values $\{p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}) : i \in S_{\text{merge}}\}$, the supremum in (17) is attained at one of these points. In fact, for any δ not equal to one of these values, shifting δ slightly to the right (toward the next jump) increases the denominator without changing the numerator, decreasing the ratio. For this reason, it suffices to take the maximum over the observed p-values, yielding the empirical merging factor

$$\widehat{m}^* = \max_{\substack{i \in S_{\text{merge}} \\ p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}) > 0}} \frac{\widehat{F}_i}{p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)})}.$$

In some cases, we are only interested in coverage above a certain significance level α as in (9). Then we only need to find the first point on the empirical CDF that surpasses α, or

$$\bar{\alpha} = \min_{\substack{i \in S_{\text{merge}} \\ \widehat{F}_i \geq \alpha}} \widehat{F}_i,$$

and then compute the scaling factor to be

$$\widehat{m}^\dagger = \max_{\substack{i \in S_{\text{merge}} \\ p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}) > 0 \\ \widehat{F}_i \leq \bar{\alpha}}} \frac{\widehat{F}_i}{p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)})}.$$

B Additional experiments

B.1 Synthetic data

We generate a simple homoskedastic dataset to simulate a regression task, where each input is a 16-dimensional Gaussian with random parameters, and the label is the sum of the different dimensions with additive noise (see §E.1 for more detail). We use this dataset to investigate how various factors impact the coverage and interval width of our prediction sets.

B.1.1 Coverage and size of prediction sets improve with larger $|S_{\text{merge}}|$

The merging set S_{merge} allows us to construct an empirical CDF for P_{all}. To improve stability in finite samples, we apply a conservative correction to the typical formula for the empirical CDF, and use this conservative CDF (18) to compute the correction factor $\widehat{m}^*$ (8). Our conservative CDF tends to be overly conservative when $|S_{\text{merge}}|$ is small, leading to overcoverage. However, as $|S_{\text{merge}}|$ grows and the empirical CDF approaches the true CDF, coverage becomes closer to nominal and the prediction sets become less conservative.

Figure 4 illustrates how the size of the merging set $|S_{\text{merge}}|$ impacts mean coverage. On the left plot, we use the $m^\dagger$ correction (9) for *WA targeted* (that is, $(0, \alpha)$ aggregation), and we compare different ways of assigning features to experts. On the right plot, we use a non-overlapping feature assignment, and we compare the different variants of weighted aggregation. Our results show that larger $|S_{\text{merge}}|$ leads to tighter coverage, with 160 samples being sufficient to achieve $< 3\%$ overcoverage for most feature assignment methods. Interestingly, configurations where experts have no overlapping features tend to overcover the most. We explore the effect of feature assignment in the next section.

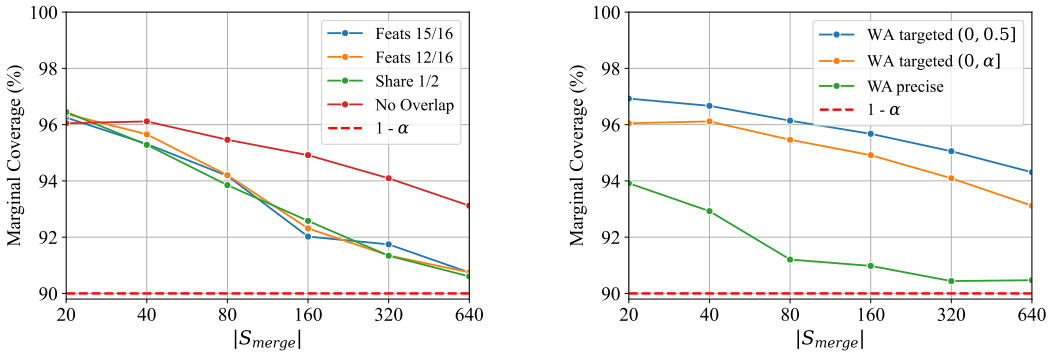

Figure 4: Mean coverage compared to the size of the merging set S_{merge}, with different feature assignments (left) and different weighted aggregation variants (right). Overall, we find that coverage improves as the merging set gets larger for all methods. We note that the merging set does not have to be prohibitively large to produce decent results: for example, most feature assignment methods overcover by only 2% with fewer than 200 samples.

B.1.2 Feature information overlap leads to higher coverage and more efficient prediction sets

To better understand how the allocation of features to experts affects the behavior of MoE weighted aggregation, we define four feature assignment methods (for our MoE of four experts):

- *Features 15/16*: each expert predicts from 15 of the 16 available features.
- *Features 12/16*: each expert predicts from 12 of the 16 available features.
- *Share 1/2*: all experts share 8 of the 16 features and partition the remaining 8 (2 features each).
- *No Overlap*: the experts partition the 16 features (4 features each).

Figure 5 shows how different feature assignment methods affect coverage and prediction set size. Broadly, we observe that greater feature overlap leads to higher coverage (exceeding the nominal level) and more efficient

prediction sets. In the MoE setting, this may be because feature sharing leads to more consistent estimations across experts, improving the reliability of the aggregated p-values and, in turn, reducing the size of the prediction sets. More generally, this may imply that the information redundancy introduced by feature overlap allows for better sample efficiency. This parallels findings in aggregation methods like cross-conformal and jackknife+, which also tend to produce smaller prediction sets than split conformal by reusing data. These results suggest that feature sharing is an important design consideration when aggregating prediction sets from multiple models.

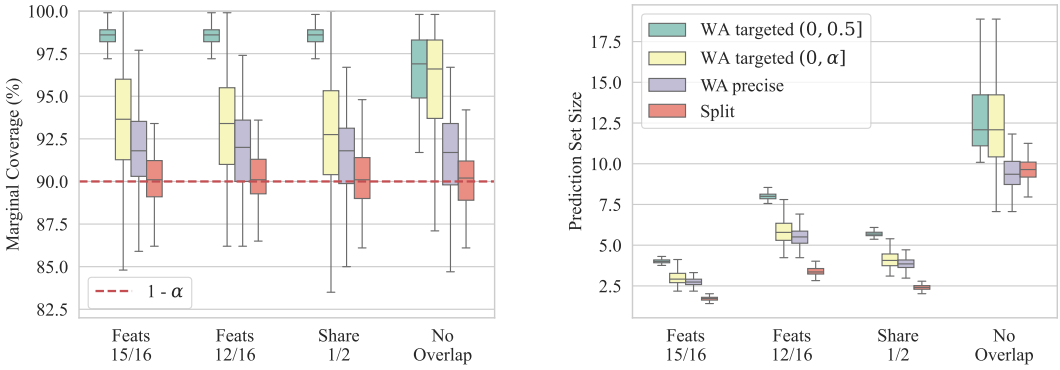

Figure 5: Coverage (left) and prediction set size (right) for different feature assignment methods and weighted aggregation methods. Our results indicate that sharing fewer features leads to tighter coverage, but sharing more features leads to more efficient (smaller) prediction sets. We also see that WA precise tends to have coverage that is closest to nominal and the most efficient prediction sets, albeit with a much looser guarantee.

B.1.3 More general coverage guarantees result in more conservative prediction sets

We now compare the different variants of weighted aggregation described in §6, focusing on how the generality of the guarantee for each variant affects its empirical coverage.

Figure 5 compares the effects of different feature assignment methods and for each weighted aggregation variant (with $\alpha' = 0.1$), and Figure 6 shows how coverage varies across α for each variant.

Unsurprisingly, WA precise—which provides the narrowest guarantee, targeting a single α—achieves coverage closest to the nominal level. In contrast, WA targeted offers more general guarantees over a range of α values, but tends to overcover, reflecting the conservativeness built into the method to accommodate worst-case behavior. We observe the same pattern on UCI data in Figure 3. These results illustrate the trade-off between the conservativeness of a method and the generality of its guarantee, and suggest that more targeted guarantees may be preferable when tighter coverage is important.

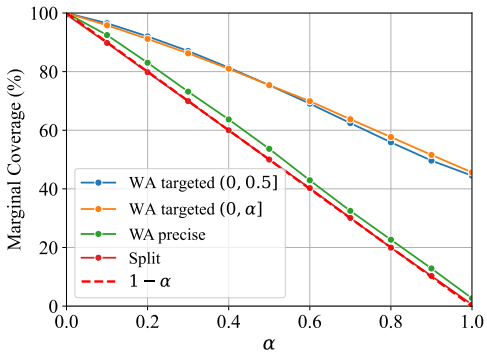

Figure 6: Mean coverage across significance levels α for the different weighted aggregation variants. WA precise achieves close to nominal coverage, while WA targeted tends to overcover (but offers more general guarantees).

B.2 Real data

B.2.1 Weighted aggregation improves local validity for classification

The experiments in Figure 7 mirror those in Figure 3, but for classification instead of regression. Like with the other set of experiments, we compare split conformal to weighted aggregation, and we see that split conformal achieves marginal coverage close to nominal but consistently undercovers on the WS slab. In contrast, weighted aggregation maintains much better WS coverage, suggesting it offers better local validity in the classification setting as well.

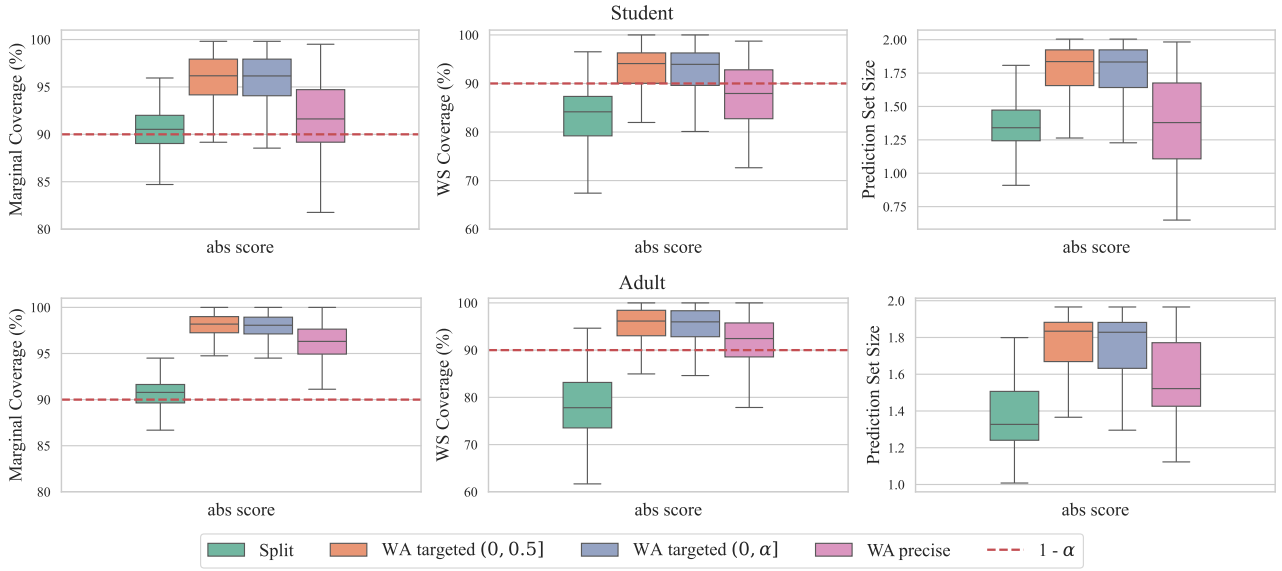

Figure 7: Local validity experiments comparing split conformal (green) to weighted aggregation (orange, purple, and pink) on the classification task. Like in Figure 3, each row corresponds to a dataset, with plots for marginal coverage, WS coverage, and prediction set size from left to right. We find that for classification as well as regression, WA improves WS coverage over split conformal, which undercovers in the WS region.

As before, we note that there is trade-off between coverage and efficiency. WA targeted is the best choice in terms of WS coverage, but WA precise balances the improved WS coverage of weighted aggregation methods with an efficiency that is closer to the split conformal. However, if marginal coverage and prediction set efficiency are the only priorities, then split conformal should be preferred.

B.2.2 Local validity over demographic groups

In this section, we evaluate our methods on the Communities and Crimes dataset (Redmond and Baveja, 2002), where the task is to predict the per capita violent crime rate of a community based on its demographic features, and our primary interest is in understanding how coverage differs across communities with varying racial compositions (Gibbs et al., 2023).

Figures 8 and 9 compare split conformal with weighted aggregation variants in terms of coverage and prediction set size across demographic groups. Unlike Figures 3 and 7, where local validity is assessed via WS coverage, these experiments evaluate local validity in terms of consistency across demographic groups. Thus, we display group-specific performance for each method, which we group further by the two types of nonconformity score, absolute residual and CQR score.

The demographic groups in Figures 8 and 9 represent communities where a particular racial demographic is in the top p-percentile of representation (Gibbs et al., 2023). Figure 8 shows results for $p = 50$, and Figure 9 for $p = 70$.

Marginal coverage and coverage across groups The "All" category in the coverage plots represents coverage across all demographic groups, or marginal coverage. As in prior experiments, we see that split conformal achieves marginal coverage closest to nominal, with WA precise close behind, while WA targeted variants tend

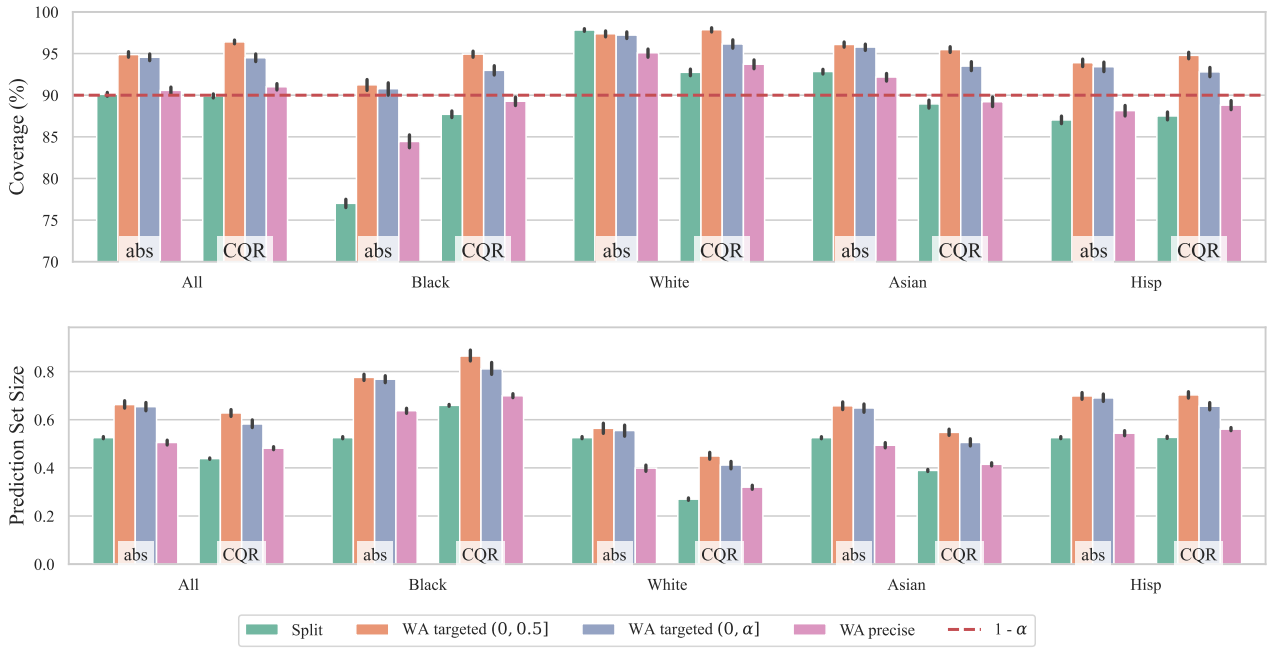

Figure 8: Coverage (top) and prediction set size (bottom) for split conformal (green) and weighted aggregation variants (orange, purple, and pink) across subgroups with top 50th percentile racial representation. Split conformal has precise marginal coverage ("All"), but WA variants have more consistent coverage across subgroups, with WA targeted meeting coverage for all subgroups. Error bars represent 95% confidence intervals.

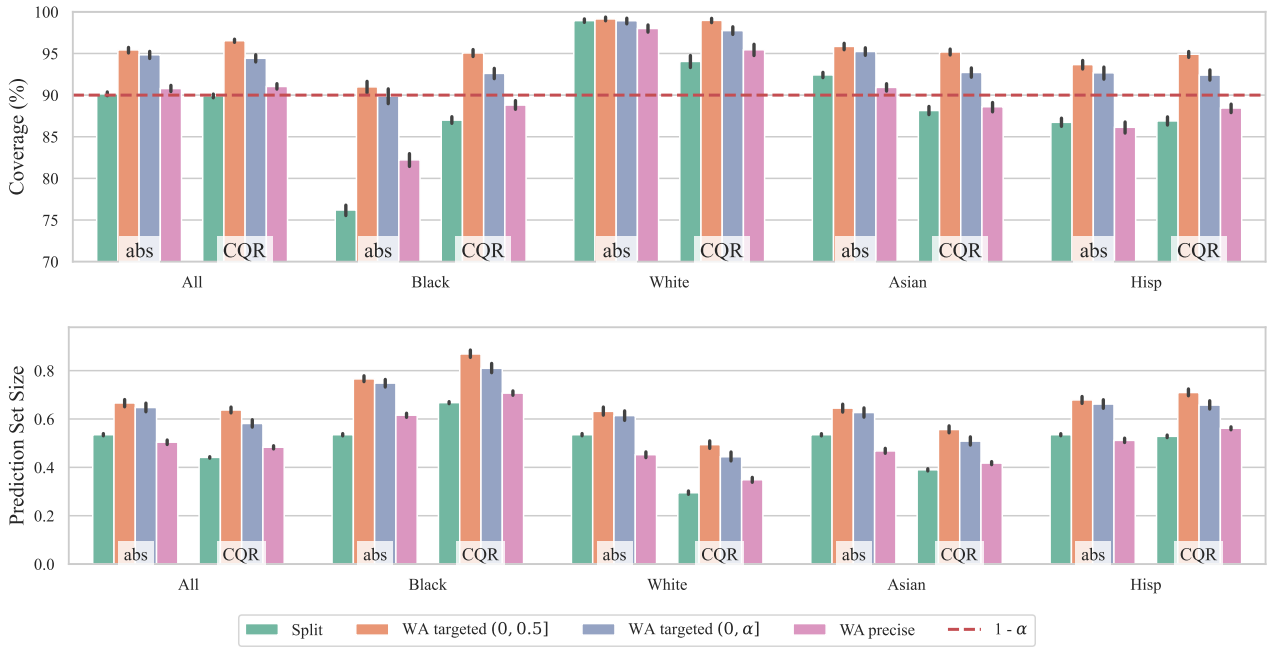

Figure 9: Coverage (top) and prediction set size (bottom) for split conformal and weighted aggregation variants across subgroups with top 70th percentile racial representation. As in Figure 8, WA targeted meets coverage for all subgroups. Error bars represent 95% confidence intervals.

to overcover. However, although split conformal enjoys precise marginal coverage, it also exhibits substantial disparities in performance across demographic groups, significantly undercovering for Black and Hispanic groups while overcovering for White. In contrast, WA variants display far less demographic variation, with WA targeted achieving coverage for all demographic groups.

Comparing nonconformity scores For all demographic groups except the Asian group, using CQR scores instead of absolute residuals improves coverage for split conformal, with the improvement being most pronounced for the Black group. However, while CQR reduces undercoverage, it is never sufficient to fully close the coverage gap for an undercovered group. Rather, its primary benefit appears to be in reducing the variability in coverage across groups. Across all methods, the variance in CQR-based coverage is lower than that of absolute residuals, suggesting that CQR contributes to more stable group-wise coverage, even if it does not fully mitigate disparities.

B.2.3 Comparing to conservative CQR

In our main results (Figure 3), weighted aggregation typically has higher marginal coverage than CQR (that is, split conformal with CQR nonconformity scores), but also achieves WS coverage that is closer to target. This raises a natural question: does weighted aggregation achieve better WS coverage simply because it is more conservative, or does this reflect genuine adaptivity to difficult regions of the data?

Experimental details To disentangle these effects, we compare weighted aggregation to a *conservative version of CQR*. For this baseline, we tune the significance level α of CQR so that its empirical marginal coverage matches that of weighted aggregation as closely as possible. (Note that because conformal prediction with a finite calibration set induces a discrete grid of attainable coverage levels, this matching cannot be exact in general; we select the α that yields the closest available coverage.)

We evaluate both methods over 200 trials and summarize the results in Figure 10. To facilitate comparisons even when marginal coverages are not perfectly matched, we also report Δ coverage, or the difference between the marginal and WS coverage per trial.[3]

Discussion of Figure 10 While matching the marginal coverage does move the WS coverage of conservative CQR closer to that of weighted aggregation, performance differences remain. Conservative CQR still undercovers on the WS slab for the Airfoil and Superconduct datasets, and its average Δ coverage remains consistently larger across all datasets. This indicates that, even when conservative CQR meets the coverage requirement on average (marginally), it fails to maintain that validity locally (on the WS). The lower Δ coverage of weighted aggregation shows a smaller gap between average and worst-case coverage, indicating a more uniform distribution of validity across the input space.

Meeting this higher marginal coverage also forces conservative CQR to inflate its prediction sets more than weighted aggregation does. In other words, when forced to be equally conservative overall, CQR improves WS coverage primarily by making sets *globally wider*. In contrast, weighted aggregation improves WS coverage by redistributing uncertainty toward the most challenging regions where coverage failures are more likely. This targeted behavior suggests that weighted aggregation is more adaptive than CQR, rather than simply more conservative.

[3]Although potentially unintuitive, empirical WS coverage can occasionally exceed empirical marginal coverage because the WS metric minimizes over a restricted family of slabs (those with at least a δ fraction of test points, and, in practice, a finite random subset). Consequently, the minimizing slab is not necessarily the region driving undercoverage, and the resulting minimum may be higher than the overall average. For this reason, Δ coverage can be occasionally less than 0.

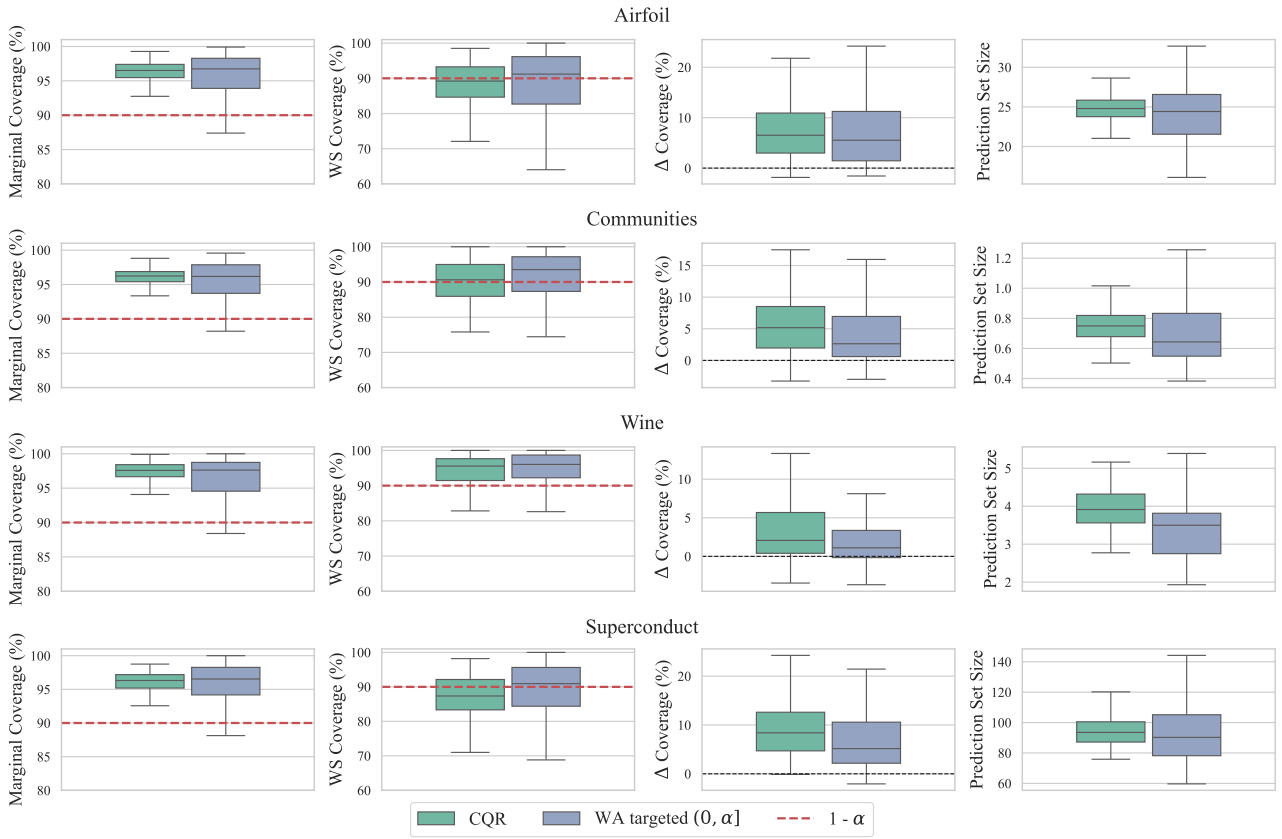

Figure 10: Comparison between WA targeted $(0, 0.1]$ and CQR at matched empirical marginal coverage. Each row corresponds to a dataset, with plots for marginal coverage, WS coverage, Δ coverage, and prediction set size from left to right. After matching marginal coverage, WA targeted $(0, 0.1]$ has more uniform coverage (that is, smaller Δ coverage) and smaller prediction set size on average.

C Comparison to other aggregation methods

C.1 Benefits of weighted p-value aggregation

Weighted p-value aggregation is particularly effective in the realistic and practically important scenario when the predictors to be aggregated are not perfectly aligned: that is, when the different predictors capture complementary aspects of a problem, so that no single predictor is uniformly superior (more predictive) across the input space. This setting is represented in Figure 1, and can occur, for example, when the predictors are trained on different modalities or specialize in different operating conditions. In this scenario, an effective aggregation method should (1) faithfully reflect whether the predictors agree or disagree, and (2) adjust to a test point by assigning more weight to the most relevant predictors at that point.

In the standard MoE formulation, weights are optimized for predictive accuracy (under mean squared error or cross-entropy loss), and reflect how relevant each expert is for a given test point. Under this setting, weighted p-value aggregation achieves both of these goals. We elaborate on this below.

C.1.1 p-value aggregation best reflects expert disagreement

One way to taxonomize aggregation methods is to consider whether they combine *calibrated* uncertainty representations (like p-values or prediction sets) or *pre-calibrated* representations (like scores). To illustrate how these differ, consider the extreme case of expert diversity where two experts have disjoint prediction sets.

Pre-calibration aggregation methods eliminate disagreement Methods that combine scores before calibration (Luo and Zhou, 2025) convert the experts' diverse behaviors into a single blended score that no longer encodes which expert supported which region. In the disjoint expert example, averaging pre-calibration scores would collapse the final prediction set into a single region that neither expert truly supports.

To see this more explicitly, let each expert use a standard convex conformal score (e.g. absolute residual or quantile regression score) so each $s_k(x, y)$ is convex in y. Then, the average score $s_{\text{avg}}(x, y) = \sum_k w_k s_k(x, y)$ is convex, and the conformal set produced from this score is the sublevel set $C_{\text{avg}} = \{y : s_{\text{avg}}(x, y) \leq Q_\alpha^+\}$, and is also convex. Thus, even if the experts' individual prediction sets are disjoint, the averaged score produces one connected prediction set which may not reflect the support of either expert.

Set selection methods eliminate disagreement Methods that select one prediction set (Yang and Kuchibhotla, 2021) necessarily discard other experts' predictions. In the disjoint expert example, the aggregate set reflects only one expert's prediction, and all complementary evidence from the other expert is lost. This means the aggregate set ignores valid, calibrated support for labels that remain plausible.

p-value aggregation preserves the degree of support from each expert In the end, the only methods that preserve expert disagreement are the methods that combine, and do not just select, *calibrated* representations—p-value aggregation, majority vote (Cherubin, 2019), and set union. However, these methods differ substantially in how much expert information they retain.

Majority vote and set union operate on binary membership. For weighted majority vote (Gasparin and Ramdas, 2024a), the prediction set is

$$C_{\text{WMV}}(x) = \left\{ y : \sum_k v_k \mathbb{1}\{p_k(x, y) > \alpha\} \geq \frac{1}{2} \right\},$$

so each expert contributes only the indicator $\mathbb{1}\{p_k(x, y) > \alpha\}$. This tracks whether an expert supports a label, but not how strongly they support it. In contrast, p-value aggregation combines the full p-value functions, with prediction set

$$C_{\text{p-value}}(x) = \left\{ y : \sum_k w_k p_k(x, y) > \alpha \right\}$$

(restated from (5)). If the ith expert assigns $p_i(x, y) = 0.95$ and the jth expert assigns $p_j(x, y) = 0.12$ at $\alpha = 0.1$, majority vote records only that both experts include (x, y), while p-value aggregation keeps their relative strength

with $w_i \cdot 0.95 + w_j \cdot 0.12$. Thus, p-value aggregation *preserves graded expert information*, and can distinguish regions of strong support from regions of weak support. This sensitivity to the degree of support from each expert allows p-value aggregation to preserve disagreement more faithfully than binary set-based methods.

Why preserving disagreement matters With complementary experts, disagreement often reflects genuine multi-modality or heterogeneity in the underlying data. Aggregation methods that collapse this structure risk including labels no expert supports, or excluding labels at least one expert favors. By retaining the full shape and strength of each expert's p-value function, p-value aggregation ensures that the final set aligns more closely with the true uncertainty represented across experts.

C.1.2 p-value aggregation allows for data-dependent weights

As we see in (1), p-variables satisfy a direct and explicit validity condition. This condition is expressed entirely in terms of the distribution function of a single random variable, and as such, is *manipulable*: we can combine p-values, examine how the combination behaves relative to this inequality, and apply a correction that restores validity. Prediction sets and nonconformity scores do not admit a similarly explicit validity criterion that allows the aggregated object to be repaired directly after reweighting, so no analogous simple correction exists for majority vote, set union, or score-based averaging.

Consider again the case where experts specialize in different regions of the input space, so that the expert that is most relevant at each point varies across the domain. Majority vote and set union cannot express this form of adaptivity. Even in weighted majority vote, the weights must be fixed in advance or otherwise independent of the test input, so the weighting rule is applied uniformly across the input space. This means every expert contributes its vote at every test point in the same way, regardless of whether that expert is informative for the local region. When an expert is poorly suited to the region around the test point, its vote can (at best) widen the prediction set unnecessarily, or even (at worst) counteract the experts that are actually reliable there.

Set selection methods have a more fundamental limitation. By construction, set selection yields a single global choice of expert, and this remains fixed for all future test points. This is incompatible with any notion of input adaptivity, which is important to settings where local conditions change.

C.2 Comparison to additional adaptive baselines

Additional adaptive baselines One of the benefits of our proposed method is that its extension to data-dependent weights allows for the weight-aggregated set to adapt to the specific test input. This property of adapting to the test input is called *local*, *adaptive*, or *conditional*.

In our experiments in §6, we compare to conformal quantile regression (CQR) (Romano et al., 2019) as our adaptive baseline. CQR is simple but powerful, and can be considered a seminal adaptive method with multiple follow-up works (Sesia and Candès, 2020; Liu et al., 2022; Jensen et al., 2022; Rossellini et al., 2024). Our experiments in §6 both compare to CQR (represented as split conformal with CQR scores), and combine our method with CQR (represented as weighted aggregation with CQR scores). To further contextualize the local validity of our approach, we now compare to two additional baselines.

- *Localized conformal prediction (LCP)* (Guan, 2019): LCP reweights calibration scores with weights $w_i \propto H(X_i, X_{n+1})$, where the localization kernel H measures the proximity between the test point X_{n+1} and each calibration point X_i. At test time, LCP forms an empirical distribution of weighted calibration scores and solves for the significance level $\tilde{\alpha}(x_{n+1})$ that restores marginal coverage after the pointwise weighting. LCP can be computationally intensive, as each test point requires kernel weights over all calibration points.
- *Locally weighted conformal prediction (LWCP)* (Lei et al., 2018): LWCP normalizes calibration residuals by fitting one model for the mean, a second model for the local error scale, and then rescaling each calibration residual by its estimated variability. Running conformal prediction on these normalized scores yields intervals whose widths adapt to local uncertainty.

We focus on comparing to methods that improve coverage across the input space in a general data-adaptive way. For this reason, we do not compare to methods whose local validity requires the user to first specify a target family of subpopulations or distributional shifts, like Gibbs et al. (2023).

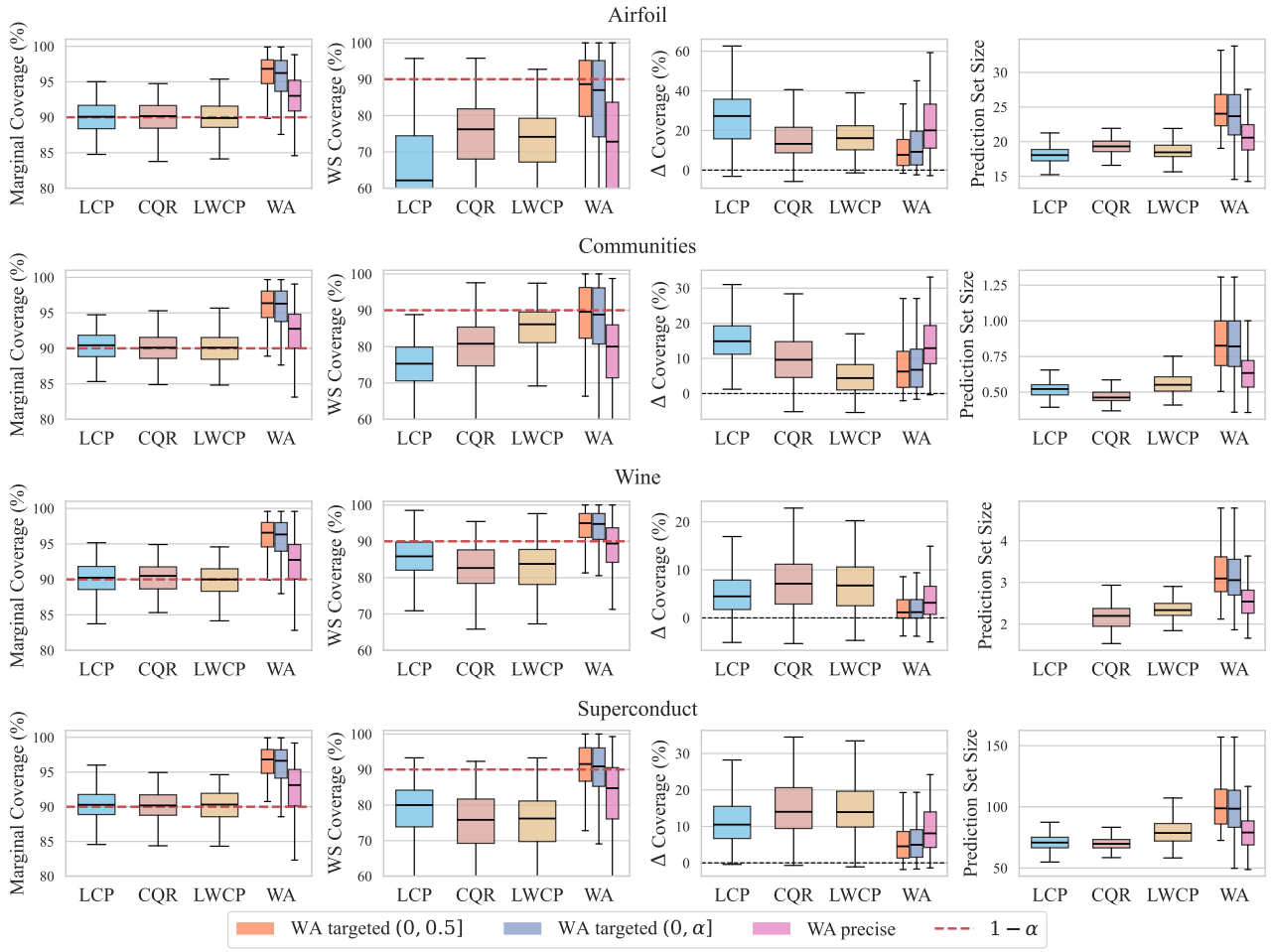

Figure 11: Comparison between weighted aggregation and adaptive baselines. Each row corresponds to a dataset, with plots for marginal coverage, WS coverage, Δ coverage, and prediction set size from left to right. Although the weighted aggregation methods (WA) are conservative marginally, it also has the most uniform coverage (that is, the smallest Δ coverage overall), and achieves closest-to-nominal coverage on the WS slab.

Weighted aggregation can be combined with other adaptive methods LCP introduces adaptivity by reweighting calibration scores, and CQR and LWCP introduce adaptivity through adaptive score functions. In contrast, our method introduces adaptivity at the *aggregation* step, by reweighting the post-calibration p-value functions in a data-dependent way. Since these adaptive methods operate on different steps of the conformal procedure, they generally do not interfere with one another. Thus, because weighted aggregation introduces adaptivity at a different step than LCP, CQR, or LWCP, it complements rather than competes with these methods, and can in principle be combined with any of them.

Experimental details For this experiment, we run each of the methods separately to isolate their individual effects. These results are therefore presented differently than in Figure 3, where we had to visualize composition between the absolute residual score / CQR score and split conformal / weighted aggregation. In Figure 11, LCP, CQR, and LWCP are separate methods that are each compared to the weighted aggregation methods (WA) directly. (More precisely, LCP, CQR, and LWCP are each implemented on top of split conformal with the full MoE predictor—each method replacing only the component it introduces, with the rest of the procedure matching split conformal using the full MoE.)

We evaluate all methods over 200 trials and report distributions of marginal coverage, WS coverage, Δ coverage, and prediction set size (as in Figure 10). Figure 11 summarizes the results.

Discussion of Figure 11 The comparisons between weighted aggregation and the adaptive baselines reflect a trade-off between marginal efficiency and adaptivity. Weighted aggregation is marginally overconservative,

whereas the other methods have marginal coverage close to target. However, the weighted aggregation methods consistently brings WS coverage close to target, while the other methods substantially undercover the WS slab. In addition, weighted aggregation almost always achieves the best Δ coverage: even though its marginal coverage is conservative, the gap between marginal and WS is almost always smaller than the other baselines, indicating that weighted aggregation reallocates uncertainty more effectively in the hardest regions (the WS slab). These results suggest that the main advantages of weighted aggregation are in its *adaptivity*, by providing more uniform coverage overall and achieving better coverage on the WS slab. We therefore suggest to use weighted aggregation when performance on challenging/minority slices is a priority.

C.3 Weighted majority vote

The idea to combine conformal prediction sets by weighted majority vote was introduced by Gasparin and Ramdas (2024a) as an extension of the majority vote method first proposed by Cherubin (2019). At first consideration, weighted majority vote appears to differ from our method: weighted majority vote performs weighted aggregation of *prediction sets*, while our method performs weighted aggregation of the *p-values associated with prediction sets*. Despite this distinction, Appendix B of Gasparin and Ramdas (2024a) observes that these two methods are, in fact, dual to each other under data-independent weights, the setting considered in their work. Nevertheless, our p-value formulation enables two key extensions that go beyond weighted majority vote.

- Our formulation allows us to apply the result of Vovk and Wang (2020) to strengthen coverage guarantees for *data-independent* weights when they are sufficiently asymmetric.
- Our formulation provides a principled extension to *data-dependent* weights by transforming the weighted average of p-variables to also be a valid p-variable. Not only does this allow us to use weights learned from data, but it also yields a form of local validity, a property not available to existing set aggregation methods.

Because our method is a dual formulation to the weighted majority vote method of Gasparin and Ramdas (2024a), we do not include it as a separate baseline to avoid redundancy.

C.4 Extending the p-variable transformation of Stutz et al. (2023) to weighted aggregation

To the best of our knowledge, weighted majority vote (Gasparin and Ramdas, 2024a,b) is the only existing method to address weighted prediction set aggregation, and the method of Stutz et al. (2023) is designed for the different problem of uncertainty in the ground truth labels. To address their problem, Stutz et al. (2023) propose sampling m labels for each calibration point and using the labels to compute m p-values, then taking the unweighted average of these p-values and applying a transformation to obtain a valid p-variable.

Although Stutz et al. (2023) only consider the unweighted average, their transformation is general enough to apply to a weighted average of p-values as well, and can therefore be adapted to our setting.

How the transformations affect the weights Both our method and the method of Stutz et al. (2023) aim to transform a random variable to a p-variable to maintain coverage guarantees. The difference between the two methods lies in the nature of the transformation. Our method applies a *linear* transformation that preserves the proportions of the weights; this can be important when the weights reflect meaningful quantities, like the weights learned by the routing network of an MoE model. In contrast, Stutz et al. (2023) apply a *nonlinear, rank-based* transformation by computing the empirical CDF of the random variable and returning its value at the observed point. That is, given a random variable X, their method estimates its CDF F and uses $F(X)$ as the resulting p-value. While this guarantees validity and preserves ordering, it does not preserve the relative scale between values and therefore discards some of the semantics of the original weights.

Comparing both transformations with our MoE setup In our method for weighted prediction set aggregation with data-dependent weights, we transform the weighted average of the individual prediction set p-values to a valid p-variable using a linear scaling. This transformation allows us to maintain a coverage guarantee for the prediction set defined by the weighted average of the p-values. In the MoE setting, this prediction set corresponds to aggregating the p-values from each expert according to the weights learned by the routing network.

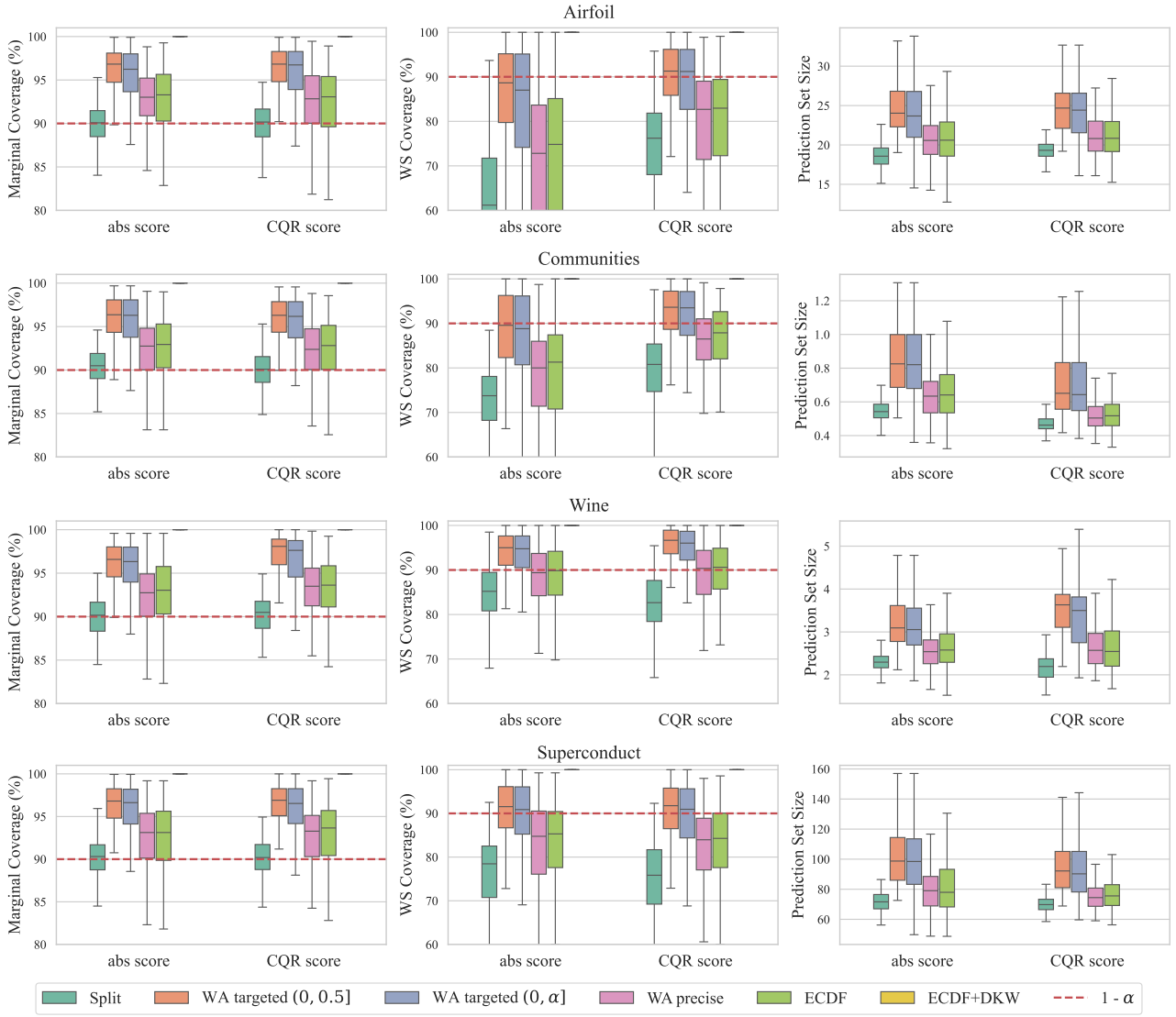

Figure 12: Regression experiments of our linear transformation method with ECDF and ECDF-DKW. Each row corresponds to a dataset, with plots for marginal coverage, WS coverage, and prediction set size from left to right. ECDF performs similarly (sometimes slightly more conservatively) to WA precise in terms of coverage and efficiency. However, ECDF-DKW is so conservative that it covers the entire label space.

To adapt the method of Stutz et al. (2023) to this context, we substitute their transformation in place of our linear scaling. We now restate their original setting and transformation in more detail to clarify how their method can be extended to our setting.

The transformation proposed by Stutz et al. (2023) was originally developed to address the problem of uncertainty in the ground truth labels. In their setting, each calibration point consists only of an input X_i; they sample m labels to get m p-values for each calibration point, and then take the unweighted average of these p-values. We denote this unweighted average as P^i_{avg}, with distribution function F. Their method transforms P^i_{avg} into a valid p-value via $F(P^i_{\text{avg}})$, and the prediction set is then the set of all labels such that the corrected p-value $F(P^i_{\text{avg}})$ corresponding to each sample is greater than some threshold α.

The coverage guarantee for this corrected p-value holds only with an asymptotic number of samples, as the empirical CDF $\widehat{F}$ approaches the true CDF F. To establish a finite-sample guarantee, Stutz et al. (2023) introduce a DKW-derived correction ϵ and define their prediction sets based on $\widehat{F} + \epsilon$. We find that, although their proposed prediction set yields an elegant $(1-\alpha)(1-\delta)$ finite-sample guarantee, the ϵ correction is extremely conservative in practice.

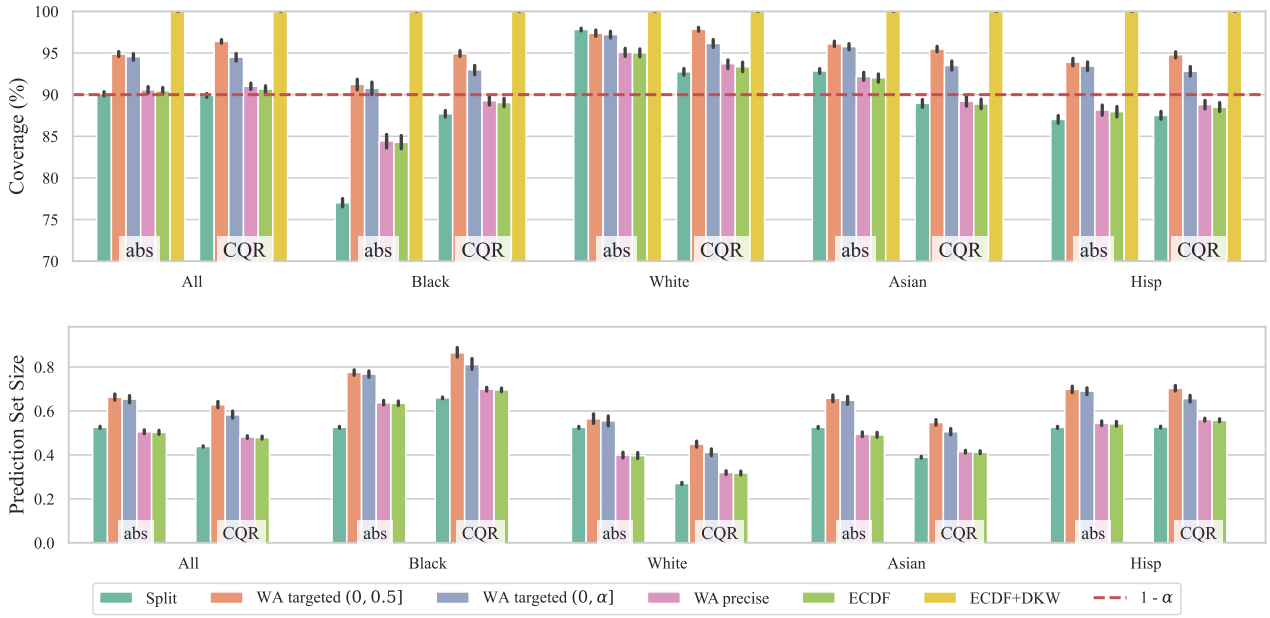

Figure 13: Coverage (top) and prediction set size (bottom) for subgroups with top 50th percentile racial representation in the UCI Communities and Crimes dataset. ECDF performs similarly to WA precise; ECDF-DKW is so conservative that it covers the entire label space. Error bars represent 95% confidence intervals.

Let us refer to the finite-sample empirical CDF method as *ECDF-DKW*, and the variant without the DKW correction as simply *ECDF*. We now present additional experiments where we recreate the main findings of our paper with the ECDF and ECDF-DKW methods.

Figure 12 recreates the regression experiments of Figure 3, with the addition of ECDF and ECDF-DKW. We note that ECDF performs very similarly to WA precise in terms of coverage (left), WS coverage (middle), and prediction set size (right), with ECDF being slightly more conservative in most cases. Like WA precise, ECDF is more conservative than split conformal and less conservative than WA targeted, although it still often falls short of the $1-\alpha$ guarantee on the WS slab. On the other hand, ECDF-DKW is so conservative that its prediction sets cover the entire label space. We represent this with 100% coverage on both the coverage plots and WS coverage plots, and we omit ECDF-DKW from the prediction set size plots.

Figure 13 recreates the Communities and Crimes experiment of Figure 8 with ECDF and ECDF-DKW. Again, we see that ECDF performs very similarly to WA precise on our demographic-conditioned view of Communities and Crimes, with similar coverage (top) and prediction set size (bottom) across all demographics—with ECDF having slightly lower coverage on most demographics, including demographics where WA precise undercovers. As before, we observe that ECDF-DKW is so conservative that it has 100% coverage and unbounded prediction sets.

Why is ECDF-DKW so conservative? For ECDF-DKW, the finite-sample variant of ECDF, Stutz et al. (2023) use DKW to add a finite-sample correction ϵ to the empirical CDF $\widehat{F}(P_{\text{all}})$, then compare this sum to the significance level α. The prediction set with finite-sample guarantees is therefore the set of all labels such that $\widehat{F}(P_{\text{avg}}) + \epsilon > \alpha$. However, if ϵ is already greater than α, then this condition is always satisfied and the prediction set includes *all* labels, becoming unbounded.

The finite-sample correction ϵ is a function of the number of samples used to compute the empirical CDF. Figure 14 shows that when the merging set size is less than 1000, ϵ is typically large enough to exceed common values of α, making unbounded sets very likely.

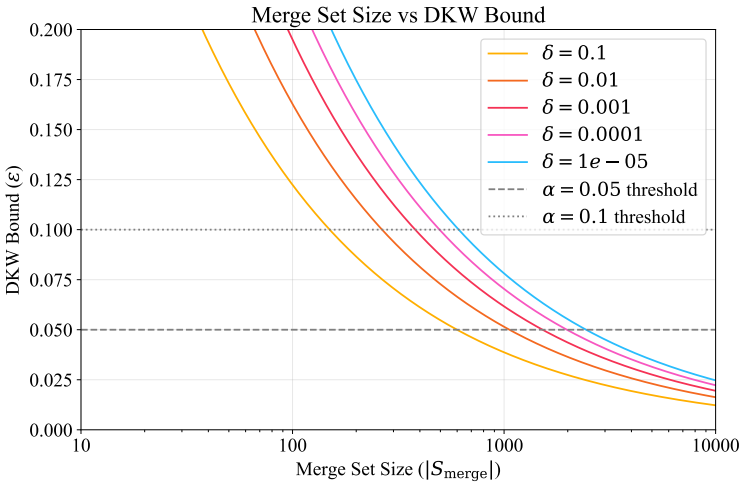

Figure 14: Finite-sample correction ϵ used in ECDF-DKW as a function of the merging set size $|S_{\text{merge}}|$ and user-specified significance level δ. This correction is the offset required to ensure the finite sample guarantee of $(1-\alpha)(1-\delta)$ in Stutz et al. (2023). Dashed lines mark α levels of 0.05 and 0.1. When $\epsilon > \alpha$, the prediction set includes all labels and becomes unbounded, yielding 100% coverage.

D A note on computational complexity

The computational complexity of our method matches the complexity of existing prediction set aggregation methods for the case of data-independent weights, and includes an additional one-time cost to compute an empirical CDF for the case of data-dependent weights.

The complexity of prediction set aggregation methods was first observed by Cherubin, who compares the computational overhead of their majority vote aggregation method with the overhead of p-value aggregation methods. They note that while majority vote requires simpler operations to determine whether each label is included in the final prediction set, it also requires that predictions be recomputed for each significance level, making it less efficient when sets must be constructed at multiple thresholds. In contrast, p-value methods allow the aggregation to be computed once and then applied to any significance level without additional computation.

These trade-offs in speed and cost may influence which method is better suited to a given application—for example, p-value aggregation may be preferable if prediction sets need to be dynamically thresholded at test time. However, the time complexity of both methods is the same: combining K prediction sets for N test objects with a label space size of $L = |\mathcal{Y}|$ has complexity $\mathcal{O}(KLN)$. This complexity is necessary for all prediction set aggregation methods, as it reflects the cost of evaluating multiple prediction sets across the label space.

With data-independent weights, our method matches this $\mathcal{O}(KLN)$ complexity directly. With data-dependent weights, the only additional computation required is a one-time estimation of a correction factor $\widehat{m}^*$ from data split S_{merge}. This step involves computing the empirical CDF of the weighted average p-values on S_{merge}, which has a complexity of $\mathcal{O}(M \log M)$ for a split of size M. Importantly, this correction is computed once and does not require retraining, and the rest of the procedure for data-dependent weights has the same $\mathcal{O}(KLN)$ cost as other prediction set aggregation methods.

E Further implementation details

E.1 Data

Synthetic data For the ablation/expository experiments in §B.1, we generate a simple homoskedastic dataset to simulate a regression task. Each input is a 16-dimensional vector drawn from a standard normal distribution, and the output label is the sum of the feature values with additive Gaussian noise. Specifically, for each sample X_i, we have

$$y_i = \sum_{j=1}^{16} X_{ij} + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ represents additive noise with standard deviation $\sigma = 0.1$. We generate both training and test datasets by independently drawing samples and computing the corresponding target labels.

Data split In all experiments, we randomly partition the data into four disjoint subsets of train, calibration, merge, and test. For the experiments on UCI data, we allocate 400 samples for train/calibration/merge with a 50/40/10 split (where the number of training samples follows Barber et al. (2021)). For methods that do not require a merge set (e.g. split conformal), we use a 50/50 train/cal split over the same 400 sample budget. This ensures that all methods use the same amount of training data.

E.2 Mixture-of-experts (MoE) model and optimization

Model architecture To isolate the effects of aggregation from model complexity, we use linear models for both the experts and the routing networks.

We define an MoE model with K experts, where the ith expert is a linear model $f_i(x; \theta_i)$ parameterized by θ_i, and the routing network is a linear model parameterized by ϕ. For each expert i, it produces a raw output $g_i(x; \phi)$. To ensure the expert weights $\{w_i(x)\}_{i=1}^K$ are positive and sum to one, we apply a softmax function to these outputs:

$$w_i(x) = \frac{\exp(g_i(x; \phi))}{\sum_{j=1}^K \exp(g_j(x; \phi))},$$

These weights determine the contribution of each expert to the final prediction.

Feature assignments In the UCI experiments, the features for each dataset are partitioned into semantically related groups, and each expert specializes in a single group (see §E.5 for feature groups). In our ablation experiments on synthetic data, we vary the degree of feature overlap between experts to explicitly study its impact on coverage in §B.1.2.

Optimization and sequential training All models are optimized using L-BFGS to minimize mean squared error (MSE). We adopt a sequential training procedure to improve stability (Nie et al., 2021; Lin et al., 2024). First, we train each expert independently on its assigned subset of features. Then, we train the routing network with the full MoE to minimize MSE of the aggregated prediction

$$\ell_{\text{routing}} = \frac{1}{|S_{\text{train}}|} \sum_{i \in S_{\text{train}}} \left(y_i - \sum_{j=1}^K w_j(x_i) f_j(x_i; \theta_j) \right)^2.$$

Note that, while experts may observe different features, all components are trained on the same train set.

Routing at inference In our experiments, the trained MoE model is utilized in two ways. For split conformal, we use the weighted sum of experts as a standard black-box predictor. For weighted aggregation, we use the weights from the routing network to scale the p-value functions of each expert.

E.3 Implementation of weighted aggregation

Computing prediction sets We leverage the structure of the p-value function to efficiently compute prediction sets.

For a fixed input x, the function $\widehat{p}_k(x, y)$ (2) is piecewise constant, with discontinuities occurring only at values of y that solve $s_k(x, y) = s_k(X_i, Y_i)$. (For an absolute residual score function $s_k(x, y) = |y - \widehat{\mu}_k(x)|$, these values correspond to $\widehat{\mu}_k(x) \pm s(X_i, Y_i)$ for all $i \in S_k$.) Thus, for each test point x, $\widehat{p}_k$ is a step function of y with finite discontinuities, and the weighted average p-value function is also a step function of y, with its discontinuities as a union of the K separate sets of discontinuities. We compute prediction sets by identifying these candidate discontinuities, evaluating the p-value at the midpoint of each interval, then taking the union of all intervals where $\widehat{p}(x, y) > \alpha$.

We use the `portion` library (Decan, 2018) to represent prediction sets and efficiently compute their Lebesgue measures without discretization.

Computing correction factors $\widehat{m}^*$, $\widehat{m}^\dagger$, $\widehat{m}^\ddagger$ All correction factors are computed using the empirical CDF of the aggregated p-values on the merging set S_{merge}.

For each point $(X_i, Y_i) \in S_{\text{merge}}$, we calculate the aggregate p-value function $P^i_{\text{all}} = p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)})$. We then construct the empirical CDF $\widehat{F}$ and compute the ratios $\widehat{F}(P^i_{\text{all}})/P^i_{\text{all}}$ for all points in the set.

- $\widehat{m}^*$ is the maximum of these ratios over $i \in S_{\text{merge}}$.
- $\widehat{m}^\dagger$ (targeted) restricts this maximum to points where $\widehat{F}(P^i_{\text{all}}) \leq \alpha'$.
- $\widehat{m}^\ddagger$ (precise) uses only the first point i where $\widehat{F}(P^i_{\text{all}}) \geq \alpha'$.

Formal definitions for $\widehat{m}^*$ and $\widehat{m}^\dagger$ are provided in §A.5.

Conservative empirical CDF To evaluate the correction factor $\widehat{m}^*$ in practice, we approximate the distribution function $F_{P_{\text{all}}}$ by its empirical counterpart on a designated merging set S_{merge}. The standard empirical CDF of the random variable $P_{\text{all}} = p_{\text{all}}(X, Y; \mathbf{W})$ is

$$\widehat{F}_{P_{\text{all}}}(\alpha) = \frac{\sum_{i \in S_{\text{merge}}} \mathbb{1}\{p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}) \leq \alpha\}}{|S_{\text{merge}}|}.$$

When the weights depend on the data, this empirical CDF is used to compute $\widehat{m}^*$. In our experiments, however, we found that the naive estimator can be unstable for small or moderate $|S_{\text{merge}}|$. To mitigate this, we use a slightly more conservative version,

$$\widehat{F}^{\text{cons}}_{P_{\text{all}}}(\alpha) = \frac{\mathbb{1}\{\min_{i \in S_{\text{merge}}} p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}) \leq \alpha\} + \sum_{i \in S_{\text{merge}}} \mathbb{1}\{p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}) \leq \alpha\}}{1 + |S_{\text{merge}}|}. \tag{18}$$

This modification ensures that the CDF accounts for the minimum observed value and thereby avoids degenerate behavior in finite samples. The correction factor is then computed as

$$\widehat{m}^* = \max_{i \in S_{\text{merge}}} \frac{\widehat{F}^{\text{cons}}_{P_{\text{all}}}(p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)}))}{p_{\text{all}}(X_i, Y_i; \mathbf{W}^{(i)})},$$

as in (8). Importantly, using $\widehat{F}^{\text{cons}}_{P_{\text{all}}}$ in place of the standard empirical CDF does not affect the theoretical guarantees of our method (see §A.3), but in practice it yields more stable and reliable estimates of $\widehat{m}^*$.

Role of the merging set S_{merge} The merging set is used to construct an empirical CDF in order to estimate the correction factor m^* that makes the aggregated p-value valid. The size of this set induces a trade-off between giving a *conservative* or *precise* estimate, which we demonstrate experimentally in §B.1.1.

A small $|S_{\text{merge}}|$ results in a noisier empirical CDF; to maintain validity, the correction factor tends to be large, leading to overconservative prediction sets. As $|S_{\text{merge}}|$ increases, the empirical CDF becomes closer to the true CDF, allowing for a more precise estimate of m^* and coverage closer to target. Our ablation study in Figure 4 indicates that a modest sample size of ≈ 160 is sufficient to reduce overcoverage to within 3%, while maintaining validity.

E.4 Evaluation metrics

Worst-slice (WS) coverage We report WS coverage to evaluate the local validity of our prediction sets.

A *slab* of the feature space is a region bounded by hyperplanes:

$$S_{v,a,b} = \{x : a \leq v^\top x \leq b\}.$$

WS coverage is the minimum empirical conditional coverage over all slabs of the feature space that contain at least some fraction δ of the test points:

$$\min_{v,a,b:\mathbb{P}\{X \in S_{v,a,b}\} \geq \delta} \mathbb{P}\{Y \in C(X) | X \in S_{v,a,b}\}.$$

Intuitively, WS coverage measures performance on the most challenging slice that still has a sufficient number of test samples. In our experiments, we approximate this metric by evaluating coverage over 1000 random projections v and slab boundaries (a, b) and reporting the minimum, following Romano et al. (2020).

E.5 Expert feature assignments for UCI experiments

For the UCI experiments, the features for each dataset are partitioned into groups of semantically related features, and each expert in the MoE then specializes in a single group of features. We list these groups and their features below, where the feature names are provided with the original dataset (Kelly et al., 2010).

Table 1: Airfoil dataset feature groups

Group	**Features**
Aerodynamics	frequency, free-stream-velocity
Geometry	attack-angle, chord-length, suction-side-displacement-thickness

Table 2: Wine dataset feature groups

Group	**Features**
Acidity	fixed_acidity, volatile_acidity, citric_acid, pH
Sugar/alcohol	residual_sugar, density, alcohol
Sulfur/salinity	chlorides, free_sulfur_dioxide, total_sulfur_dioxide, sulphates

Table 3: Communities dataset feature groups

Group	Features
Population	population, householdsize, numbUrban, pctUrban, LandArea, PopDens, agePct12t21, agePct12t29, agePct16t24, agePct65up
Race/ethnicity	racepctblack, racePctWhite, racePctAsian, racePctHisp, PctForeignBorn, PctSpeakEnglOnly, PctNotSpeakEnglWell, PctBornSameState, PctSameHouse85, PctSameCity85
Income/poverty	medIncome, medFamInc, perCapInc, whitePerCap, blackPerCap, indianPerCap, AsianPerCap, HispPerCap, NumUnderPov, PctPopUnderPov
Employment/industry	pctWWage, pctWFarmSelf, pctWInvInc, pctWSocSec, pctWPubAsst, pctWRetire, PctUnemployed, PctEmploy, PctEmplManu, PctEmplProfServ
Occupation/education	PctOccupManu, PctOccupMgmtProf, PctWorkMomYoungKids, PctWorkMom, PctUsePubTrans, PctLess9thGrade, PctNotHSGrad, PctBSorMore, MalePctDivorce, MalePctNevMarr
Family structure	FemalePctDiv, TotalPctDiv, PersPerFam, PctFam2Par, PctKids2Par, PctYoungKids2Par, PctTeen2Par, PctLargHouseFam, PctLargHouseOccup, PctSameState85
Housing characteristics	PersPerOccupHous, PersPerOwnOccHous, PersPerRentOccHous, PctPersOwnOccup, PctHousNoPhone, PctHousLess3BR, MedNumBR, HousVacant, PctHousOccup, PctHousOwnOcc
Housing quality	PctPersDenseHous, PctVacantBoarded, PctVacMore6Mos, MedYrHousBuilt, PctWOFullPlumb, OwnOccLowQuart, OwnOccMedVal, OwnOccHiQuart, RentLowQ, RentMedian
Homelessness	RentHighQ, MedRent, MedRentPctHousInc, MedOwnCostPctInc, MedOwnCostPctIncNoMtg, NumInShelters, NumStreet, NumIlleg, PctIlleg, LemasPctOfficDrugUn
Immigration	NumImmig, PctImmigRecent, PctImmigRec5, PctImmigRec8, PctImmigRec10, PctRecentImmig, PctRecImmig5, PctRecImmig8, PctRecImmig10

Table 4: Superconductivity dataset feature groups

Feature Group	Features
Atomic mass	mean_atomic_mass, wtd_mean_atomic_mass, gmean_atomic_mass, wtd_gmean_atomic_mass, entropy_atomic_mass, wtd_entropy_atomic_mass, range_atomic_mass, wtd_range_atomic_mass, std_atomic_mass, wtd_std_atomic_mass, number_of_element
Atomic radius	mean_atomic_radius, wtd_mean_atomic_radius, gmean_atomic_radius, wtd_gmean_atomic_radius, entropy_atomic_radius, wtd_entropy_atomic_radius, range_atomic_radius, wtd_range_atomic_radius, std_atomic_radius, wtd_std_atomic_radius
Density	mean_Density, wtd_mean_Density, gmean_Density, wtd_gmean_Density, entropy_Density, wtd_entropy_Density, range_Density, wtd_range_Density, std_Density, wtd_std_Density
Electron affinity	mean_ElectronAffinity, wtd_mean_ElectronAffinity, gmean_ElectronAffinity, wtd_gmean_ElectronAffinity, entropy_ElectronAffinity, wtd_entropy_ElectronAffinity, range_ElectronAffinity, wtd_range_ElectronAffinity, std_ElectronAffinity, wtd_std_ElectronAffinity
FIE	mean_fie, wtd_mean_fie, gmean_fie, wtd_gmean_fie, entropy_fie, wtd_entropy_fie, range_fie, wtd_range_fie, std_fie, wtd_std_fie
Fusion heat	mean_FusionHeat, wtd_mean_FusionHeat, gmean_FusionHeat, wtd_gmean_FusionHeat, entropy_FusionHeat, wtd_entropy_FusionHeat, range_FusionHeat, wtd_range_FusionHeat, std_FusionHeat, wtd_std_FusionHeat
Thermal conductivity	mean_ThermalConductivity, wtd_mean_ThermalConductivity, gmean_ThermalConductivity, wtd_gmean_ThermalConductivity, entropy_ThermalConductivity, wtd_entropy_ThermalConductivity, range_ThermalConductivity, wtd_range_ThermalConductivity, std_ThermalConductivity, wtd_std_ThermalConductivity
Valence	mean_Valence, wtd_mean_Valence, gmean_Valence, wtd_gmean_Valence, entropy_Valence, wtd_entropy_Valence, range_Valence, wtd_range_Valence, std_Valence, wtd_std_Valence

Table 5: Adult dataset feature groups

Group	Features
Demographics	age, race, sex
Education	education, education-num
Economic status	fnlwgt, capital-gain, capital-loss, hours-per-week
Family relationship	marital-status, relationship

Table 6: Student dataset feature groups

Group	**Features**
Personal details	Marital Status, Gender, Age at enrollment, Nationality, International, Mother's qualification, Father's qualification, Mother's occupation, Father's occupation
Academic details	Application mode, Application order, Previous qualification, Previous qualification (grade), Admission grade, Daytime/evening attendance, Course, Displaced, Educational special needs
Performance	Curricular units 1st sem (credited), Curricular units 1st sem (enrolled), Curricular units 1st sem (evaluations), Curricular units 1st sem (approved), Curricular units 1st sem (grade), Curricular units 1st sem (without evaluations), Curricular units 2nd sem (credited), Curricular units 2nd sem (enrolled), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (approved), Curricular units 2nd sem (grade), Curricular units 2nd sem (without evaluations)
Economic and financial	Scholarship holder, Tuition fees up to date, Debtor, Unemployment rate, Inflation rate, GDP