Journal of Machine Learning Research 22 (2021) 1-42 Submitted 7/20; Revised 2/21; Published 3/21

Knowing what You Know: valid and validated confidence sets in multiclass and multilabel prediction

Maxime Cauchois MAXCAUCH@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305-4020, USA

Suyash Gupta SUYASH28@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305-4020, USA

John C. Duchi JDUCHI@STANFORD.EDU
Department of Statistics and Department of Electrical Engineering
Stanford University
Stanford, CA 94305-4020, USA

Editor: Amir Globerson

Abstract

We develop conformal prediction methods for constructing valid predictive confidence sets in multiclass and multilabel problems without assumptions on the data generating distribution. A challenge here is that typical conformal prediction methods—which give marginal validity (coverage) guarantees—provide uneven coverage, in that they address easy examples at the expense of essentially ignoring difficult examples. By leveraging ideas from quantile regression, we build methods that always guarantee correct coverage but additionally provide (asymptotically consistent) conditional coverage for both multiclass and multilabel prediction problems. To address the potential challenge of exponentially large confidence sets in multilabel prediction, we build tree-structured classifiers that efficiently account for interactions between labels. Our methods can be bolted on top of any classification model—neural network, random forest, boosted tree—to guarantee its validity. We also provide an empirical evaluation, simultaneously providing new validation methods, that suggests the more robust coverage of our confidence sets.

***K*eywords** Multilabel and Multiclass Classification, Conformal Inference, Validity, Quantile Regression, Graphical Models

1. Introduction

The average accuracy of a machine-learned model by itself is insufficient to trust the model's application; instead, we should ask for valid confidence in its predictions. Valid here does not mean "valid under modeling assumptions," or "trained to predict confidence," but honest validity, independent of the underlying distribution. In particular, for a supervised learning task with inputs $x \in \mathcal{X}$, targets $y \in \mathcal{Y}$, and a given confidence level $\alpha \in (0, 1)$, we seek confidence sets $C(x)$ such that $P(Y \in C(X)) \geq 1 - \alpha$; that is, we *cover* the true target

Y with a given probability $1-\alpha$. Given the growing importance of statistical learning in real-world applications—autonomous vehicles (Kalra and Paddock, 2016), skin lesion identification (Esteva et al., 2017; Oakden-Rayner et al., 2020), loan repayment prediction (Hardt et al., 2016)—such validity is essential.

The typical approach in supervised learning is to learn a scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ where high scores $s(x, y)$ mean that y is more likely for a given x. Potential examples include any strictly increasing function of the conditional probability $P(Y = y|X = x)$ in a classification problem, but also more generally $-\ell(\widehat{y}(x), y)$, where $\widehat{y} : \mathcal{X} \to \mathcal{Y}$ is a predictor, and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function measuring the distance between the prediction and the true response. Conceptually, a "good" score function should rank the elements of $\mathcal{Y}$ by decreasing likelihood conditionally on x, but the coverage validity of the resulting confidence sets should not depend upon it. Given such a score, a natural goal for prediction with confidence is to compute a quantile function q_α satisfying

$$P\left(s(x, Y) \geq q_\alpha(x) \mid X = x\right) \geq 1 - \alpha, \tag{1}$$

where $\alpha > 0$ is some *a-priori* acceptable error level. We could then output conditionally valid confidence sets for each $x \in \mathcal{X}$ at level $1 - \alpha$ by returning

$$\left\{y \in \mathcal{Y} \mid s(x, y) \geq q_\alpha(x)\right\}.$$

Unfortunately, such conditional coverage is impossible without either vacuously loose thresholds (Vovk, 2012) or strong modeling assumptions (Vovk, 2012; Barber et al., 2019), but this idea forms the basis for our approach.

To address this impossibility, *conformal inference* (Vovk et al., 2005) resolves instead to a marginal coverage guarantee: given n observations and a desired confidence level $1 - \alpha$, conformal methods construct confidence sets $C(x)$ such that for a new pair (X_{n+1}, Y_{n+1}) from the same distribution, $Y_{n+1} \in C(X_{n+1})$ with probability at least $1 - \alpha$, where the probability is *jointly* over X and Y. Conformal inference algorithms can build upon arbitrary predictors, neural networks, random forests, kernel methods and treat them as black boxes, "conformalizing" them post-hoc.

This distribution-free coverage is only achievable marginally, and standard conformal predictions for classification achieve it (as we see later) by providing good coverage on easy examples at the expense of miscoverage on harder instances. We wish to provide more uniform coverage, and we address this in both multiclass—where each example belongs to a single class—and multilabel—where each example may belong to several classes—classification problems. We combine the ideas of conformal prediction (Vovk et al., 2005) with an approach to fit a quantile function q on the scores $s(x, y)$ of the prediction model, which Romano et al. (2019) originate for regression problems, and build feature-adaptive quantile predictors that output sets of labels, allowing us to guarantee valid marginal coverage (independent of the data generating distribution) while better approximating the conditional coverage (1). A challenge is to evaluate whether we indeed do provide better than marginal coverage, so we provide new validation methodology to test this as well.

1.1 Conformal inference in classification

For multiclass problems, we propose a method that fits a quantile function on the scores, conformalizing it on held-out data. While this immediately provides valid marginal coverage,

the accuracy of the quantile function—how well it approximates the conditional quantiles—determines conditional coverage performance. Under certain consistency assumptions on the learned scoring functions and quantiles as sample size increases, we show in Section 2 that we recover the conditional coverage (1) asymptotically.

The multilabel case poses new statistical and computational challenges, as a K-class problem entails 2^K potential responses. In this case, we seek efficiently representable inner and outer sets $C_{\text{in}}(x)$ and $C_{\text{out}}(x) \subset \{1, \ldots, K\}$ such that

$$\mathbb{P}(C_{\text{in}}(X) \subset Y \subset C_{\text{out}}(X)) \geq 1 - \alpha. \tag{2}$$

We propose two approaches to guarantee the containments (2). The first directly fits inner and outer sets by solving two separate quantile regression problems. The second begins by observing that labels are frequently correlated—think, for example, of chairs, which frequently co-occur with a table—and learns a tree-structured graphical model (Koller and Friedman, 2009) to efficiently address such correlation. We show how to build these on top of any predictive model. In an extension when the sets (2) provide too imprecise confidence sets, we show how to also construct a small number of sets $C^{(i)}_{\text{in/out}}$ that similarly satisfy $\mathbb{P}(\cup_i\{C^{(i)}_{\text{in}}(X) \subset Y \subset C^{(i)}_{\text{out}}(X)) \geq 1 - \alpha$ while guaranteeing $C_{\text{in}}(x) \subset \cup_i C^{(i)}_{\text{in}}(x)$ and $C_{\text{out}}(x) \supset \cup_i C^{(i)}_{\text{out}}(x)$.

1.2 Related work and background

Vovk et al. (2005) introduce (split-)conformal inference, which splits the first n samples of the exchangeable pairs $\{(X_i, Y_i)\}_{i=1}^{n+1}$ into two sets (say, each of size n_1 and n_2 respectively) where the first training set ($\mathcal{I}_1$) is used to learn a scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and the second validation set ($\mathcal{I}_2$) to "conformalize" the scoring function and construct a confidence set over potential labels (or targets) $\mathcal{Y}$ of the form

$$C(x) \coloneqq \{y \in \mathcal{Y} \mid s(x, y) \geq t\}$$

for some threshold t. The basic split-conformal method chooses the $(1 + 1/n_2)(1 - \alpha)$-empirical quantile $\hat{Q}^{\text{marg}}_{1-\alpha}$ of the negative scores $\{-s(X_i, Y_i)\}_{i\in\mathcal{I}_2}$ (on the validation set) and defines

$$C(x) \coloneqq \{y \in \mathcal{Y} \mid s(x, y) \geq -\hat{Q}^{\text{marg}}_{1-\alpha}\}. \tag{3}$$

The argument that these sets C provide valid coverage is beautifully simple and is extensible given any scoring function s: letting $S_i = -s(X_i, Y_i)$, if $\hat{Q}^{\text{marg}}_{1-\alpha}$ is the $(1 + 1/n_2)\,(1 - \alpha)$-quantile of $\{S_i\}_{i\in\mathcal{I}_2}$ and the pairs $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, we have

$$\begin{aligned}\mathbb{P}(Y_{n+1} \in C(X_{n+1})) &= \mathbb{P}(s(X_{n+1}, Y_{n+1}) \geq -\hat{Q}^{\text{marg}}_{1-\alpha}) \\ &= \mathbb{P}\left(\text{Rank of } S_{n+1} \text{ in } \{S_i\}_{i\in\mathcal{I}_2\cup\{n+1\}} \leq \lceil (n_2 + 1)(1 - \alpha)\rceil\right) \geq 1 - \alpha.\end{aligned}$$

We refer to the procedure (3) as the *Marginal* conformal prediction method. Such a "conformalization" scheme typically outputs a confidence set by listing all the labels that it contains, which is feasible in a K-class multiclass problem, but more challenging in a K-class multilabel one, as the number of configurations (2^K) grows exponentially.

While conformal inference guarantees marginal coverage without assumption on the distribution generating the data, Vovk (2012) shows it is virtually impossible to attain distribution-free conditional coverage, and Barber et al. (2019) prove that in regression, one can only achieve a weaker form of approximately-conditional coverage without conditions on the underlying distribution. Because of this theoretical limitation, work in conformal inference often focuses on minimizing confidence set sizes or guaranteeing different types of coverage. For instance, Sadinle et al. (2019) propose conformal prediction algorithms for multiclass problems that minimize the expected size of the confidence set and conformalize the scores separately for each class, providing class-wise coverage. In the same vein, Hechtlinger et al. (2019) use density estimates of $p(x \mid y)$ as conformal scores to build a "cautious" predictor, the idea being that it should output an empty set when the new sample differs too much from the original distribution. A more recent approach—published as we were completing the original draft of the current paper—are Izbicki et al.'s distribution-split and conditional distribution-splitting methods (2020), which attempt to achieve conditional coverage in a regression problem (predicting targets $y \in \mathbb{R}$) by using quantiles of a density estimator of $y \mid x$ fit on partitions of the feature space $\mathcal{X}$.

Several papers also develop conformal inference methods for multi-label problems, though their foci are often distinct from ours. Papadopoulos (2014) proposes a method that builds non-conformity scores that weight interaction terms between labels—as our scoring functions using trees in Section 3.3 do—but uses the implied confidence set only implicitly, so that actually representing the set appears computationally challenging; Paisios et al. (2019) develop the method in the context of multi-label text classification. In other work, Lambrou and Papadopoulos (2016), Wang et al. (2014), and Wang et al. (2015) develop methods for multi-label classification, which they term the Binary Relevance Multi-Label Conformal Predictor (MLCP) and the Instance Reproduction MLCP (the latter of which empirically returns tighter confidence sets), which use a non-conformity score akin to the minimum score $s(x, y)$ of a positive label. These are similar to our more naive conformalized inner/outer method (Alg. 3), in that they do not correct for interactions or structure in the labels, though they return the equivalent of "outer sets" in our framework. In contrast, we also wish to identify labels that *should* be present, so that we develop a type of inner/outer nested confidence sets, and our methods also incorporate interactions between labels in ways that are both computationally efficient and explicitly representable, which appears to not be present in current literature. In work building off of the initial post of this paper to the `arXiv`, Romano et al. (2020) construct conformal confidence sets for multi-class problems by leveraging pivotal p-value-like quantities, which provide conditional coverage when models are well-specified. In regression problems, Romano et al. (2019) conformalize a quantile predictor, which allows them to build marginally valid confidence sets that are adaptive to feature heterogeneity and empirically smaller on average than purely marginal confidence sets. We adapt this regression approach to classification tasks, learning quantile functions to construct valid—potentially conditionally valid—confidence predictions.

1.3 Notation

$\mathcal{P}$ is a set of distributions on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{-1, 1\}^K$ is a discrete set of labels, $K \geq 2$ is the number of classes, and $\mathcal{X} \subset \mathbb{R}^d$. We assume that we observe a finite sequence

$\{(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} P$ from some distribution $P = P_X \otimes P_{Y|X} \in \mathcal{P}$, where P_X denotes the marginal distribution over X and $P_{Y|X}$ is the Markov kernel describing the conditional distribution of Y given X. We wish to predict a confidence set for a new example $X_{n+1} \sim P_X$. $\mathbb{P}$ stands over the randomness of both the new sample (X_{n+1}, Y_{n+1}) and the full procedure. We tacitly identify the vector $Y \in \mathcal{Y} = \{-1, 1\}^K$ with the subset of $[K] = \{1, \ldots, K\}$ it represents, the notation $\mathcal{X} \rightrightarrows [K]$ indicates a set-valued mapping between $\mathcal{X}$ and $[K]$, and $\preceq$ is the partial order induced by the nonnegative orthant $\mathbb{R}_+^K$. We define $\|f\|_{L^2(P)}^2 \coloneqq \int f^2 dP$.

2. Conformal multiclass classification

Algorithm 1: Split Conformalized Quantile Classification (CQC).

Input: Sample $\{(X_i, Y_i)\}_{i=1}^n$, index sets $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3 \subset [n]$, fitting algorithm $\mathcal{A}$, quantile functions $\mathcal{Q}$, and desired confidence level α

1. Fit scoring function via
$$\widehat{s} \coloneqq \mathcal{A}\left((X_i, Y_i)_{i \in \mathcal{I}_1}\right). \tag{4}$$

2. Fit quantile function via
$$\widehat{q}_\alpha \in \operatorname*{argmin}_{q \in \mathcal{Q}} \left\{ \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \rho_\alpha\left(\widehat{s}(X_i, Y_i) - q(X_i)\right) \right\} \tag{5}$$

3. Calibrate by computing non-conformity scores $S_i = \widehat{q}_\alpha(X_i) - \widehat{s}(X_i, Y_i)$, defining
$$Q_{1-\alpha}(S, \mathcal{I}_3) \coloneqq (1 + 1/n_3)(1-\alpha) \text{ empirical quantile of } \{S_i\}_{i \in \mathcal{I}_3}$$
and return prediction set function
$$\widehat{C}_{n,1-\alpha}(x) \coloneqq \left\{k \in [K] \mid \widehat{s}(x, k) \geq \widehat{q}_\alpha(x) - Q_{1-\alpha}(S, \mathcal{I}_3)\right\}. \tag{6}$$

We begin with multiclass classification problems, developing *Conformalized Quantile Classification* (CQC) to construct finite sample marginally valid confidence sets. CQC is similar to Romano et al. (2019)'s Conformalized Quantile Regression (CQR): we estimate a quantile function of the scores, which we use to construct valid confidence sets after conformalization. We split the data into subsets $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$ with sample sizes n_1, n_2 and n_3, where $\mathcal{I}_3$ is disjoint from $\mathcal{I}_1 \cup \mathcal{I}_2$. Algorithm 1 outlines the basic idea: we use the set $\mathcal{I}_1$ for fitting the scoring function with an (arbitrary) learning algorithm $\mathcal{A}$, use $\mathcal{I}_2$ for fitting a quantile function from a family $\mathcal{Q} \subset \mathcal{X} \to \mathbb{R}$ of possible quantile functions to the resulting scores, and use $\mathcal{I}_3$ for calibration. In the algorithm, we recall the "pinball loss" (Koenker and Bassett Jr., 1978) $\rho_\alpha(t) = (1-\alpha)\left[-t\right]_+ + \alpha\left[t\right]_+$, which satisfies $\operatorname{argmin}_{q \in \mathbb{R}} \mathbb{E}[\rho_\alpha(Z - q)] = \inf\{q \mid \mathbb{P}(Z \leq q) \geq \alpha\}$ for any random variable Z.

2.1 Finite sample validity of CQC

We begin by showing that Alg. 1 enjoys the coverage guarantees we expect, which is more or less immediate by the marginal guarantees for the method (3). We include the proof for completeness and because its cleanness highlights the ease of achieving validity.

None of our guarantees for Alg. 1 explicitly requires that we fit the scoring function and the quantile function on disjoint subsets $\mathcal{I}_1$ and $\mathcal{I}_2$.

We assume for simplicity that the full sample $\{(X_i, Y_i)\}_{i=1}^{n+1}$ is exchangeable, though we only require the sample $\{(X_i, Y_i)\}_{i\in\mathcal{I}_3\cup\{n+1\}}$ be exchangeable conditionally on $\{(X_i, Y_i)\}_{i\in\mathcal{I}_1\cup\mathcal{I}_2}$. In general, so long as the score $\widehat{s}$ and quantile $\widehat{q}_{1-\alpha}$ functions are measurable with respect to $\{(X_i, Y_i)\}_{i\in\mathcal{I}_1\cup\mathcal{I}_2}$ and a σ-field independent of $\{(X_i, Y_i)\}_{i\in\mathcal{I}_3\cup\{n+1\}}$, then Theorem 1 remains valid if the exchangeability assumption holds for the instances in $\mathcal{I}_3 \cup \{n+1\}$.

Theorem 1. *Assume that $\{(X_i, Y_i)\}_{i=1}^{n+1} \sim P$ are exchangeable, where P is an arbitrary distribution. Then the prediction set $\widehat{C}_{n,1-\alpha}$ of the split CQC Algorithm 1 satisfies*

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

Proof The argument is due to Romano et al. (2019, Thm. 1). Observe that $Y_{n+1} \in \widehat{C}_{n,1-\alpha}(X_{n+1})$ if and only if $S_{n+1} \leq Q_{1-\alpha}(S, \mathcal{I}_3)$. Define the σ-field $\mathcal{F}_{12} = \sigma\left\{(X_i, Y_i) \mid i \in \mathcal{I}_1 \cup \mathcal{I}_2\right\}$. Then

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_{n,1-\alpha}(X_{n+1}) \mid \mathcal{F}_{12}) = \mathbb{P}(S_{n+1} \leq Q_{1-\alpha}(S, \mathcal{I}_3) \mid \mathcal{F}_{12}).$$

We use the following lemma.

Lemma 2 (Lemma 2, Romano et al. (2019))**.** *Let $Z_1, \ldots, Z_{n+1}$ be exchangeable random variables and $\widehat{Q}_n(\cdot)$ be the empirical quantile function of $Z_1, \ldots Z_n$. Then for any $\alpha \in (0, 1)$,*

$$\mathbb{P}\big(Z_{n+1} \leq \widehat{Q}_n((1 + n^{-1})\alpha)\big) \geq \alpha.$$

If $Z_1, \ldots, Z_n$ are almost surely distinct, then

$$\mathbb{P}\big(Z_{n+1} \leq \widehat{Q}_n((1 + n^{-1})\alpha)\big) \leq \alpha + \frac{1}{n}.$$

As the original sample is exchangeable, so are the non-conformity scores S_i for $i \in \mathcal{I}_3$, conditionally on $\mathcal{F}_{12}$. Lemma 2 implies $\mathbb{P}(S_{n+1} \leq Q_{1-\alpha}(S, \mathcal{I}_3) \mid \mathcal{F}_{12}) \geq 1 - \alpha$, and taking expectations over $\mathcal{F}_{12}$ yields the theorem. □

The conditional distribution of scores given X is discrete, so the confidence set may be conservative: it is possible that for any q such that $P(s(X, Y) \geq q \mid X) \geq 1 - \alpha$, we have $P(s(X, Y) \geq q \mid X) \geq 1 - \epsilon$ for some $\epsilon \ll \alpha$. Conversely, as the CQC procedure 1 seeks $1-\alpha$ marginal coverage, it may sacrifice a few examples to bring the coverage down to $1 - \alpha$. Moreover, there may be no unique quantile function for the scores. One way to address these issues is to estimate a quantile function on $\mathcal{I}_2$ (recall step (5) of the CQC method) so that we can guarantee higher coverage (which is more conservative, but is free in the $\epsilon \ll \alpha$ case).

An alternative to achieve exact $1-\alpha$ asymptotic coverage and a unique quantile, which we outline here, is to randomize scores without changing the relative order. Let $Z_i \overset{\text{iid}}{\sim} \pi$ for some distribution π with continuous density supported on the entire real line, and let $\sigma > 0$ be a noise parameter. Then for any scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we define (for a given $x \in \mathcal{X}$, $z \in \mathbb{R}$ and any $y \in \mathcal{Y}$)

$$s^\sigma(x, y, z) \coloneqq s(x, y) + \sigma z.$$

As $s^\sigma(x, y, z) - s^\sigma(x, y', z) = s(x, y) - s(x, y')$, this maintains the ordering of label scores, only giving the score function a conditional density. Further, this mechanism adapts to any choice of score function, and in particular does not make any assumption on the original model s. Now, consider replacing the quantile estimator (5) with the randomized estimator

$$\widehat{q}^\sigma_\alpha \in \operatorname*{argmin}_{q \in \mathcal{Q}} \left\{ \frac{1}{n_2} \sum_{i \in \mathcal{I}_2} \rho_\alpha(\widehat{s}^\sigma(X_i, Y_i, Z_i) - q(X_i)) \right\}, \tag{7}$$

and let $S^\sigma_i \coloneqq \widehat{q}^\sigma_\alpha(X_i) - \widehat{s}^\sigma(X_i, Y_i, Z_i)$ be the corresponding non-conformity scores. Similarly, replace the prediction set (6) with

$$\widehat{C}^\sigma_{n,1-\alpha}(x, z) \coloneqq \{k \in [K] \mid \widehat{s}^\sigma(x, k, z) \geq \widehat{q}^\sigma_\alpha(x) - Q_{1-\alpha}(S^\sigma, \mathcal{I}_3)\}. \tag{8}$$

where $Q_{1-\alpha}(S^\sigma, \mathcal{I}_3)$ is the $(1-\alpha)(1+1/n_3)$-th empirical quantile of $\{S^\sigma_i\}_{i \in \mathcal{I}_3}$. Then for a new input $X_{n+1} \in \mathcal{X}$, we simulate a new independent variable $Z_{n+1} \sim \pi$, and give the confidence set $\widehat{C}^\sigma_{n,1-\alpha}(X_{n+1}, Z_{n+1})$.

As the next result shows, this gives nearly perfectly calibrated coverage.

Corollary 3. *Assume that $\{(X_i, Y_i)\}_{i=1}^{n+1} \sim P$ are exchangeable, where P is an arbitrary distribution. Let the estimators* (7) *and* (8) *replace the estimators* (5) *and* (6) *in the CQC Algorithm 1, respectively. Then the prediction set $\widehat{C}^\sigma_{n,1-\alpha}$ satisfies*

$$1 - \alpha \leq \mathbb{P}\{Y_{n+1} \in \widehat{C}^\sigma_{n,1-\alpha}(X_{n+1}, Z_{n+1})\} \leq 1 - \alpha + \frac{1}{1 + n_3}.$$

Proof The argument is identical to that for Theorem 1, except that we apply the second part of Lemma 2 to achieve the upper bound, as the scores are a.s. distinct. □

Remark 4. *The upper bound in Corollary 3 is tight and unavoidable for any smoothed conformal predictor and its split counterpart due to finite sample. Exact $1-\alpha$ coverage can only be achieved in the limit as we have growing number of samples in the calibration set.*

2.2 Asymptotic consistency of CQC method

Under appropriate assumptions typical in proving the consistency of prediction methods, Conformalized Quantile Classification (CQC) guarantees conditional coverage asymptotically. To set the stage, assume that as $n \uparrow \infty$, the fit score functions $\widehat{s}_n$ in Eq. (4) converge to a fixed $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ (cf. Assumption 1). Let

$$q_\alpha(x) \coloneqq \inf \{z \in \mathbb{R} \mid \alpha \leq P(s(x, Y) \leq z \mid X = x)\}$$

be the α-quantile function of the limiting scores $s(x, Y)$ and for $\sigma > 0$ define

$$q_\alpha^\sigma(x) \coloneqq \inf \left\{ z \in \mathbb{R} \mid \alpha \le P\left(s^\sigma(x, Y, Z) \le z \mid X = x\right) \right\},$$

where $Z \sim \pi$ (for a continuous distribution π as in the preceding section) is independent of x, y, to be the α-quantile function of the noisy scores $s^\sigma(x, Y, Z) = s(x, Y) + \sigma Z$. With these, we can make the following natural definitions of our desired asymptotic confidence sets.

Definition 5. *The* randomized-oracle *and* super-oracle *confidence sets are* $C_{1-\alpha}^\sigma(X, Z) \coloneqq \{k \in [K] \mid s^\sigma(X, k, Z) \ge q_\alpha^\sigma(X)\}$ *and* $C_{1-\alpha}(X) = \{k \in [K] \mid s(X, k) \ge q_\alpha(X\}$*, respectively.*

Our aim will be to show that the confidence sets of the split CQC method 1 (or its randomized variant) converge to these confidence sets under appropriate consistency conditions.

To that end, we consider the following consistency assumption.

Assumption 1 (Consistency of scores and quantile functions)**.** *The score functions* $\widehat{s}$ *and quantile estimator* $\widehat{q}_\alpha^\sigma$ *are mean-square consistent, so that as* $n_1, n_2, n_3 \to \infty$*,*

$$\|\widehat{s} - s\|_{L^2(P_X)}^2 \coloneqq \int_{\mathcal{X}} \|\widehat{s}(x, \cdot) - s(x, \cdot)\|_\infty^2 dP_X(x) \xrightarrow{p} 0$$

and

$$\|\widehat{q}_\alpha^\sigma - q_\alpha^\sigma\|_{L^2(P_X)}^2 \coloneqq \int_{\mathcal{X}} (\widehat{q}_\alpha^\sigma(x) - q_\alpha^\sigma(x))^2 dP_X(x) \xrightarrow{p} 0.$$

Assumption 1 states that the score and quantile estimators converge in mean-squared error under P_X towards the population counterparts . Precisely, $\|\widehat{s} - s\|_{L^2(P_X)}^2$ and $\|\widehat{q}_\alpha^\sigma - q_\alpha^\sigma\|_{L^2(P_X)}^2$ are random variables that depend on $\{(X_i, Y_i)\}_{i=1}^n$ through the random functions $\widehat{s}$ and $\widehat{q}$, and measure their distance in $L^2(P_X)$ to s and q respectively. Note that, for a fixed $x \in \mathcal{X}$, we have $s(x, \cdot) \in \mathbb{R}^K$, and by convenience we use the ℓ_∞-norm to measure the pointwise distance between $\widehat{s}(x, \cdot)$ and $s(x, \cdot)$. With this assumption, we have the following theorem, whose proof we provide in Appendix A.1.

Theorem 6. *Let Assumption 1 hold. Then the confidence sets* $\widehat{C}_{n,1-\alpha}^\sigma$ *satisfy*

$$\lim_{n \to \infty} \mathbb{P}(\widehat{C}_{n,1-\alpha}^\sigma(X_{n+1}, Z_{n+1}) \ne C_{1-\alpha}^\sigma(X_{n+1}, Z_{n+1})) = 0.$$

Unlike other work (Sadinle et al., 2019; Romano et al., 2020) in conformal inference, validity of Theorem 6 does not rely on the scores being consistent for the log-conditional probabilities. Instead, we require the $(1 - \alpha)$-quantile function to be consistent for the scoring function at hand, which is weaker (though still a strong assumption). Of course, the ideal scenario occurs when the limiting score function s is optimal (cf. Bartlett et al., 2006; Zhang, 2004; Tewari and Bartlett, 2007), so that $s(x, y) > s(x, y')$ whenever $P(Y = y \mid X = x) > P(Y = y' \mid X = x)$. Under this additional condition, the super-oracle confidence set $C_{1-\alpha}(X)$ in Def. 5 is the smallest conditionally valid confidence set at level $1-\alpha$, and in that sense is the optimal $1-\alpha$ confidence set. Conveniently, our randomization is consistent as $\sigma \downarrow 0$: the super oracle confidence set $C_{1-\alpha}$ contains $C_{1-\alpha}^\sigma$ (with high probability). We provide the proof of the following result in Appendix A.2.

Proposition 7. *The confidence sets $C^\sigma_{1-\alpha}$ satisfy*

$$\lim_{\sigma \to 0} \mathbb{P}(C^\sigma_{1-\alpha}(X_{n+1}, Z_{n+1}) \subseteq C_{1-\alpha}(X_{n+1})) = 1.$$

Because we always have $\mathbb{P}(Y \in C^\sigma_{1-\alpha}(X, Z)) \geq 1 - \alpha$, Proposition 7 shows that we maintain validity while potentially shrinking the confidence sets.

Remark 8. *Since the confidence sets $C^\sigma_{1-\alpha}(X, Z)$ are randomized and thus often smaller than their deterministic counterparts, there is an increased probability of returning an empty confidence set, especially for small values of Z. Such situation can be confusing for the practitioner, but an easy workaround to avoid it is to return instead the singleton $\{y\} \subset \mathcal{Y}$ that maximizes $y' \mapsto s(x, y')$, which can only improve coverage.*

3. The multilabel setting

In multilabel classification, we observe a single feature vector $x \in \mathcal{X}$ and wish to predict a vector $y \in \{-1, 1\}^K$ where $y_k = 1$ indicates that label k is present and $y_k = -1$ indicates its absence. For example, in object detection problems (Redmon et al., 2016; Tan et al., 2019; Lin et al., 2014), we wish to detect several entities in an image, while in text classification tasks (Lodhi et al., 2002; McCallum and Nigam, 1998; Joulin et al., 2016), a single document can potentially share multiple topics. To conformalize such predictions, we wish to output an aggregated set of predictions $\{\widehat{y}(x)\} \subset \{-1, 1\}^K$—a collection of -1-1-valued vectors—that contains the true configuration $Y = (Y_1, \ldots, Y_K)$ with probability at least $1 - \alpha$.

The multilabel setting poses statistical and computational challenges. On the statistical side, multilabel prediction engenders a multiple testing challenge: even if each task has an individual confidence set $C_k(x)$ such that $P(Y_k \in C_k(X)) \geq 1 - \alpha$, in general we can only conclude that $P\left(Y \in C_1(X) \times \cdots \times C_K(X)\right) \geq 1 - K\alpha$; as all predictions share the same features x, we wish to leverage correlation through x. Additionally, as we discuss in the introduction (recall Eq. (3)), we require a scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Given a predictor $\widehat{y} : \mathcal{X} \to \{-1, 1\}^k$, a naive scoring function for multilabel problems is to use $s(x, y) = 1\{\widehat{y}(x) = y\}$, but this fails, as the confidence sets contain either all configurations or a single configuration. On the computational side, the total number of label configurations (2^K) grows exponentially, so a "standard" multiclass-like approach outputting confidence sets $\widehat{C}(X) \subset \mathcal{Y} = \{-1, 1\}^K$, although feasible for small values of K, is computationally impractical even for moderate K.

We instead propose using inner and outer set functions $\widehat{C}_{\text{in}}, \widehat{C}_{\text{out}} : \mathcal{X} \rightrightarrows [K]$ to efficiently describe a confidence set on $\mathcal{Y}$, where we require they satisfy the coverage guarantee

$$\mathbb{P}\left(\widehat{C}_{\text{in}}(X) \subset Y \subset \widehat{C}_{\text{out}}(X)\right) \geq 1 - \alpha, \tag{9a}$$

or equivalently, we learn two functions $\widehat{y}_{\text{in}}, \widehat{y}_{\text{out}} : \mathcal{X} \to \mathcal{Y} = \{-1, 1\}^K$ such that

$$\mathbb{P}\left(\widehat{y}_{\text{in}}(X) \preceq Y \preceq \widehat{y}_{\text{out}}(X)\right) \geq 1 - \alpha. \tag{9b}$$

We thus say that coverage is *valid* if the inner set exclusively contains positive labels and the outer set contains at least all positive labels. For example, $\widehat{y}_{\text{in}}(x) = \mathbf{0}$ and $\widehat{y}_{\text{out}}(x) = \mathbf{1}$

are always valid, though uninformative, while the smaller the set difference between $\widehat{C}_{\text{in}}(X)$ and $\widehat{C}_{\text{out}}(X)$ the more confident we may be in a single prediction. As we mention in the introduction, we extend the inner/outer coverage guarantees (9) to construct to unions of such rectangles to allow more nuanced coverage; see Sec. 3.1.1.

In the remainder of this section, we propose methods to conformalize multilabel predictors in varying generality, using tree-structured graphical models to both address correlations among labels and computational efficiency. We begin in Section 3.1 with a general method for conformalizing an arbitrary scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ on multilabel vectors, which guarantees validity no matter the score. We then show different strategies to efficiently implement the method, depending on the structure of available scores, in Section 3.2, showing how "tree-structured" scores allow computational efficiency while modeling correlations among the task labels y. Finally, in Section 3.3, we show how to build such a tree-structured score function s from both arbitrary predictive models (e.g. Boutell et al., 2004; Zhang and Zhou, 2006) and those more common multilabel predictors—in particular, those based on neural networks—that learn and output per-task scores $s_k : \mathcal{X} \to \mathbb{R}$ for each task (Herrera et al., 2016; Read et al., 2011).

3.1 A generic split-conformal method for multilabel sets

We begin by assuming we have a general score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for $\mathcal{Y} = \{-1, 1\}^K$ that evaluates the quality of a given set of labels $y \in \mathcal{Y}$ for an instance x; in the next subsection, we describe how to construct such scores from multilabel prediction methods, presenting our general method first. We consider the variant of the CQC method 1 in Alg. 2.

Algorithm 2: Split Conformalized Inner/Outer method for classification (CQioC)

Input: Sample $\{(X_i, Y_i)\}_{i=1}^n$, disjoint index sets $\mathcal{I}_2, \mathcal{I}_3 \subset [n]$, quantile functions $\mathcal{Q}$, desired confidence level α, and score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

1. Fit quantile function $\widehat{q}_\alpha \in \operatorname{argmin}_{q \in \mathcal{Q}} \{\sum_{i \in \mathcal{I}_2} \rho_\alpha(s(X_i, Y_i) - q(X_i))\}$, as in (5).

2. Compute non-conformity scores $S_i = \widehat{q}_\alpha(X_i) - s(X_i, Y_i)$, define $\widehat{Q}_{1-\alpha}$ as $(1 + 1/|\mathcal{I}_3|) \cdot (1 - \alpha)$ empirical quantile of $S = \{S_i\}_{i \in \mathcal{I}_3}$, and return prediction set function
$$\widehat{C}_{\text{io}}(x) := \{y \in \mathcal{Y} \mid \widehat{y}_{\text{in}}(x) \preceq y \preceq \widehat{y}_{\text{out}}(x)\}, \tag{10}$$
where $\widehat{y}_{\text{in}}$ and $\widehat{y}_{\text{out}}$ satisfy
$$\begin{aligned} \widehat{y}_{\text{in}}(x)_k = 1 \text{ if and only if } \max_{y \in \mathcal{Y}: y_k = -1} s(x, y) < \widehat{q}_\alpha(x) - \widehat{Q}_{1-\alpha} \\ \widehat{y}_{\text{out}}(x)_k = -1 \text{ if and only if } \max_{y \in \mathcal{Y}: y_k = 1} s(x, y) < \widehat{q}_\alpha(x) - \widehat{Q}_{1-\alpha}. \end{aligned} \tag{11}$$

There are two considerations for Algorithm 2: its computational efficiency and its validity guarantees. Deferring the efficiency questions to the coming subsections, we begin with the latter. The naive approach is to simply use the "standard" or implicit conformalization approach, used for regression or classification, by defining
$$\widehat{C}_{\text{imp}}(x) := \left\{y \in \mathcal{Y} \mid s(x, y) \geq \widehat{q}_\alpha(x) - \widehat{Q}_{1-\alpha}\right\}, \tag{12}$$

where $\widehat{q}_\alpha$ and $\widehat{Q}_{1-\alpha}$ are as in the CQioC method 2. This does guarantee validity, as we have the following corollary of Theorem 1.

Corollary 9. *Assume* $\{(X_i, Y_i)\}_{i=1}^{n+1} \stackrel{\text{iid}}{\sim} P$*, where* P *is an arbitrary distribution. Then for any confidence set* $\widehat{C} : \mathcal{X} \rightrightarrows \mathcal{Y}$ *satisfying* $\widehat{C}(x) \supset \widehat{C}_{\text{imp}}(x)$ *for all* x*,*

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

Instead of the inner/outer set $\widehat{C}_{\text{io}}$ in Eq. (10), we could use any confidence set $\widehat{C}(x) \supset \widehat{C}_{\text{imp}}(x)$ and maintain validity. Unfortunately, as we note above, the set $\widehat{C}_{\text{imp}}$ may be exponentially complex to represent and compute, necessitating a simplifying construction, such as our inner/outer approach. Conveniently, the set (10) we construct via $\widehat{y}_{\text{in}}$ and $\widehat{y}_{\text{out}}$ satisfying the conditions (11) satisfies Corollary 9. Indeed, we have the following proposition.

Proposition 10. *Let* $\widehat{y}_{\text{in}}$ *and* $\widehat{y}_{\text{out}}$ *satisfy the conditions* (11)*. Then the confidence set* $\widehat{C}_{\text{io}}(x)$ *in Algorithm 2 is the smallest set containing* $\widehat{C}_{\text{imp}}(x)$ *and admitting the form* (10)*.*

Proof The conditions (11) immediately imply

$$\widehat{y}_{\text{in}}(x)_k = \min_{y \in \widehat{C}_{\text{imp}}(x)} y_k \quad \text{and} \quad \widehat{y}_{\text{out}}(x)_k = \max_{y \in \widehat{C}_{\text{imp}}(x)} y_k,$$

which shows that $\widehat{C}_{\text{imp}}(x) \subset \widehat{C}_{\text{io}}(x)$. On the other hand, suppose that $\tilde{y}_{\text{in}}(x)$ and $\tilde{y}_{\text{out}}(x)$ are configurations inducing a confidence set $\tilde{C}(x)$ that satisfies $\widehat{C}_{\text{imp}}(x) \subset \tilde{C}(x)$. Then for any label k included in $\tilde{y}_{\text{in}}(x)$, all configurations $y \in \widehat{C}_{\text{imp}}(x)$ satisfy $y_k = 1$, as $\widehat{C}_{\text{imp}}(x) \subset \tilde{C}(x)$, so we must have $\tilde{y}_{\text{in}}(x)_k \leq \widehat{y}_{\text{in}}(x)_k$. The argument to prove $\tilde{y}_{\text{out}}(x)_k \geq \widehat{y}_{\text{out}}(x)_k$ is similar. □

The expansion from $\widehat{C}_{\text{imp}}(x)$ to $\widehat{C}_{\text{io}}(x)$ may increase the size of the confidence set, most notably in cases when labels repel each other. As a paradoxical worst-case, if $\widehat{C}_{\text{imp}}(x)$ includes only each of the K single-label configurations then $\widehat{C}_{\text{io}} = \{-1, 1\}^K$. In such cases, refinement of the inner/outer sets may be necessary; we outline an approach that considers unions of such sets in Sec. 3.1.1 to come. Yet in problems for which we have strong predictors, we typically do not expect "opposing" configurations y and $-y$ to both belong to $\widehat{C}_{\text{io}}(x)$, which limits the increase in practice; moreover, in the case that the standard implicit confidence set $\widehat{C}_{\text{imp}}$ is a singleton, there is a single $y \in \mathcal{Y}$ satisfying $s(x, y) \geq \widehat{q}_\alpha(x) - \widehat{Q}_{1-\alpha}$ and by definition $\widehat{y}_{\text{in}}(x) = \widehat{y}_{\text{out}}(x)$, so that $\widehat{C}_{\text{io}} = \widehat{C}_{\text{imp}}$.

3.1.1 Unions of inner and outer sets

As we note above, it can be beneficial to approximate the implicit set (12) more carefully; here, we consider a union of easily representable sets. The idea is that if two tasks always have opposing labels, the confidence sets should reflect that, yet it is possible that the naive condition (11) fails this check. For example, consider a set $\widehat{C}_{\text{imp}}(x)$ for which any configuration $y \in \widehat{C}_{\text{imp}}(x)$ satisfies $y_1 = -y_2$, but which contains labels with both $y_1 = 1$ and $y_1 = -1$. In this case, necessarily $\widehat{y}_{\text{in}}(x)_1 = \widehat{y}_{\text{in}}(x)_2 = -1$ and $\widehat{y}_{\text{out}}(x)_1 = \widehat{y}_{\text{out}}(x)_2 = 1$. If instead we construct two sets of inner and outer vectors $\widehat{y}_{\text{in}}^{(i)}, \widehat{y}_{\text{out}}^{(i)}$ for $i = 1, 2$, where

$$\widehat{y}_{\text{in}}^{(1)}(x)_1 = -\widehat{y}_{\text{in}}^{(1)}(x)_2 = 1 \quad \text{and} \quad \widehat{y}_{\text{in}}^{(2)}(x)_1 = -\widehat{y}_{\text{in}}^{(2)}(x)_2 = -1,$$

then choose the remaining labels $k = 3, \ldots, K$ so that $\widehat{y}_{\text{in}}^{(1)}(x) \preceq y \preceq \widehat{y}_{\text{in}}^{(1)}(x)$ for all $y \in \widehat{C}_{\text{imp}}(x)$ satisfying $y_1 = -y_2 = 1$, and vice-versa for $\widehat{y}_{\text{in}}^{(2)}$ and $\widehat{y}_{\text{out}}^{(2)}$, we may evidently reduce the size of the confidence set $\widehat{C}_{\text{io}}(x)$ by half while maintaining validity.

Extending this idea, let $I \subset [K]$ denote a set of indices. We consider inner and outer sets $\widehat{y}_{\text{in}}$ and $\widehat{y}_{\text{out}}$ that index all configurations of the labels $y_I = (y_i)_{i \in I} \in \{\pm 1\}^m$, so that analogizing the condition (11), we define the 2^m inner and outer sets

$$\begin{aligned}
\widehat{y}_{\text{in}}(x, y_I)_k &= \min\{y'_k \mid y' \in \widehat{C}_{\text{imp}}(x), y'_I = y_I\} \\
&= \begin{cases} -1 & \text{if } \max_{y': y'_k = -1, y'_I = y_I} s(x, y') \geq \widehat{q}_\alpha(x) - \widehat{Q}_{1-\alpha} \\ 1 & \text{if } \max_{y': y'_k = 1, y'_I = y_I} s(x, y') \geq \widehat{q}_\alpha(x) - \widehat{Q}_{1-\alpha} \text{ and preceding fails} \\ +\infty & \text{otherwise,} \end{cases}
\end{aligned} \tag{13a}$$

and similarly

$$\widehat{y}_{\text{out}}(x, y_I)_k = \max\{y'_k \mid y' \in \widehat{C}_{\text{imp}}(x), y'_I = y_I\}. \tag{13b}$$

For any $I \subset [K]$ with $|I| = m$, we can then define the index-based inner/outer confidence set

$$\widehat{C}_{\text{io}}(x, I) := \cup_{y_I \in \{\pm 1\}^m} \left\{y \in \mathcal{Y} \mid \widehat{y}_{\text{in}}(x, y_I) \preceq y \preceq \widehat{y}_{\text{out}}(x, y_I)\right\}, \tag{14}$$

which analogizes the function (10). When m is small, this union of rectangles is efficiently representable, and gives a tighter approximation to $\widehat{C}_{\text{imp}}$ than does the simpler representation (10); indeed, if for some pair $(i, j) \in I$ we have $y_i = -y_j$ for all $y \in \widehat{C}_{\text{imp}}(x)$, but for which there are vectors $y \in \widehat{C}_{\text{imp}}(x)$ realizing $y_i = 1$ and $y_i = -1$, then $|\widehat{C}_{\text{io}}(x, I)| \leq |\widehat{C}_{\text{io}}(x)|/2$. Moreover, no matter the choice I of the index set, we have the containment $\widehat{C}_{\text{io}}(x, I) \supset \widehat{C}_{\text{imp}}(x)$, so that Corollary 9 holds and $\widehat{C}_{\text{io}}$ provides valid marginal coverage. The sets (13) are efficiently computable for the scoring functions s we consider (cf. Sec. 3.2).

The choice of the indices I over which to split the rectangles requires some care. A reasonable heuristic is to obtain the inner/outer vectors $\widehat{y}_{\text{in}}(x)$ and $\widehat{y}_{\text{out}}(x)$ in Alg. 2, and if they provide too large a confidence set, select a pair of variables $I = (i, j)$ for which $\widehat{y}_{\text{in}}(x)_{i,j} = (-1, -1)$ while $\widehat{y}_{\text{out}}(x)_{i,j} = (1, 1)$. To choose the pair, the heuristic is to find those pairings most likely to yield empty confidence sets in the collections (13), which is why we suggest (and use in our experiments) the most negatively correlated pair of labels in the training data that satisfy this joint inclusion;

3.2 Efficient construction of inner and outer confidence sets

With the validity of any inner/outer construction verifying the conditions (11) established, we turn to two approaches to efficiently satisfy these. The first focuses on the scenario where a prediction method provides individual scoring functions $s_k : \mathcal{X} \to \mathbb{R}$ for each task $k \in [K]$ (as frequent for complex classifiers, such as random forests or deep networks, see e.g. Herrera et al., 2016; Read et al., 2011), while the second considers the case when we have a scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that is tree-structured in a sense we make precise; in the next section, we will show how to construct such tree-structured scores using any arbitrary multilabel prediction method.

3.2.1 A direct inner/outer method using individual task scores

We assume here that we observe individual scoring functions $s_k : \mathcal{X} \to \mathbb{R}$ for each task. We can construct inner and outer sets using only these scores while neglecting label correlations by learning threshold functions $t_{\text{in}} \geq t_{\text{out}} : \mathcal{X} \to \mathbb{R}$, where we would like to have the (true) labels y_k satisfy

$$\text{sign}(s_k(x) - t_{\text{in}}(x)) \leq y_k \leq \text{sign}(s_k(x) - t_{\text{out}}(x)). \tag{15}$$

In Algorithm 3, we accomplish this via quantile threshold functions on the maximal and minimal values of the scores s_k for positive and negative labels, respectively. We produce a score function $s : \mathcal{X} \times \mathcal{Y}$ that takes higher values for configurations $y \in \mathcal{Y}$ such that $s_k(x)$ is small when $y_k = -1$ and large when $y_k = 1$, while ensuring that the threshold functions satisfy Eq. (15) with probability $1 - \alpha$.

Algorithm 3: Split Conformalized Direct Inner/Outer method for Classification (CDioC).

Input: Sample $\{(X_i, Y_i)\}_{i=1}^n$, disjoint index sets $\mathcal{I}_2, \mathcal{I}_3 \subset [n]$, quantile functions $\mathcal{Q}$, desired confidence level α, and K score functions $s_k : \mathcal{X} \to \mathbb{R}$.

1. Fit threshold functions (noting the sign conventions)

$$\begin{aligned} \hat{t}_{\text{in}} &= -\operatorname*{argmin}_{t \in \mathcal{Q}} \left\{ \sum_{i \in \mathcal{I}_2} \rho_{\alpha/2}\Big(- \max_k \{s_k(X_i) \mid Y_{i,k} = -1\} - t(X_i) \Big) \right\} \\ \hat{t}_{\text{out}} &= \operatorname*{argmin}_{t \in \mathcal{Q}} \left\{ \sum_{i \in \mathcal{I}_2} \rho_{\alpha/2}\Big(\min_k \{s_k(X_i) \mid Y_{i,k} = 1\} - t(X_i) \Big) \right\} \end{aligned} \tag{16}$$

2. Define score $s(x, y) = \min\{\min_{k:y_k=1} s_k(x) - \hat{t}_{\text{out}}(x), \hat{t}_{\text{in}}(x) - \max_{k:y_k=-1} s_k(x)\}$ and compute conformity scores $S_i := -s(X_i, Y_i)$.

3. Let $\widehat{Q}_{1-\alpha}$ be the $(1 + 1/|\mathcal{I}_3|)(1-\alpha)$-empirical quantile of $S = \{S_i\}_{i \in \mathcal{I}_3}$ and

$$t_{\text{in}}(x) := \hat{t}_{\text{in}}(x) + \widehat{Q}_{1-\alpha} \quad \text{and} \quad t_{\text{out}}(x) := \hat{t}_{\text{out}}(x) - \widehat{Q}_{1-\alpha}.$$

4. Define $\widehat{y}_{\text{in}}(x)_k = \text{sign}(s_k(x) - t_{\text{in}}(x))$ and $\widehat{y}_{\text{out}}(x)_k = \text{sign}(s_k(x) - t_{\text{out}}(x))$ and return prediction set $\widehat{C}_{\text{io}}$ as in Eq. (10).

The method is a slightly modified instantiation of the CQioC method in Alg. 2 that allows easier computation. We can also see that it guarantees validity.

Corollary 11. *Assume that* $\{(X_i, Y_i)\}_{i=1}^{n+1} \overset{\text{iid}}{\sim} P$*, where* P *is an arbitrary distribution. Then the confidence set* $\widehat{C}_{\text{io}}$ *in Algorithm 3 satisfies*

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

Proof We show that the definitions of $\widehat{y}_{\text{in}}$ and $\widehat{y}_{\text{out}}$ in Alg. 3 are special cases of the condition (11), which then allows us to apply Corollary 9. We focus on the inner set, as

the outer is similar. Suppose in Alg. 3 that $\widehat{y}_{\text{in}}(x)_k = 1$. Then $s_k(x) - \hat{t}_{\text{in}}(x) - \widehat{Q}_{1-\alpha} \geq 0$, which implies that $\hat{t}_{\text{in}}(x) - s_k(x) \leq -\widehat{Q}_{1-\alpha}$, and for any $y \in \mathcal{Y}$ satisfying $y_k = -1$, we have

$$\hat{t}_{\text{in}}(x) - \max_{l:y_l=-1} s_l(x) \leq -\widehat{Q}_{1-\alpha}.$$

For the scores $s(x, y)$ in line 2 of Alg. 3, we then immediately obtain $s(x, y) \leq -\widehat{Q}_{1-\alpha}$ for any $y \in \mathcal{Y}$ with $y_k = -1$. This is the first condition (11), so that (performing, *mutatis-mutandis*, the same argument with $\widehat{y}_{\text{out}}(x)_k$) Corollary 9 implies the validity of $\widehat{C}_{\text{io}}$ in Algorithm 3. □

3.2.2 A prediction method for tree-structured scores

We now turn to efficient computation of the inner and outer vectors (11) when the scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is tree-structured. By this we mean that there is a tree $\mathcal{T} = ([K], E)$, with nodes $[K]$ and edges $E \subset [K]^2$, and an associated set of pairwise and singleton factors $\psi_e : \{-1, 1\}^2 \times \mathcal{X} \to \mathbb{R}$ for $e \in E$ and $\varphi_k : \{-1, 1\} \times \mathcal{X} \to \mathbb{R}$ for $k \in [K]$ such that

$$s(x, y) = \sum_{k=1}^{K} \varphi_k(y_k, x) + \sum_{e=(k,l)\in E} \psi_e(y_k, y_l, x). \tag{17}$$

Such a score allows us to consider interactions between tasks k, l while maintaining computational efficiency, and we will show how to construct such a score function both from arbitrary multilabel prediction methods and from those with individual scores as above. When we have a tree-structured score (17), we can use efficient message-passing algorithms (Dawid, 1992; Koller and Friedman, 2009) to compute the collections (maximum marginals) of scores

$$\mathcal{S}_- := \left\{ \max_{y\in\mathcal{Y}:y_k=-1} s(x, y) \right\}_{k=1}^{K} \text{ and } \mathcal{S}_+ := \left\{ \max_{y\in\mathcal{Y}:y_k=1} s(x, y) \right\}_{k=1}^{K} \tag{18}$$

in time $O(K)$, from which it is immediate to construct $\widehat{y}_{\text{in}}$ and $\widehat{y}_{\text{out}}$ as in the conditions (11). We outline the approach in Appendix A.3, as it is not the central theme of this paper, though this efficiency highlights the importance of the tree-structured scores for practicality.

3.3 Building tree-structured scores

With the descriptions of the generic multilabel conformalization method in Alg. 2 and that we can efficiently compute predictions using tree-structured scoring functions (17), we now turn to constructing such scoring functions from predictors, which trade between label dependency structure, computational efficiency, and accuracy of the predictive function. We begin with a general case of an arbitrary predictor function, then describe a heuristic graphical model construction when individual label scores are available (as we assume in Alg. 3).

3.3.1 From arbitrary predictions to scores

We begin with the most general case that we have access only to a predictive function $\widehat{y} : \mathcal{X} \to \mathbb{R}^K$. This prediction function is typically the output of some learning algorithm,

and in the generality here, may either output real-valued scores $\widehat{y}_k(x) \in \mathbb{R}$ or simply output $\widehat{y}_k(x) \in \{-1, 1\}$, indicating element k's presence.

We compute a regularized scoring function based on a tree-structured graphical model (cf. Koller and Friedman, 2009) as follows. Given a tree $\mathcal{T} = ([K], E)$ on the labels $[K]$ and parameters $\alpha \in \mathbb{R}^K$, $\beta \in \mathbb{R}^E$, we define

$$s_{\mathcal{T},\alpha,\beta}(x, y) := \sum_{k=1}^{K} \alpha_k y_k \widehat{y}_k(x) + \sum_{e=(k,l)\in E} \beta_e y_k y_l \tag{19}$$

for all $(x, y) \in \mathcal{X} \times \{-1, 1\}^K$, where we recall that we allow $\widehat{y}_k(x) \in \mathbb{R}$. We will find the tree $\mathcal{T}$ assigning the highest (regularized) scores to the true data $(X_i, Y_i)_{i=1}^n$ using efficient dynamic programs reminiscent of the Chow-Liu algorithm (Chow and Liu, 1968). To that end, we use Algorithm 4.

Algorithm 4: Method to find optimal tree from arbitrary predictor $\widehat{y}$.

Input: Sample $\{(X_i, Y_i)\}_{i\in\mathcal{I}_1}$, regularizers $r_1, r_2 : \mathbb{R} \to \mathbb{R}$, predictor $\widehat{y} : \mathcal{X} \to \mathbb{R}^K$. Set

$$(\widehat{\mathcal{T}}, \widehat{\alpha}, \widehat{\beta}) := \operatorname*{argmax}_{\mathcal{T}=([K],E),\alpha,\beta} \left\{ \sum_{i=1}^{n} s_{\mathcal{T},\alpha,\beta}(X_i, Y_i) - \sum_{k=1}^{K} r_1(\alpha_k) - \sum_{e\in E} r_2(\beta_e) \right\}. \tag{20}$$

and return score function $s_{\widehat{\mathcal{T}},\widehat{\alpha},\widehat{\beta}}$ of form (19).

Because the regularizers r_1, r_2—the squared ℓ_2-norm in our experiments—decompose along the edges and nodes of the tree, we can implement Alg. 4 using a maximum spanning tree algorithm. Indeed, recall (Boyd and Vandenberghe, 2004) the familiar convex conjugate $r^*(t) := \sup_\alpha\{\alpha t - r(\alpha)\}$. Then immediately

$$\begin{aligned}\sup_{\alpha,\beta} \left\{ \sum_{i=1}^{n} s_{\mathcal{T},\alpha,\beta}(X_i, Y_i) - \sum_{k=1}^{K} r_1(\alpha_1) - \sum_{e\in E} r_2(\beta_e) \right\} \\ = \sum_{k=1}^{K} r_1^* \left(\sum_{i=1}^{n} Y_{i,k} \widehat{y}_k(X_i) \right) + \sum_{e=(k,l)\in E} r_2^* \left(\sum_{i=1}^{n} Y_{i,k} Y_{i,l} \right),\end{aligned}$$

which decomposes along the edges of the putative tree. As a consequence, we may solve problem (20) by finding the maximum weight spanning tree in a graph with edge weights $r_2^*(\sum_{i=1}^n Y_{i,k} Y_{i,l})$ for each edge (k, l), then choosing α, β to maximize the objective (20), which is a collection of 1-dimensional convex optimization problems.

3.3.2 From single-task scores to a tree-based probabilistic model

While Algorithm 4 will work regardless of the predictor it is given—which may simply output a vector $\widehat{y} \in \{-1, 1\}^K$, as in Alg. 3 it is frequently the case that multilabel methods output scores $s_k : \mathcal{X} \to \mathbb{R}$ for each task. To that end, a natural strategy is to model the distribution of $Y \mid X$ directly via a tree-structured graphical model (Lafferty et al.,

2001). Similar to the score in Eq. (17), we define interaction factors $\psi : \{-1, 1\}^2 \to \mathbb{R}^4$ by $\psi(-1, -1) = e_1$, $\psi(1, -1) = e_2$, $\psi(-1, 1) = e_3$ and $\psi(1, 1) = e_4$, the standard basis vectors, and marginal factors $\varphi_k : \{-1, 1\} \times \mathcal{X} \to \mathbb{R}^2$ with

$$\varphi_k(y_k, x) := \frac{1}{2} \begin{bmatrix} (y_k - 1) \cdot s_k(x) \\ (y_k + 1) \cdot s_k(x) \end{bmatrix},$$

incorporating information $s_k(x)$ provides on y_k. For a tree $\mathcal{T} = ([K], E)$, the label model is

$$p_{\mathcal{T},\alpha,\beta}(y \mid x) \propto \exp\Big(\sum_{e=(k,l)\in E} \beta_e^T \psi(y_k, y_l) + \sum_{k=1}^K \alpha_k^T \varphi_k(y_k, x) \Big), \tag{21}$$

where (α, β) is a set of parameters such that, for each edge $e \in E$, $\beta_e \in \mathbb{R}^4$ and $\mathbf{1}^T \beta_e = 0$ (for identifiability), while $\alpha_k \in \mathbb{R}^2$ for each label $k \in [K]$. Because we view this as a "bolt-on" approach, applicable to *any* method providing scores s_k, we include only pairwise label interaction factors independent of x, allowing singleton factors to depend on the observed feature vector x through the scores s_k.

The log-likelihood $\log p_{\mathcal{T},\alpha,\beta}$ is convex in (α, β) for any fixed tree $\mathcal{T}$, and the Chow-Liu decomposition (Chow and Liu, 1968) of the likelihood of a tree $\mathcal{T} = ([K], E)$ gives

$$\log p_{\mathcal{T},\alpha,\beta}(y \mid x) = \sum_{k=1}^K \log p_{\mathcal{T},\alpha,\beta}(y_k \mid x) + \sum_{e=(k,l)\in E} \log \frac{p_{\mathcal{T},\alpha,\beta}(y_k, y_l \mid x)}{p_{\mathcal{T},\alpha,\beta}(y_k \mid x) p_{\mathcal{T},\alpha,\beta}(y_l \mid x)}, \tag{22}$$

that is, the sum of the marginal log-likelihoods and pairwise mutual information terms, conditional on $X = x$. Given a sample $(X_i, Y_i)_{i=1}^n$, the goal is to then solve

$$\underset{\mathcal{T},\alpha,\beta}{\text{maximize}} \, L_n(\mathcal{T}, \alpha, \beta) := \sum_{i=1}^n \log p_{\mathcal{T},\alpha,\beta}(Y_i \mid X_i). \tag{23}$$

When there is no conditioning on x, the pairwise mutual information terms $\log \frac{p_{\mathcal{T},\alpha,\beta}(y_k,y_l)}{p(y_k)p(y_l)}$ are independent of the tree $\mathcal{T}$ (Chow and Liu, 1968). We heuristically compute empirical conditional mutual informations between each pair (k, l) of tasks, choosing the tree $\mathcal{T}$ that maximizes these values to approximate problem (23) in Algorithm 5, using the selected tree $\widehat{\mathcal{T}}$ to choose α, β maximizing $L_n(\widehat{\mathcal{T}}, \alpha, \beta)$. (In the algorithm we superscript Y to make task labels versus observations clearer.)

Algorithm 5: Chow-Liu-type approximate Maximum Likelihood Tree and Scoring Function

Input: Sample $\{(X^{(i)}, Y^{(i)})\}_{i \in \mathcal{I}_1}$, and K score functions $s_k : \mathcal{X} \to \mathbb{R}$.
For each pair $e = (k, l) \in [K]^2$

1. Define the single-edge tree $\mathcal{T}_e = (\{k, l\}, \{e\})$
2. Fit model (22) for tree $\mathcal{T}_e$ via $(\widehat{\alpha}, \widehat{\beta}) \coloneqq \operatorname{argmax}_{\alpha,\beta} L_n(\mathcal{T}_e, \alpha, \beta)$
3. Estimate edge empirical mutual information

$$\widehat{I}_e \coloneqq \sum_{i=1}^{n} \log \left(\frac{p_{\mathcal{T}_e, \widehat{\alpha}, \widehat{\beta}}(Y_k^{(i)}, Y_l^{(i)} \mid X^{(i)})}{p_{\mathcal{T}_e, \widehat{\alpha}, \widehat{\beta}}(Y_k^{(i)} \mid X^{(i)}) p_{\mathcal{T}_e, \widehat{\alpha}, \widehat{\beta}}(Y_l^{(i)} \mid X^{(i)})} \right)$$

Set $\widehat{\mathcal{T}} = \textsc{MaxSpanningTree}((\widehat{I}_e)_{e \in [K]^2})$ and $(\widehat{\alpha}, \widehat{\beta}) = \operatorname{argmax}_{\alpha,\beta} L_n(\widehat{\mathcal{T}}, \alpha, \beta)$.
Return scoring function

$$s_{\widehat{\mathcal{T}}, \widehat{\alpha}, \widehat{\beta}}(x, y) \coloneqq \sum_{e=(k,l) \in E} \widehat{\beta}_e^T \psi(y_k, y_l) + \sum_{k=1}^{K} \widehat{\alpha}_k^T \varphi_k(y_k, x)$$

The algorithm takes time roughly $O(nK^2 + K^2 \log(K))$, as each optimization step 2 solves an $O(1)$-dimensional concave maximization problem, which is straightforward via a Newton method (or gradient descent). The approach does not guarantee recovery of the correct tree structure even if the model is well-specified, as we neglect information coming from labels other than k and l in the estimates $\widehat{I}_e$ for edge $e = (k, l)$, although we expect the heuristic to return sufficiently reasonable tree structures. In any case, the actual scoring function s it returns still allows efficient conformalization and valid predictions via Alg. 2, regardless of its accuracy; a more accurate tree will simply allow smaller and more accurate confidence sets (11).

4. Experiments

Our main motivation is to design methods with more robust conditional coverage than the "marginal" split-conformal method (3). Accordingly, the methods we propose in Sections 2 and 3 fit conformalization scores that depend on features x and, in some cases, model dependencies among y variables. Our experiments consequently focus on more robust notions of coverage than the nominal marginal coverage the methods guarantee, and we develop a new evaluation metric for validation and testing of coverage, looking at connected subsets of the space $\mathcal{X}$ and studying coverage over these. Broadly, we expect our proposed methods to maintain coverage of (nearly) $1 - \alpha$ across subsets; the experiments are consistent with this expectation. We include a few additional plots in supplementary appendices for completeness.

4.1 Measures beyond marginal coverage

Except in simulated experiments, we cannot compute conditional coverage of each instance, necessitating approximations that provide more conditional-like measures of coverage—where methods providing weaker marginal coverage may fail to uniformly cover—while still allowing efficient computation. To that end, we consider coverage over slabs

$$S_{v,a,b} \coloneqq \left\{x \in \mathbb{R}^d \mid a \le v^T x \le b\right\},$$

where $v \in \mathbb{R}^d$ and $a < b \in \mathbb{R}$, which satisfy these desiderata. For a direction v and threshold $0 < \delta \le 1$, we consider the worst coverage over all slabs containing δ mass in $\{X_i\}_{i=1}^n$, defining

$$\mathrm{WSC}_n(\widehat{C}, v) \coloneqq \inf_{a<b} \left\{P_n(Y \in \widehat{C}(X) \mid a \le v^T X \le b) \quad \text{s.t.} \quad P_n(a \le v^T X \le b) \ge \delta\right\}, \quad (24)$$

where P_n denotes the empirical distribution on $(X_i, Y_i)_{i=1}^n$, which is efficiently computable in $O(n)$ time (min Chung and Lu, 2003). As long as the mapping $\widehat{C} : \mathcal{X} \rightrightarrows \mathcal{Y}$ is constructed independently of P_n, we can show that these quantities concentrate. Indeed, let us temporarily assume that the confidence set has the form $\widehat{C}(x) = \{y \mid s(x,y) \ge q(x)\}$ for an arbitrary scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and threshold functions q. Let $V \subset \mathbb{R}^d$; we abuse notation to let $\mathrm{VC}(V)$ be the VC-dimension of the set of halfspaces it induces, where we note that $\mathrm{VC}(V) \le \min\{d, \log_2 |V|\}$. Then for some numerical constant C, for all $t > 0$

$$\sup_{v \in V, a \le b : P_n(X \in S_{v,a,b}) \ge \delta} \left\{|P_n(Y \in \widehat{C}(X) \mid X \in S_{v,a,b}) - P(Y \in \widehat{C}(X) \mid X \in S_{v,a,b})|\right\} \quad (25)$$
$$\le C\sqrt{\frac{\mathrm{VC}(V)\log n + t}{\delta n}} \le C\sqrt{\frac{\min\{d, \log|V|\}\log n + t}{\delta n}}$$

with probability at least $1 - e^{-t}$. (See Appendix A.4 for a brief derivation of inequality (25) and a few other related inequalities.)

Each of the confidence sets we develop in this paper satisfy $\widehat{C}(x) \supset \{y \mid s(x,y) \ge q(x)\}$ for some scoring function s and function q. Thus, if $\widehat{C} : \mathcal{X} \rightrightarrows \mathcal{Y}$ effectively provides conditional coverage at level $1 - \alpha$, we should observe that

$$\inf_{v \in V} \mathrm{WSC}_n(\widehat{C}, v) \ge 1 - \alpha - O(1)\sqrt{\frac{\mathrm{VC}(V)\log n}{\delta n}} \ge 1 - \alpha - O(1)\sqrt{\frac{\min\{d, \log|V|\}\log n}{\delta n}}.$$

In each of our coming experiments, we draw $M = 1000$ samples v_j uniformly on $\mathbb{S}^{d-1}$, computing the worst-slab coverage (24) for each v_j. In the multiclass case, we expect our conformal quantile classification (CQC, Alg. 1) method to provide larger worst-slab coverage than the standard marginal method (3), while in the multilabel case, we expect that the combination of tree-based scores (Algorithms 4 or 5) with the conformalized quantile inner/outer classification (CQioC, Alg. 2) should provide larger worst-slab coverage than the conformalized direct inner/outer classification (CDioC, Alg. 3). In both cases, we expect that our more sophisticated methods should provide confidence sets of comparable size to the marginal methods. In multiclass experiments, for comparison, we additionally

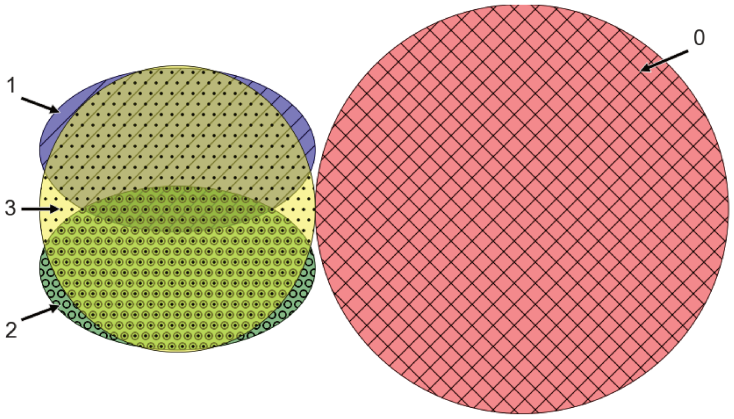

Figure 1. Gaussian mixture with $\mu_0 = (1, 0)$, $\mu_1 = (-\frac{1}{2}, \frac{\sqrt{3}}{2})$, $\mu_2 = (-\frac{1}{2}, -\frac{\sqrt{3}}{2})$, and $\mu_3 = (-\frac{1}{2}, 0)$; and $\Sigma_0 = \text{diag}(0.3, 1)$, $\Sigma_1 = \Sigma_2 = \text{diag}(0.2, 0.4)$, and $\Sigma_3 = \text{diag}(0.2, 0.5)$.

include the Generalized Inverse Quantile method (GIQ, Alg. 1 in Romano et al., 2020), which appeared after the initial version of the current paper appeared on the `arXiv`, and similarly targets improved conditional coverage. Unlike in the multiclass case, we know of no baseline method for multilabel problems, as the "marginal" method (3) is computationally inefficient when the number of labels grows. For this reason, we focus on the methods in this paper, highlighting the potential advantages of each one while comparing them to an oracle (conditionally perfect) confidence set in simulation.

We present five experiments: two with simulated data sets and three with real data sets—CIFAR-10 (Krizhevsky and Hinton, 2009), ImageNet (Deng et al., 2009) and Pascal VOC 2012 (Everingham et al., 2012)—, and both in multiclass and multilabel settings. In each experiment, a single trial corresponds to a realized random split of the data between training, validation and calibration sets $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$, and in each figure, the red dotted line represents the desired level of coverage $1-\alpha$. Unless otherwise specified, we summarize results via boxplots that display the lower and upper quartiles as the hinges of the box, the median as a bold line, and whiskers that extend to the 5% and 95% quantiles of the statistic of interest (typically coverage or average confidence set size).

4.2 Simulation

4.2.1 More uniform coverage on a multiclass example

Our first simulation experiment allows us to compute the conditional coverage of each sample and evaluate our CQC method 1. We study its performance on small sub-populations, a challenge for traditional machine learning models (Duchi and Namkoong, 2018; Hashimoto et al., 2018). In contrast to the traditional split-conformal algorithm (method (3), cf. Vovk et al., 2005), we expect the quantile estimator (5) in the CQC method 1 to better account for data heterogeneity, maintaining higher coverage on subsets of the data, in particular in regions of the space where multiple classes coexist.

To test this, we generate $n = 15,000$ data points $\{X_i, Y_i\}_{i\in[N]}$ i.i.d. from a Gaussian mixture with one majority group and three minority ones,

$$Y \sim \text{Mult}(\pi) \text{ and } X \mid Y = y \sim \mathsf{N}(\mu_y, \Sigma_y). \tag{26}$$

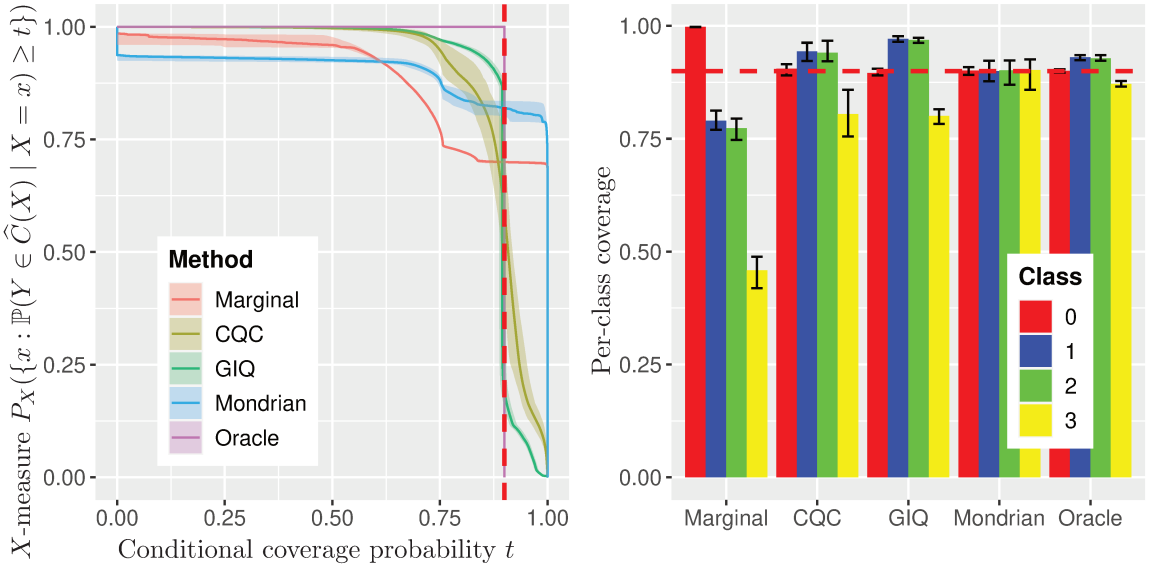

Figure 2. Simulation results on multiclass problem. Left: X-probability of achieving a given level t of conditional coverage versus coverage t, i.e., $t \mapsto P_X(\mathbb{P}(Y \in \widehat{C}_{\text{Method}}(X) \mid X) \geq t)$. The ideal is to observe $t \mapsto 1\{t \leq 1-\alpha\}$. Right: class-wise coverage $\mathbb{P}(Y \in \widehat{C}_{\text{Method}}(X) \mid Y = y)$ on the distribution (26) (as in Fig. 1) for each method. Confidence bands and error bars display the range of the statistic over $M = 20$ trials.

where $\pi = (.7, .1, .1, .1)$ (see Fig. 1). We purposely introduce more confusion for the three minority groups (1, 2 and 3), whereas the majority (0) has a clear linear separation. We choose $\alpha = 10\%$, the same as the size of the smaller sub-populations, then we apply our CQC method and compare it to both the Marginal (3) and the Oracle (which outputs the smallest randomized conditionally valid $1-\alpha$ confidence set) methods. In this example, we provide for completeness a comparison with the "Mondrian" conformal method Balasubramanian et al. (2014), which conformalizes the scores for each label, and hence provides a class-wise coverage guarantee of the form

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

in contrast to feature-conditional coverage goal, harder to achieve.

The results in Figure 2 are consistent with our expectations. The randomized oracle method provides exact $1-\alpha$ conditional coverage, but CQC appears to provide more robust coverage for individual classes than the marginal method. While all methods have comparable confidence set sizes and maintain $1-\alpha$ coverage marginally (see also supplementary Fig. 12), the left plot shows the CQC and GIQ methods provide consistently better conditional coverage than the the Mondrian method, whose emphasis is on label-wise coverage, and the marginal method. The latter (3) has a coverage close to 1 for 70% of the examples (typically all examples from the majority class) and so undercovers the remaining 30% minority examples, in distinction from the CQC and GIQ methods, whose aims for conditional coverage yield better coverage for minority classes (see right plot of Fig. 2).

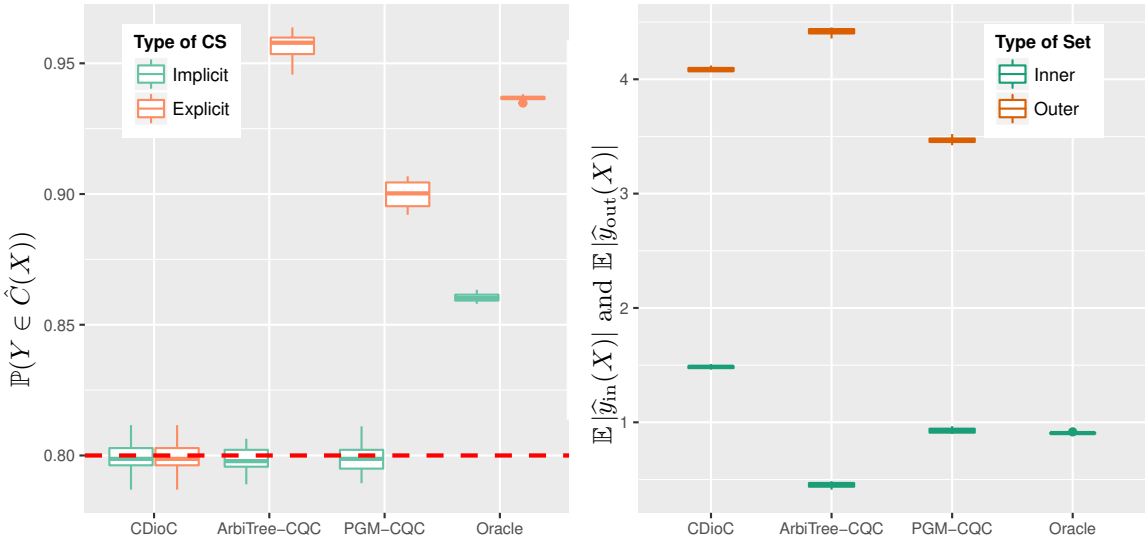

Figure 3. Results for simulated multilabel experiment with label distribution (27). Methods are the true oracle confidence set; the conformalized direct inner/outer method (CDioC, Alg. 3) and tree-based methods with implicit confidence sets $\widehat{C}_{\rm imp}$ or explicit inner/outer sets $\widehat{C}_{\rm io}$, labeled ArbiTree and PGM (see Sec. 4.2.2). Left: Marginal coverage probability at level $1-\alpha = .8$ for different methods and types of confidence sets (CS). Right: confidence set sizes of methods.

4.2.2 Improved coverage with graphical models

Our second simulation addresses the multilabel setting, where we have a predictive model outputting scores $s_k(x) \in \mathbb{R}$ for each task k. As a baseline comparison, we compute oracle confidence sets (the smallest $1-\alpha$-conditionally valid (non-randomized) confidence set in the implicit case and the smaller inner and outer sets containing it in the explicit case).

We run three methods, First, the direct Inner/Outer method (CDioC), Alg. 3. Second, we use the graphical model score from the likelihood model in Alg. 5 to choose a scoring function $s_{\mathcal{T}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which we call the PGM-CQC method; we then either use the CQC method 1 with this scoring function directly, that is, the implicit confidence set (recall (12)) $\widehat{C}_{\rm imp}(x) = \{y \in \mathcal{Y} \mid s_{\mathcal{T}}(x,y) \geq \hat{q}(x) - \widehat{Q}_{1-\alpha}\}$ or the explicit $\widehat{C}_{\rm io}$ set of Eqs. (10)–(11). Finally, we do the same except that we use the arbitrary predictor method (Alg. 4) to construct the score $s_{\mathcal{T}}$, where we use the $\{\pm 1\}^K$ assignment $\widehat{y}$ instead of scores as input predictors, which we term ArbiTree-CQC.

We consider a misspecified logistic regression model, where hidden confounders induce correlation between labels. Because of the dependency structure, we expect the tree-based methods to output smaller and more robust confidence sets than the direct CDioC method 3. Comparing the two score-based methods (CDioC and PGM-CQC) with the scoreless tree method ArbiTree-CQC is perhaps unfair, as the latter uses less information—only signs of predicted labels. Yet we expect the ArbiTree-CQC method to leverage the correlation between labels and mitigate this disadvantage, at least relative to CDioC.

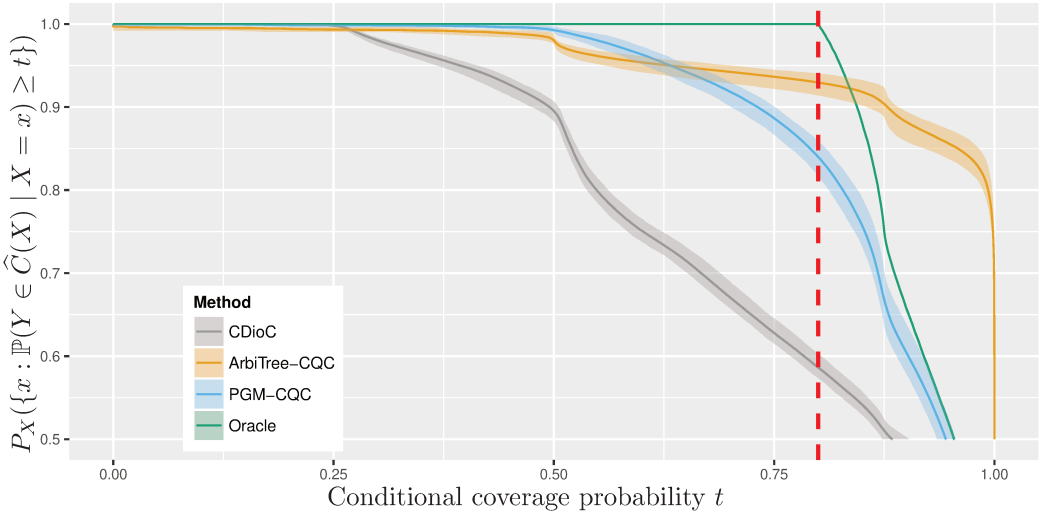

Figure 4. Simulated multilabel experiment with label distribution (27). The plot shows the X-probability of achieving a given level t of conditional coverage versus coverage t, i.e., $t \mapsto P_X(\mathbb{P}(Y \in \widehat{C}_{\text{Method}}(X) \mid X) \geq t)$, using explicit confidence sets $\widehat{C}_{\text{io}}$. The ideal is to observe $t \mapsto 1\{t \leq 1-\alpha\}$. Confidence bands display the range of the statistic over $M = 20$ trials.

Our setting follows, where we consider $K = 5$ tasks and dimension $d = 2$. For each experiment, we choose a tree $\mathcal{T} = ([K], E)$ uniformly at random, where each edge $e \in E$ has a correlation strength value $\tau_e \sim \mathsf{Uni}[-5, 5]$, and for every task k we sample a vector $\theta_k \sim \mathsf{Uni}(r\mathbb{S}^{d-1})$ with radius $r = 5$. For each observation, we draw $X \overset{\text{iid}}{\sim} \mathsf{N}(0, I_d)$ and draw independent uniform hidden variables $H_e \in \{-1, 1\}$ for each edge e of the tree. Letting E_k be the set of edges adjacent to k, we draw $Y_k \in \{-1, 1\}$ from the logistic model

$$\mathbb{P}\left(Y_k = y_k \mid X = x, H = h\right) \propto \exp\left\{-y_k\left(1 + x^T\theta_k + \sum_{e \in E_k} \tau_e h_e\right)\right\}. \tag{27}$$

We simulate $n_{\text{total}} = 40{,}000$ data points, using $n_{\text{tr}} = 10{,}000$ for training, $n_{\text{v}} = 5{,}000$ for validation, $n_{\text{c}} = 5{,}000$ for calibration, and $n_{\text{te}} = 20{,}000$ for testing.

The methods CDioC and PGM-CQC require per-task scores s_k, so for each method we fit K separate logistic regressions of Y_k against X (leaving the hidden variables H unobserved) on the training data. We then use the fit parameters $\widehat{\theta}_k \in \mathbb{R}^d$ to define scores $s_k(x) = \widehat{\theta}_k^T x$ (for the methods CDioC and PGM-CQC) and the "arbitrary" predictor $\widehat{y}_k(x) = \text{sign}(s_k(x))$ (for method ArbiTree-CQC). We use a one-layer fully-connected neural network with 16 hidden neurons as the class $\mathcal{Q}$ in the quantile regression step (5) of the methods; no matter our choice of $\mathcal{Q}$, the final conformalization step guarantees (marginal) validity.

Figure 3, 5, and 4 show our results. In Fig. 3, we give the marginal coverage of each method (left plot), including both the implicit $\widehat{C}_{\text{imp}}$ and explicity $\widehat{C}_{\text{io}}$ confidence sets, and the average confidence set sizes for the explicit confidence set $\widehat{C}_{\text{io}}$ in the right plot. The explicit sets $\widehat{C}_{\text{io}}$ in the ArbiTree-CQC and PGM-CQC methods both overcover, though not substantially more than the oracle method; the sizes of $\widehat{C}_{\text{io}}$ for the PGM-CQC and oracle methods are similar (see supplementary Fig. 13). On the other hand, the PGM-

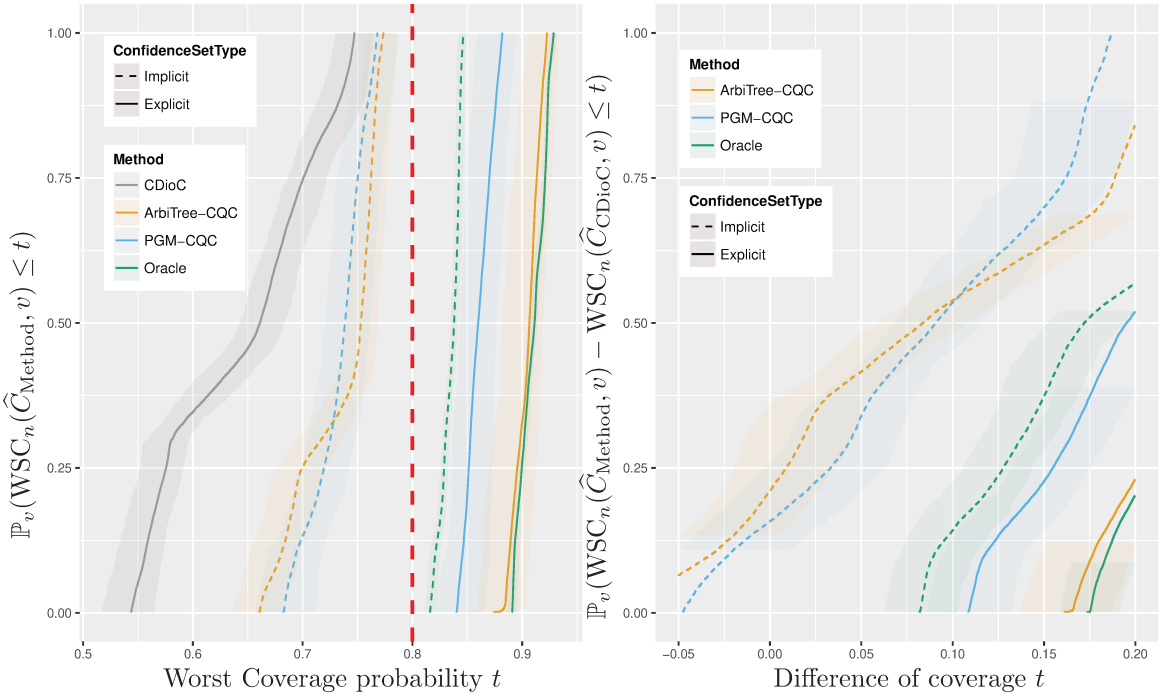

Figure 5. Cumulative distribution of worst-slab coverage (24) (with $\delta = 20\%$) on mis-specifed logistic regression model (27) over $M = 1000$ i.i.d. choices of direction $v \in \mathbb{S}^{d-1}$. We expect to have $\mathbb{P}_v(\mathrm{WSC}_n(\widehat{C}, v) \leq t) = 1\{t \geq 1 - \alpha\}$ if $\widehat{C}$ provides exact $1 - \alpha$-conditional coverage. Left: CDF of $\mathrm{WSC}_n(\widehat{C}, v)$. Right: CDF of difference $\mathrm{WSC}_n(\widehat{C}_{\mathrm{Method}_1}, v) - \mathrm{WSC}_n(\widehat{C}_{\mathrm{Method}_2}, v)$, Method_2 is always the CDioC method (Alg. 3), and the other four are ArbiTree-CQC and PGM-CQC with both implicit $\widehat{C}_{\mathrm{imp}}$ and explicit $\widehat{C}_{\mathrm{io}}$ confidence sets.

CQC explicit confidence sets (which expand the implicit set $\widehat{C}_{\mathrm{imp}}$ as in (11)) cover more than the direct Inner/Outer method CDioC. The confidence sets of the scoreless method ArbiTree-CQC are wider, which is consistent with the limitations of a method using only the predictions $\widehat{y}_k = \mathrm{sign}(s_k)$.

Figures 4 and 5 consider each method's approximation to conditional coverage; the former with exact calculations and the latter via worst-slab coverage measures (24). Both plots dovetail with our expectations that the PGM-CQC and ArbiTree-CQC methods are more robust and feature-adaptive. Indeed, Figure 4 shows that the PGM-CQC method provides at least coverage $1 - \alpha$ for 75% of the examples, against only 60% for the CDioC method, and has an overall consistently higher coverage. In Figure 5, each of the $M = 10^3$ experiments corresponds to a draw of $v \overset{\mathrm{iid}}{\sim} \mathsf{Uni}(\mathbb{S}^{d-1})$, then evaluating the worst-slab coverage (24) with $\delta = .2$. In the left plot, we show its empirical cumulative distribution across draws of v for each method, which shows a substantial difference in coverage between the CDioC method and the others. We also perform direct comparisons in the right plot: we draw the same directions v for each method, and then (for the given v) evaluate the difference $\mathrm{WSC}_n(\widehat{C}, v) - \mathrm{WSC}_n(\widehat{C}', v)$, where $\widehat{C}$ and $\widehat{C}'$ are the confidence sets each method

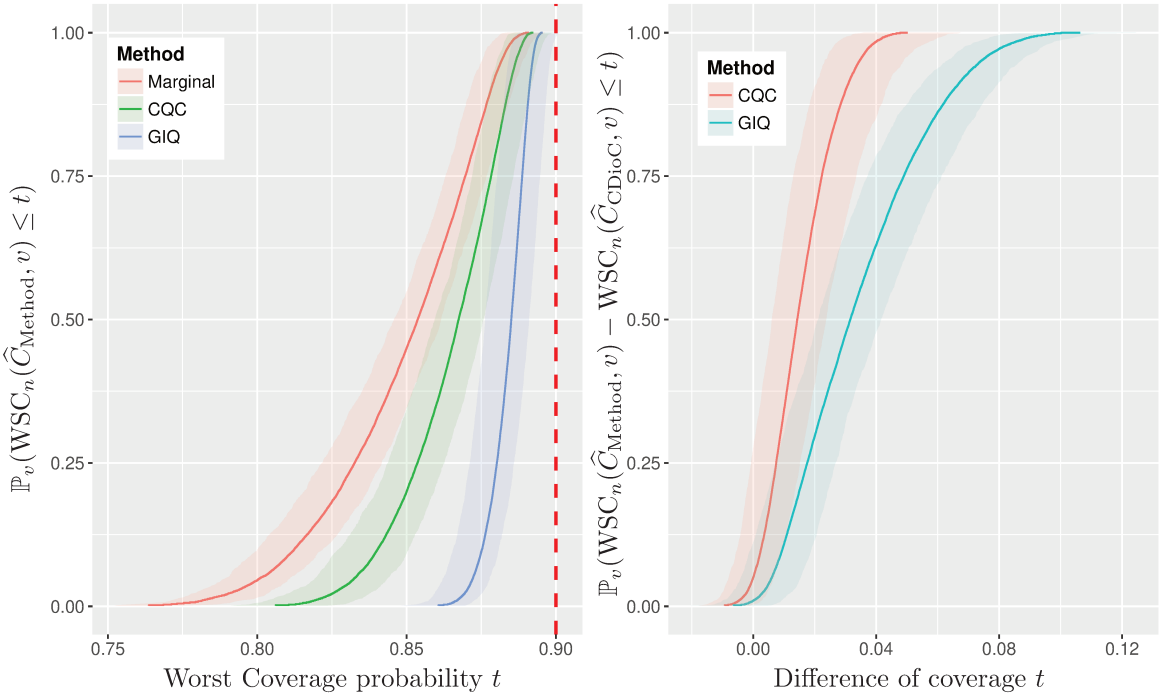

Figure 6. Worst-slab coverage for ImageNet-10 with $\delta = .2$ over $M = 1000$ draws $v \overset{\text{iid}}{\sim} \mathsf{Uni}(\mathbb{S}^{d-1})$. The dotted line is the desired (marginal) coverage. Left: CDF of worst-slab coverage. Right: CDF of coverage difference $\mathrm{WSC}_n(\widehat{C}_{\mathrm{CQC}}, v) - \mathrm{WSC}_n(\widehat{C}_{\mathrm{Marginal}}, v)$.

produces, respectively. Thus, we see that both tree-based methods always provide better worst-slab coverage, whether we use the implicit confidence sets $\widehat{C}_{\mathrm{imp}}$ or the larger direct inner/outer (explicit) confidence sets $\widehat{C}_{\mathrm{io}}$, though in the latter case, some of the difference likely comes from the differences in marginal coverage. The worst-slab coverage is consistent with the true conditional coverage in that the relative ordering of method performance is consistent, suggesting its usefulness as a proxy.

4.3 More robust coverage on CIFAR 10 and ImageNet data sets

In our first real experiments, we study two multiclass image classification problems with the benchmark CIFAR-10 (Krizhevsky and Hinton, 2009) and ImageNet (Deng et al., 2009) data sets. We use similar approaches to construct our feature vectors and scoring functions. With the CIFAR-10 data set, which consists of $n = 60{,}000$ 32×32 images across 10 classes, we use $n_{\mathrm{tr}} = 50{,}000$ of them to train the full model, a validation set of size $n_{\mathrm{v}} = 6{,}000$ for fitting quantile functions and hyperparameter tuning, $n_{\mathrm{c}} = 3{,}000$ for calibration and the last $n_{\mathrm{te}} = 1{,}000$ for testing. We train a standard ResNet50 (He et al., 2016) architecture for 200 epochs, obtaining test set accuracy $92.5 \pm 0.5\%$ (on the 10,000 examples it was not trained on), and use the $d = 256$-dimensional output of the final layer of the ResNet as the inputs X to the quantile estimation. For the ImageNet classification problem, we load a pre-trained Inception-ResNetv2 (Szegedy et al., 2017) architecture, achieving top-1 test accuracy $80.3 \pm 0.5\%$, using the $d = 1536$-dimensional output of the final layer as features X.

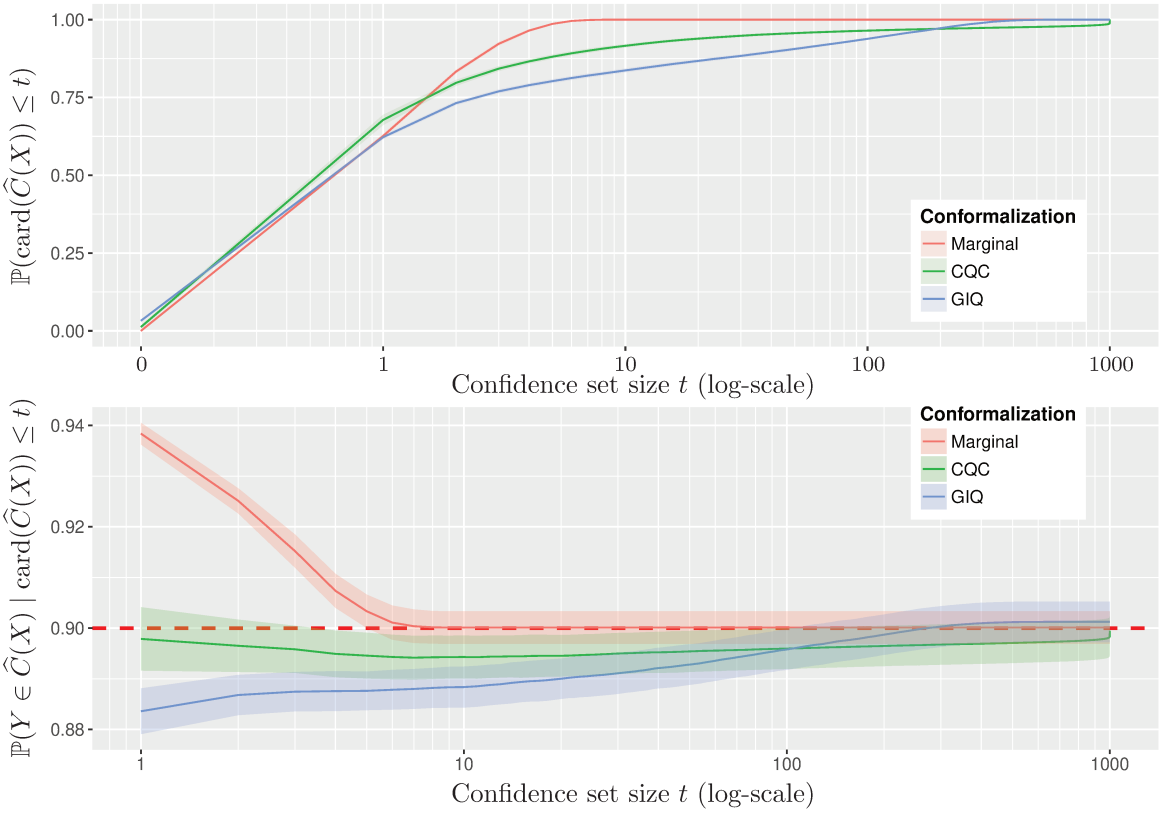

Figure 7. ImageNet results over 20 trials. Methods are the marginal method (Alg. 3), the CQC method (Alg. 1), and the GIQ method (Alg. 1, Romano et al., 2020). Top: cumulative distribution of confidence set size. Bottom: probability of coverage conditioned on confidence set size.

Splitting the original ImageNet validation set containing 50,000 instances into 3 different sets, we fit our quantile function on $n_{\text{v}} = 30{,}000$ examples, calibrate on $n_{\text{c}} = 10{,}000$ and test on the last $n_{\text{te}} = 10{,}000$ images.

We apply the CQC method 1 with $\alpha = 5\%$ for CIFAR-10 and $\alpha = 10\%$ for ImageNet and compare it to the benchmark marginal method (3) and GIQ (Romano et al., 2020). We expect the marginal method to typically output small confidence sets, as the neural network's accuracy is close to the confidence level $1 - \alpha$; this allows (typically) predicting a single label while maintaining marginal coverage. Supplementary figures 15 and 16 show this. The worst-slab coverage (24) tells a different story. In Figure 6, we compare worst-slab coverage over $M = 1000$ draws of $v \overset{\text{iid}}{\sim} \mathsf{Uni}(\mathbb{S}^{d-1})$ on the ImageNet data set (see also Fig. 17 for the identical experiment with CIFAR-10). The CQC and GIQ methods provide significant 3–5% and 5–7% better coverage, respectively, in worst-slab coverage over the marginal method. The CQC and GIQ methods provide stronger gains on ImageNet than on CIFAR-10, which we attribute to the relative easiness of CIFAR-10: the accuracy of the classifier is high, allowing near coverage by Dirac measures.

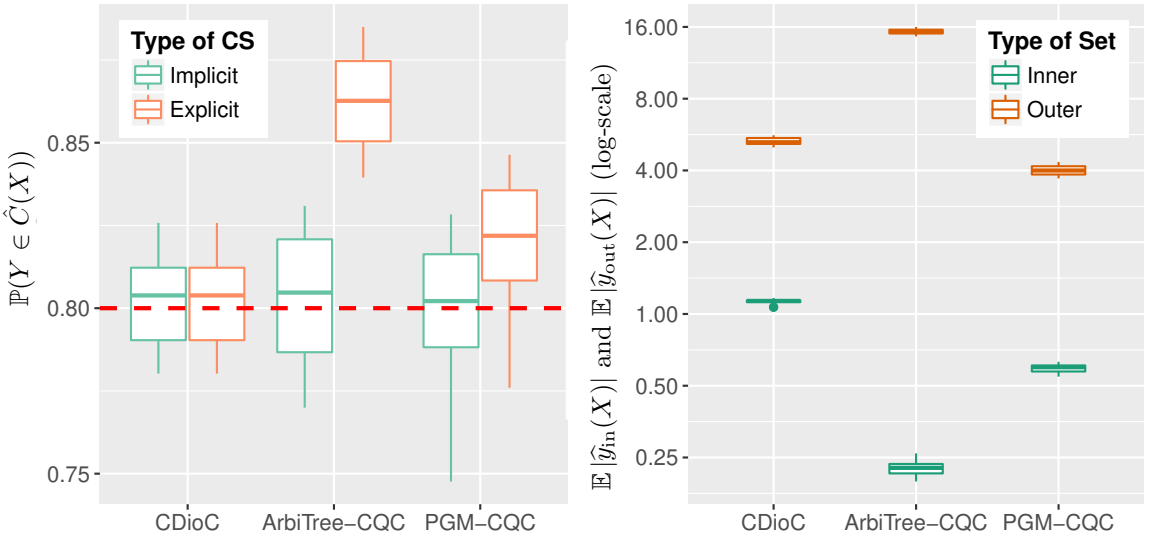

Figure 8. Pascal-VOC data set (Everingham et al., 2012) results over 20 trials. Methods are the conformalized direct inner/outer method (CDioC, Alg. 3) and tree-based methods with implicit confidence sets $\widehat{C}_{\text{imp}}$ or explicit inner/outer sets $\widehat{C}_{\text{io}}$, labeled ArbiTree and PGM (see Sec. 4.2.2). Left: Marginal coverage probability at level $1 - \alpha = .8$. Right: confidence set sizes of methods.

Figure 7 compares the confidence set sizes and probability of coverage given confidence set size for the marginal, CQC, and GIQ methods. We summarize briefly. The CQC method gives confidence sets of size at most 2 for 80% of the instances—comparable to the marginal method and more frequently than the GIQ method (which yields 75% examples with $|\widehat{C}(X)| \le 2$). Infrequently, the CQC and GIQ methods yield very large confidence sets, with $|\widehat{C}(X)| \ge 200$ about 5% of the time for the GIQ method and the completely informative $|\widehat{C}(X)| = 1000$ about 2% of the time for the CQC method. While the average confidence set size $\mathbb{E}[|\widehat{C}(X)|]$ is smaller for the marginal method (cf. supplementary Fig. 16), this is evidently a very incomplete story. The bottom plot in Fig. 7 shows the behavior we expect for a marginal method given a reasonably accurate classifier: it overcovers for examples x with $\widehat{C}(x)$ small. GIQ exhibits the opposite behavior, overcovering when $\widehat{C}(x)$ is large and undercovering for small $\widehat{C}(x)$, while GIQ provides nearly $1-\alpha$ coverage roughly independent of confidence set size, as one would expect for a method with valid conditional coverage. (In supplementary Fig. 14, we see similar but less-pronounced behavior on CIFAR-10.)

4.4 A multilabel image recognition data set

Our final set of experiments considers the multilabel image classification problems in the PASCAL VOC 2007 and VOC 2012 data sets (Everingham et al., 2007, 2012), which consist of $n_{2012} = 11540$ and $n_{2007} = 9963$ distinct 224×224 images, where the goal is to predict the presence of entities from $K = 20$ different classes, (e.g. birds, boats, people). We compare the direct inner outer method (CDioC) 3 with the split-conformalized inner/outer method (CQioC) 2, where we use the tree-based score functions that Algorthims 4 and 5

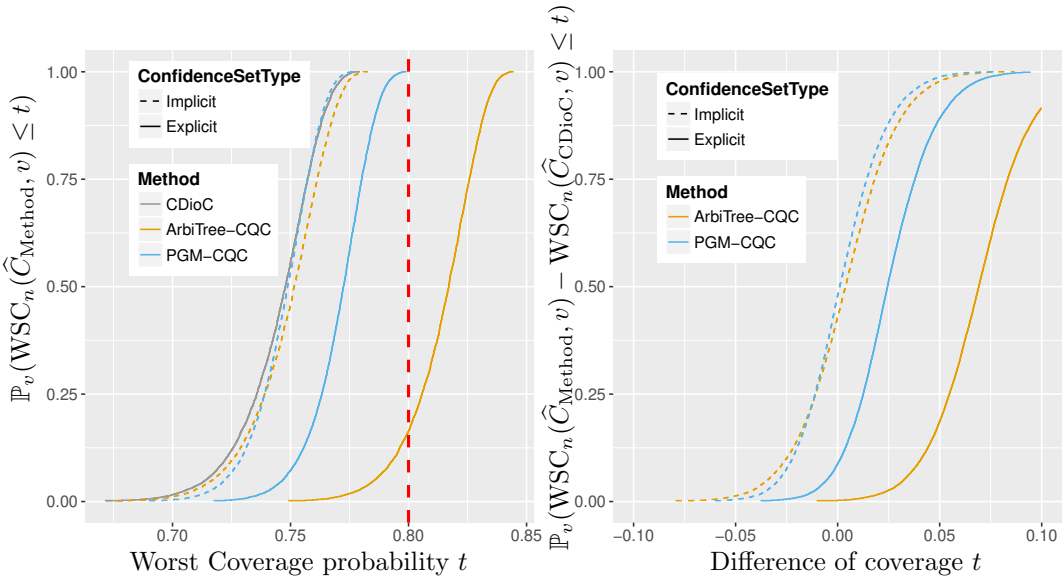

Figure 9. Worst-slab coverage (24) for Pascal-VOC with $\delta = .2$ over $M = 1000$ draws $v \overset{\text{iid}}{\sim} \mathsf{Uni}(\mathbb{S}^{d-1})$. For tree-structured methods ArbiTree-CQC and PGM-CQC, we compute the worst-slab coverage using implicit confidence sets $\widehat{C}_{\text{imp}}$ and explicit inner/outer sets $\widehat{C}_{\text{io}}$. The dotted line is the desired (marginal) coverage. Left: distribution of worst-slab coverage. Right: distribution of the coverage difference $\mathrm{WSC}_n(\widehat{C}_{\mathrm{Method}}, v) - \mathrm{WSC}_n(\widehat{C}_{\mathrm{CDioC}}, v)$ for $\mathrm{Method}_i \in \{\text{ArbiTree-CQC}, \text{PGM-CQC}\}$ with implicit $\widehat{C}_{\text{imp}}$ or explicit inner/outer $\widehat{C}_{\text{io}}$ confidence sets.

output. For the PGM method, which performs best in practice, we additionally compare the performance of standard inner and outer sets (see Alg. 2), to the refinement that we describe in section 3.1.1, where we instead output a confidence set as a union of inner and outer sets. Here, we choose $m = 2$, which corresponds to outputting a union of 4 inner/outer sets in equation 14, and select the indices I according to the heuristics that we describe in that same section.

In this problem, we use the $d = 2048$ dimensional output of a ResNet-101 with pretrained weights on the ImageNet data set (Deng et al., 2009) as our feature vectors X. For each of the K classes, we fit a binary logistic regression model $\widehat{\theta}_k \in \mathbb{R}^d$ of $Y_k \in \{\pm 1\}$ against X, then use $s_k(x) = \widehat{\theta}_k^T x$ as the scores for the PGM-CQC method (Alg. 5). We use $\widehat{y}_k(x) = \mathrm{sign}(s_k(x))$ for the ArbiTree-CQC method 4. The fit predictors have F1-score 0.77 on held-out data, so we do not expect our confidence sets to be uniformly small while maintaining the required level of coverage, in particular for ArbiTree-CQC. We use a validation set of $n_{\text{v}} = 3493$ images to fit the quantile functions (5) as above using a one layer fully-connected neural network with 16 hidden neurons and tree parameters.

The results from Figure 9 are consistent with our hypotheses that the tree-based models should improve robustness of the coverage. Indeed, while all methods ensure marginal coverage at level $\alpha = .8$ (see Fig. 8), Figure 9 shows that worst-case slabs (24) for the tree-based methods have closer to $1-\alpha$ coverage. In particular, for most random slab directions v, the tree-based methods have higher worst-slab coverage than the direct inner/outer method (CDioC, Alg. 3). At the same time, both the CDioC method and PGM-CQC method (using $\widehat{C}_{\text{io}}$) provide similarly-sized confidence sets, as the inner set of the PGM-CQC method is

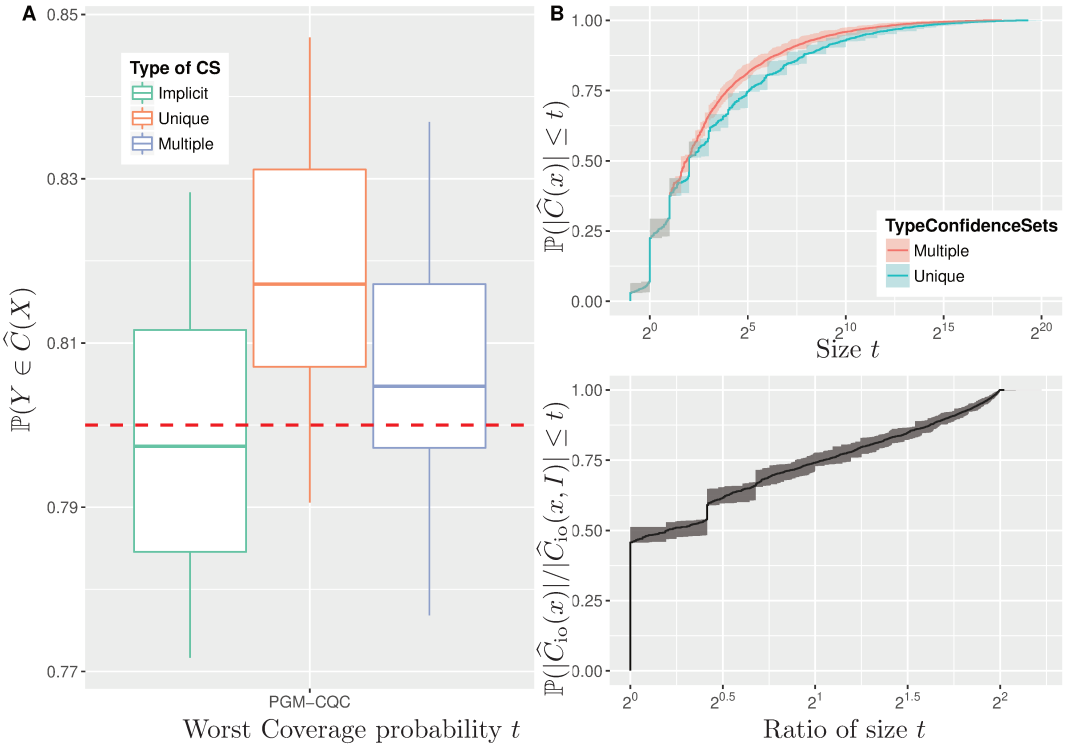

Figure 10. Comparison between standard inner/outer sets (10) ("Unique" in the figure) and union of inner/outer sets (14) ("Multiple"), on Pascal-VOC data set. The dotted line is the desired (marginal) coverage. The label index set I contains two labels for each instance x (see Sec. 3.1.1). A: summary box plot of coverage for each type of confidence set over 20 trials. B: cumulative distribution of the number of total configurations in each confidence set. C: cumulative distribution of the ratio of size of confidence sets (10) and (14) (with $m = 2$).

typically smaller, as is its outer set (cf. Fig. 11). The ArbiTree-CQC method performs poorly on this example: its coverage is at the required level, but the confidence sets are too large and thus essentially uninformative.

Finally, Figure 10 investigates refining the confidence sets as we suggest in Section 3.1.1 by taking pairs of the most negatively correlated labels i, j satisfying $\widehat{y}_{\text{in}}(x)_{(i,j)} = -1$ and $\widehat{y}_{\text{out}}(x)_{(i,j)} = 1$. Moving from a single inner/outer set to the union of 4 inner/outer sets non only mitigates increased coverage, moving closer to the target level $1 - \alpha$ and to the coverage of the implicit confidence sets, but in addition, for more than half of the examples, it decreases (by a factor up to $4 = 2^m$) the number of configurations in the confidence set.

5. Conclusions

As long as we have access to a validation set, independent of (or at least exchangeable with) the sets used to fit a predictive model, split conformal methods guarantee marginal validity, which gives great freedom in modeling. It is thus of interest to fit models more adaptive to the signal inputs x at hand—say, by quantile predictions as we have done, or other methods yet to be discovered—that can then in turn be conformalized. As yet we have limited

understanding of what "near" conditional coverage might be possible: Barber et al. (2019) provide procedures that can guarantee coverage uniformly across subsets of X-space as long as those subsets are not too complex, but it appears computationally challenging to fit the models they provide, and our procedures empirically appear to have strong coverage across slabs of the data as in Eq. (24). Our work, then, is a stepping stone toward more uniform notions of validity, and we believe exploring approaches to this—perhaps by distributionally robust optimization (Delage and Ye, 2010; Duchi and Namkoong, 2018), perhaps by uniform convergence arguments (Barber et al., 2019, Sec. 4)—will be both interesting and essential for trusting applications of statistical machine learning.

Acknowledgments

We would like to acknowledge support for this project from NSF CAREER award CCF-1553086, ONR Young Investigator Program award N00014-19-2288, and NSF award HDR-1934578 (the Stanford Data Science Collaboratory).

Appendix A. Technical proofs and appendices

A.1 Proof of Theorem 6

Our proof adapts arguments similar to those that Sesia and Candès (2019) use in the regression setting, repurposing and modifying them for classification. We have that $\|\widehat{s} - s\|_{L^2(P_X)} \xrightarrow{p} 0$ and $\|\widehat{q}^\sigma_\alpha - q^\sigma_\alpha\|_{L^2(P_X)} \xrightarrow{p} 0$. We additionally have that

$$\widehat{Q} := Q_{1-\alpha}(E^\sigma, \mathcal{I}_3) \xrightarrow{p} 0. \tag{28}$$

This is item (ii) in the proof of Theorem 1 of Sesia and Candès (2019) (see Appendix A of their paper), which proves the convergence (28) of the marginal quantile of the error precisely when $\widehat{q}^\sigma_\alpha$ and $\widehat{s}$ are L^2-consistent in probability (when the randomized quantity $s(x, Y) + \sigma Z$ has a density), as in our Assumption 1.

Recalling that $\widehat{s}, \widehat{q}$ tacitly depend on the sample size n, let $\epsilon > 0$ be otherwise arbitrary, and define the sets

$$B_n := \left\{x \in \mathcal{X} \mid \|\widehat{s}(x, \cdot) - s(x, \cdot)\|_\infty > \epsilon^2 \text{ or } |\widehat{q}^\sigma_\alpha(x) - q^\sigma_\alpha(x)| > \epsilon^2\right\}.$$

Then $B_n \subset \mathcal{X}$ is measurable, and by Markov's inequality,

$$P_X(B_n) \le \frac{\|\widehat{s} - s\|_{L^2(P_X)}}{\epsilon} + \frac{\|\widehat{q}^\sigma_\alpha - q^\sigma_\alpha\|_{L^2(P_X)}}{\epsilon},$$

so

$$\mathbb{P}(P_X(B_n) \ge \epsilon) \le \mathbb{P}(\|\widehat{s} - s\|_{L^2(P_X)} > \epsilon^2) + \mathbb{P}(\|\widehat{q}^\sigma_\alpha - q^\sigma_\alpha\|_{L^2(P_X)} > \epsilon^2) \to 0. \tag{29}$$

Thus, the measure of the sets B_n tends to zero in probability, i.e., $P_X(B_n) \xrightarrow{p} 0$.

Now recall the shorthand (28) that $\widehat{Q} = Q_{1-\alpha}(E^\sigma, \mathcal{I}_3)$. Let us consider the event that $\widehat{C}^\sigma_{1-\alpha}(x, z) \neq C^\sigma_{1-\alpha}(x, z)$. If this is the case, then we must have one of

$$\begin{aligned} A_1(x, k, z) &:= \left\{\widehat{s}(x, k) + \sigma z \ge \widehat{q}^\sigma_\alpha(x) - \widehat{Q} \text{ and } s(x, k) + \sigma z < q^\sigma_\alpha(x)\right\} \text{ or} \\ A_2(x, k, z) &:= \left\{\widehat{s}(x, k) + \sigma z < \widehat{q}^\sigma_\alpha(x) - \widehat{Q} \text{ and } s(x, k) + \sigma z \ge q^\sigma_\alpha(x)\right\}. \end{aligned} \tag{30}$$

We show that the probability of the set A_1 is small; showing that the probability of set A_2 is small is similar. Using the convergence (28), let us assume that $|\widehat{Q}| \le \epsilon$, and suppose that $x \notin B_n$. Then for A_1 to occur, we must have both $s(x, k) + \epsilon + \sigma z \ge q^\sigma_\alpha(x) - 2\epsilon$ and $s(x, k) + \sigma z < q^\sigma_\alpha(x)$, or

$$q^\sigma_\alpha(x) - s(x, k) - 3\epsilon \le \sigma z < q^\sigma_\alpha(x) - s(x, k).$$

As Z has a bounded density, we have $\limsup_{\epsilon \to 0} \sup_{a \in \mathbb{R}} P_Z(a \le \sigma Z \le a + 3\epsilon) = 0$, or (with some notational abuse) $\limsup_{\epsilon \to 0} \sup_{x \notin B_n} P_Z(A_1(x, k, Z)) = 0$.

Now, let $\mathcal{F}_n = \sigma(\{X_i\}_{i=1}^n, \{Y_i\}_{i=1}^n, \{Z_i\}_{i=1}^n)$ be the σ-field of the observed sample. Then by the preceding derivation (*mutatis mutandis* for the set A_2 in definition (30)) for any $\eta > 0$, there is an $\epsilon > 0$ (in the definition of B_n) such that

$$\begin{aligned} &\sup_{x \in \mathcal{X}} \mathbb{P}\left(\widehat{C}^\sigma_{1-\alpha}(x, Z_{n+1}) \neq C^\sigma_{1-\alpha}(x, Z_{n+1}) \mid \mathcal{F}_n\right) 1\{x \notin B_n\} 1\left\{|\widehat{Q}| \le \epsilon\right\} \\ &\quad \le \sup_{x \notin B_n} \sum_{k=1}^K \mathbb{P}\left(A_1(x, k, Z_{n+1}) \text{ or } A_2(x, k, Z_{n+1}) \mid \mathcal{F}_n\right) 1\left\{|\widehat{Q}| \le \epsilon\right\} \le \eta. \end{aligned}$$

In particular, by integrating the preceding inequality, we have

$$\mathbb{P}\left(\widehat{C}^{\sigma}_{1-\alpha}(X_{n+1}, Z_{n+1}) \neq \widehat{C}(X_{n+1}, Z_{n+1}), X_{n+1} \notin B_n, |\widehat{Q}| \leq \epsilon\right) \leq \eta.$$

As $\mathbb{P}(|\widehat{Q}| \leq \epsilon) \to 1$ and $\mathbb{P}(X_{n+1} \notin B_n) \to 1$ by the convergence guarantees (28) and (29), we have the theorem.

A.2 Proof of Proposition 7

We wish to show that $\lim_{\sigma \to 0} \mathbb{P}\left(C^{\sigma}_{1-\alpha}(X, Z) \subset C_{1-\alpha}(X)\right) = 1$, which, by a union bound over $k \in [K]$, is equivalent to showing

$$\mathbb{P}\left(s(X,k) + \sigma Z \geq q^{\sigma}_{\alpha}(X), s(X,k) < q_{\alpha}(X)\right) \underset{\sigma \to 0}{\longrightarrow} 0$$

for all $k \in [K]$. Fix $k \in [K]$, and define the events

$$A \coloneqq \{s(X,k) < q_{\alpha}(X)\}$$

and

$$B^{\sigma} \coloneqq \{s(X,k) + \sigma Z \geq q^{\sigma}_{\alpha}(X)\}.$$

Now, for any $\delta > 0$, consider $A_{\delta} \coloneqq \{s(X,k) \leq q_{\alpha}(X) - \delta\}$. On the event $A_{\delta} \cap B^{\sigma}$, it must hold that

$$\delta \leq q_{\alpha}(X) - q^{\sigma}_{\alpha}(X) + \sigma Z.$$

The following lemma—whose proof we defer to Section A.2.1—shows that the latter can only occur with small probability.

Lemma 12. *With probability 1 over X, the quantile function satisfies*

$$\liminf_{\sigma \to 0} q^{\sigma}_{\alpha}(X) \geq q_{\alpha}(X),$$

and hence $\limsup_{\sigma \to 0} \{q_{\alpha}(X) - q^{\sigma}_{\alpha}(X) + \sigma Z\} \leq 0$ *almost surely.*

Lemma 12 implies that

$$\mathbb{P}\left(q_{\alpha}(X) - q^{\sigma}_{\alpha}(X) + \sigma Z \geq \delta\right) \underset{\sigma \to 0}{\longrightarrow} 0,$$

which, in turn, shows that, for every fixed $\delta > 0$,

$$\mathbb{P}\left(A_{\delta} \cap B^{\sigma}\right) \underset{\sigma \to 0}{\longrightarrow} 0.$$

To conclude the proof, fix $\epsilon > 0$. The event A_{δ} increases to A when $\delta \to 0$, so there exists $\delta > 0$ so that

$$\mathbb{P}\left(A \setminus A_{\delta}\right) \leq \epsilon.$$

Finally,

$$\limsup_{\sigma \to 0} \mathbb{P}\left(A \cap B^{\sigma}\right) \leq \mathbb{P}\left(A \setminus A_{\delta}\right) + \limsup_{\sigma \to 0} \mathbb{P}\left(A_{\delta} \cap B^{\sigma}\right) \leq \epsilon,$$

as $\limsup_{\sigma \to 0} \mathbb{P}\left(A_{\delta} \cap B^{\sigma}\right) = 0$. We conclude the proof by sending $\epsilon \to 0$.

A.2.1 Proof of Lemma 12

Fix $x \in \mathcal{X}$. Let $F_{\sigma,x}$ and F_x be the respective cumulative distribution functions of $s^\sigma(x, Y, Z)$ and $s(x, Y)$ conditionally on $X = x$, and define the (left-continuous) inverse CDFs

$$F^{-1}_{\sigma,x}(u) = \inf\{t \in \mathbb{R} : u \le F_{\sigma,x}(t)\} \quad \text{and} \quad F^{-1}_x(u) = \inf\{t \in \mathbb{R} : u \le F_x(t)\}.$$

We use a standard lemma about the convergence of inverse CDFs, though for lack of a proper reference, we include a proof in Section A.2.2.

Lemma 13. *Let $(F_n)_{n\ge1}$ be a sequence of cumulative distribution functions converging weakly to F, with inverses F_n^{-1} and F^{-1}. Then for each $u \in (0, 1)$,*

$$F^{-1}(u) \le \liminf_{n\to\infty} F_n^{-1}(u) \le \limsup_{n\to\infty} F_n^{-1}(u) \le F^{-1}(u+). \tag{31}$$

As $F_{\sigma,x}$ converges weakly to F_x as $\sigma \to 0$, Lemma 13 implies that

$$F^{-1}_x(\alpha) \le \liminf_{\sigma\to0} F^{-1}_{\sigma,x}(\alpha).$$

But observe that $q^\sigma_\alpha(x) = F^{-1}_{\sigma,x}(\alpha)$ and $q_\alpha(x) = F^{-1}_x(\alpha)$, so that we have the desired result $q_\alpha(x) \le \liminf_{\sigma\to0} q^\sigma_\alpha(x)$.

A.2.2 Proof of Lemma 13

We prove only the first inequality, as the last inequality is similar. Fix $u \in (0, 1)$ and $\epsilon > 0$. F is right-continuous and non-decreasing, so its set of continuity points is dense in $\mathbb{R}$; thus, there exists a continuity point w of F such that $w < F^{-1}(u) \le w + \epsilon$.

Since $w < F^{-1}(u)$, it must hold that $F(w) < u$, by definition of F^{-1}. As w is a continuity point of F, $\lim_{n\to\infty} F_n(w) = F(w) < u$, which means that $F_n(w) < u$ for large enough n, or equivalently, that $w < F_n^{-1}(u)$. We can thus conclude that

$$\liminf F_n^{-1}(u) \ge w \ge F^{-1}(u) - \epsilon.$$

Taking $\epsilon \to 0$ proves the first inequality.

A.3 Efficient computation of maximal marginals for condition (11)

We describe a more or less standard dynamic programming approach to efficiently compute the maximum marginals (18) (i.e. maximal values of a tree-structured score $s : \mathcal{Y} = \{-1, 1\}^K \to \mathbb{R}$), referring to standard references on max-product message passing (Dawid, 1992; Wainwright and Jordan, 2008; Koller and Friedman, 2009) for more. Let $\mathcal{T} = ([K], E)$ be a tree with nodes $[K]$ and undirected edges E, though we also let edges($\mathcal{T}$) and nodes($\mathcal{T}$) denote the edges and nodes of the tree $\mathcal{T}$. Assume

$$s(y) = \sum_{k=1}^K \varphi_k(y_k) + \sum_{e=(k,l)\in E} \psi_e(y_k, y_l).$$

To compute the *maximum marginals* $\max_{y\in\mathcal{Y}}\{s(y) \mid y_k = \hat{y}_k\}$, we perform two message passing steps on the tree: an upward and downward pass. Choose a node r arbitrarily to

be the root of $\mathcal{T}$, and let $\mathcal{T}_{\rm down}$ be the directed tree whose edges are E with r as its root (which is evidently unique); let $\mathcal{T}_{\rm up}$ be the directed tree with all edges reversed from $\mathcal{T}_{\rm down}$.

A maximum marginal message passing algorithm then computes a single downward and a single upward pass through each tree, each in topological order of the tree. The downard messages $m_{l\to k} : \{-1,1\} \to \mathbb{R}$ are defined for $\hat{y} \in \{-1,1\}$ by

$$m_{l\to k}(\hat{y}) = \max_{y_l \in \{-1,1\}} \left\{ \varphi_l(y_l) + \psi_{(l,k)}(y_l, \hat{y}) + \sum_{i:(i\to l)\in {\rm edges}(\mathcal{T}_{\rm down})} m_{i\to l}(y_l) \right\},$$

while the upward pass is defined similarly except that $\mathcal{T}_{\rm up}$ replaces $\mathcal{T}_{\rm down}$. After a single downward and upward pass through the tree, which takes time $O(K)$, the invariant of message passing on the tree (Koller and Friedman, 2009, Ch. 13.3) then guarantees that for each $k \in [K]$,

$$\max_{y\in\mathcal{Y}} \{s(y) \mid y_k = \hat{y}_k\} = \varphi_k(\hat{y}_k) + \sum_{e=(l,k)\in\mathcal{T}} m_{l\to k}(\hat{y}_k), \tag{32}$$

where we note that there exists exactly one message to k (whether from the downward or upward pass) from each node l neighboring k in $\mathcal{T}$.

We can evidently then compute all of these maximal values (the sets $\mathcal{S}_\pm$ in Eq. (18)) simultaneously in time of the order of the number of edges in the tree $\mathcal{T}$, or $O(K)$, as each message $m_{l\to k}$ can appear in at most one of the maxima (32), and there are $2K$ messages.

A.4 Concentration of coverage quantities

We sketch a derivation of inequality (25); see also (Barber et al., 2019, Theorem 5) for related arguments.

We begin with a technical lemma that is the basis for our result. In the lemma, we abuse notation briefly, and let $\mathcal{F} \subset \mathcal{Z} \to \{0,1\}$ be a collection of functions with VC-dimension d. We define $Pf = \int f(z)dP(z)$ and $P_n f = \frac{1}{n}\sum_{i=1}^n f(Z_i)$, as is standard.

Lemma 14 (Relative concentration bounds, e.g. Boucheron et al. (2005), Theorem 5.1)**.** *Let* $\mathrm{VC}(\mathcal{F}) \le d$. *There is a numerical constant* C *such that for any* $t > 0$, *with probability at least* $1 - e^{-t}$,

$$\sup_{f\in\mathcal{F}} \left\{ |Pf - P_n f| - C\sqrt{\min\{Pf, P_n f\}\frac{d\log n + t}{n}} \right\} \le C\frac{d\log n + t}{n}.$$

Proof By Boucheron et al. (2005, Thm. 5.1) for $t > 0$, with probability at least $1 - e^{-t}$ we have

$$\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}} \le C\sqrt{\frac{d\log n + t}{n}} \quad \text{and} \quad \sup_{f\in\mathcal{F}} \frac{P_n f - Pf}{\sqrt{P_n f}} \le C\sqrt{\frac{d\log n + t}{n}}.$$

Let $\varepsilon_n = C\sqrt{(d\log n + t)/n}$ for shorthand. Then the second inequality is equivalent to the statement that for all $f \in \mathcal{F}$,

$$P_n f - Pf \le \frac{1}{2\eta} P_n f + \frac{\eta}{2}\varepsilon_n^2 \quad \text{for all } \eta > 0.$$

Rearranging the preceding display, we have

$$\left(1 - \frac{1}{2\eta}\right)(P_n f - Pf) \le \frac{1}{2\eta} Pf + \frac{\eta}{2}\varepsilon_n^2.$$

If $\sqrt{Pf} \ge \varepsilon_n$, we set $\eta = \sqrt{Pf}/\varepsilon_n$ and obtain $\frac{1}{2}(P_n f - Pf) \le \sqrt{Pf}\varepsilon_n$, while if $\sqrt{Pf} < \varepsilon_n$, then setting $\eta = 1$ yields $\frac{1}{2}(P_n f - Pf) \le \frac{1}{2}(Pf + \varepsilon_n^2) \le \frac{1}{2}(\sqrt{Pf}\varepsilon_n + \varepsilon_n^2)$. In either case,

$$P_n f - Pf \le C\left[\sqrt{Pf\frac{d\log n + t}{n}} + \frac{d\log n + t}{n}\right].$$

A symmetric argument replacing each P_n with P (and vice versa) gives the lemma. □

We can now demonstrate inequality (25). Let $V \subset \mathbb{R}^d$ and $\mathcal{V} := \{\{x \in \mathbb{R}^d \mid v^T x \le 0\}\}_{v \in V}$ the collection of halfspaces it induces. The collection $\mathcal{S} = \{S_{v,a,b}\}_{v \in V, a<b}$ of slabs has VC-dimension $\mathrm{VC}(\mathcal{S}) \le O(1)\mathrm{VC}(\mathcal{V})$. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $c : \mathcal{S} \to \mathbb{R}$ be arbitrary functions. If for $S \subset \mathcal{X}$ we define $S^+ := \{(x, y) \mid x \in S, f(x, y) \ge c(S)\}$ and the collection $\mathcal{S}^+ := \{S^+ \mid S \in \mathcal{S}\}$, then $\mathrm{VC}(\mathcal{S}^+) \le \mathrm{VC}(\mathcal{S}) + 1$ (Barber et al., 2019, Lemma 5). As a consequence, for the conformal sets inequality (25) specifies, for any $t > 0$ we have with probability at least $1 - e^{-t}$ that

$$\begin{aligned}
&\left|P_n(Y \in \widehat{C}(X), X \in S_{v,a,b}) - P(Y \in \widehat{C}(X), X \in S_{v,a,b})\right| \\
&\le O(1)\left[\sqrt{\min\{P(Y \in \widehat{C}(X), X \in S_{v,a,b}), P_n(Y \in \widehat{C}(X), X \in S_{v,a,b})\}\frac{\mathrm{VC}(\mathcal{V})\log n + t}{n}}\right] \\
&\qquad + O(1)\left[\frac{\mathrm{VC}(\mathcal{V})\log n + t}{n}\right]
\end{aligned}$$

simultaneously for all $v \in V, a < b \in \mathbb{R}$, and similarly

$$\begin{aligned}
&|P_n(X \in S_{v,a,b}) - P(X \in S_{v,a,b})| \\
&\le O(1)\left[\sqrt{\min\{P(X \in S_{v,a,b}), P_n(X \in S_{v,a,b})\}\frac{\mathrm{VC}(\mathcal{V})\log n + t}{n}} + \frac{\mathrm{VC}(\mathcal{V})\log n + t}{n}\right].
\end{aligned}$$

Now, we use the following simple observation. For any ε and nonnegative α, β, γ with $\alpha \le \gamma$ and $2|\varepsilon| \le \gamma$,

$$\left|\frac{\alpha}{\gamma + \varepsilon} - \frac{\beta}{\gamma}\right| \le \frac{|\alpha - \beta|}{\gamma} + \frac{|\alpha\epsilon|}{\gamma^2 + \gamma\epsilon} \le \frac{|\alpha - \beta|}{\gamma} + \frac{2|\epsilon|}{\gamma}.$$

Thus, as soon as $\delta \ge \frac{\mathrm{VC}(\mathcal{V})\log n + t}{n}$, we have with probability at least $1 - e^{-t}$ that

$$\begin{aligned}
&|P_n(Y \in \widehat{C}(X) \mid X \in S_{v,a,b}) - P(Y \in \widehat{C}(X) \mid X \in S_{v,a,b})| \\
&= \left|\frac{P_n(Y \in \widehat{C}(X), X \in S_{v,a,b})}{P_n(X \in S_{v,a,b})} - \frac{P(Y \in \widehat{C}(X), X \in S_{v,a,b})}{P(X \in S_{v,a,b})}\right| \\
&\le O(1)\left[\sqrt{\frac{\mathrm{VC}(\mathcal{V})\log n + t}{\delta n}} + \frac{\mathrm{VC}(\mathcal{V})\log n + t}{\delta n}\right]
\end{aligned}$$

simultaneously for all $v \in V, a < b \in \mathbb{R}$, where we substitute $\gamma = P(X \in S_{v,a,b})$, $\alpha = P(Y \in \widehat{C}(X), X \in S_{v,a,b})$, $\beta = P_n(Y \in \widehat{C}(X), X \in S_{v,a,b})$, and $\varepsilon = (P_n - P)(X \in S_{v,a,b})$. Note that if $\frac{\mathrm{VC}(\mathcal{V}) \log n + t}{\delta n} \geq 1$, the bound is vacuous in any case.

Appendix B. Additional Figures

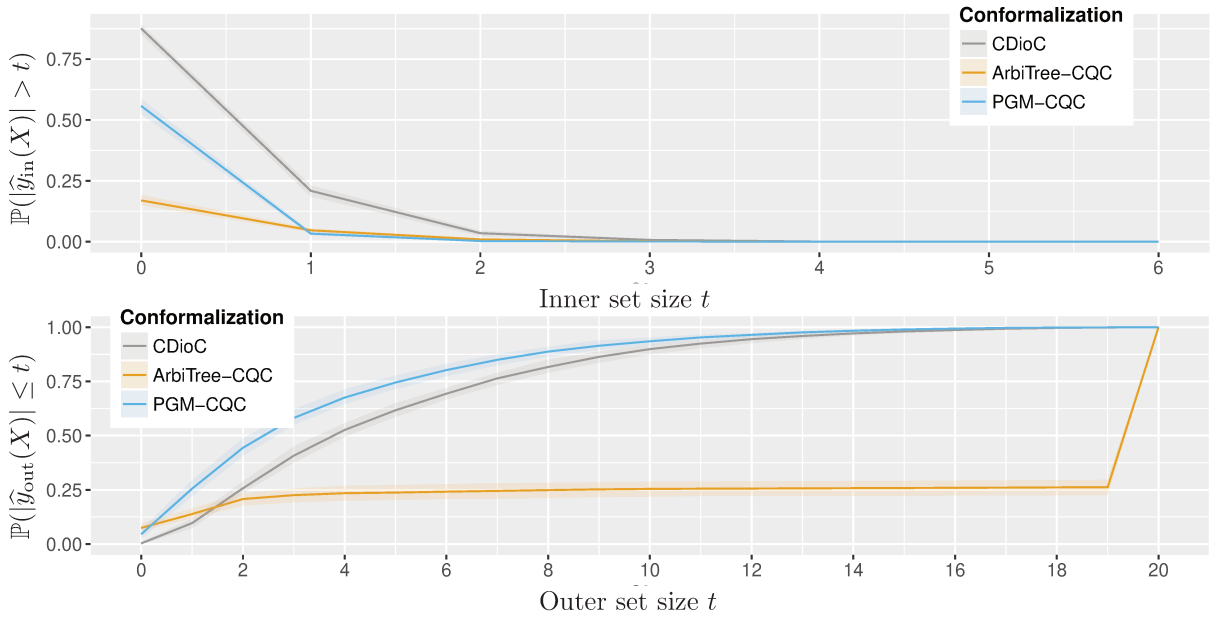

Figure 11. Results for PASCAL VOC data set (Everingham et al., 2012) over $M = 20$ trials. Methods are the conformalized direct inner/outer method (CDioC), Alg. 3; and tree-based methods with implicit confidence sets $\widehat{C}_{\mathrm{imp}}$ or explicit inner/outer sets $\widehat{C}_{\mathrm{io}}$, labeled ArbiTree and PGM (see description in Sec. 4.2.2). Top: 1− empirical c.d.f of sizes for the inner sets $\widehat{y}_{\mathrm{in}}$. Bottom: empirical c.d.f of sizes for the outer sets $\widehat{y}_{\mathrm{out}}$.

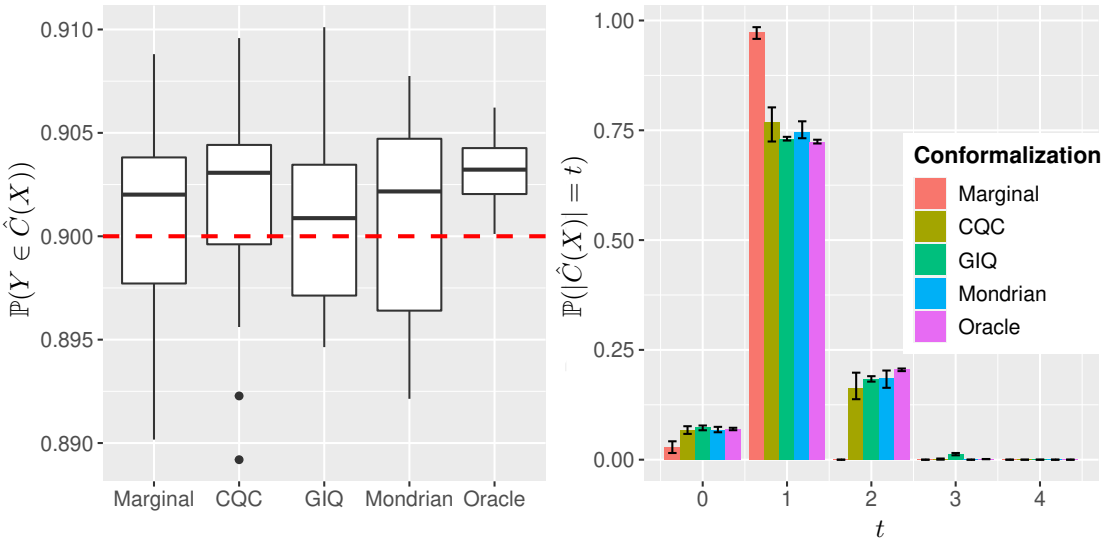

Figure 12. Marginal coverage and distribution of the confidence set size in the multiclass simulation (26) over $M = 20$ trials.

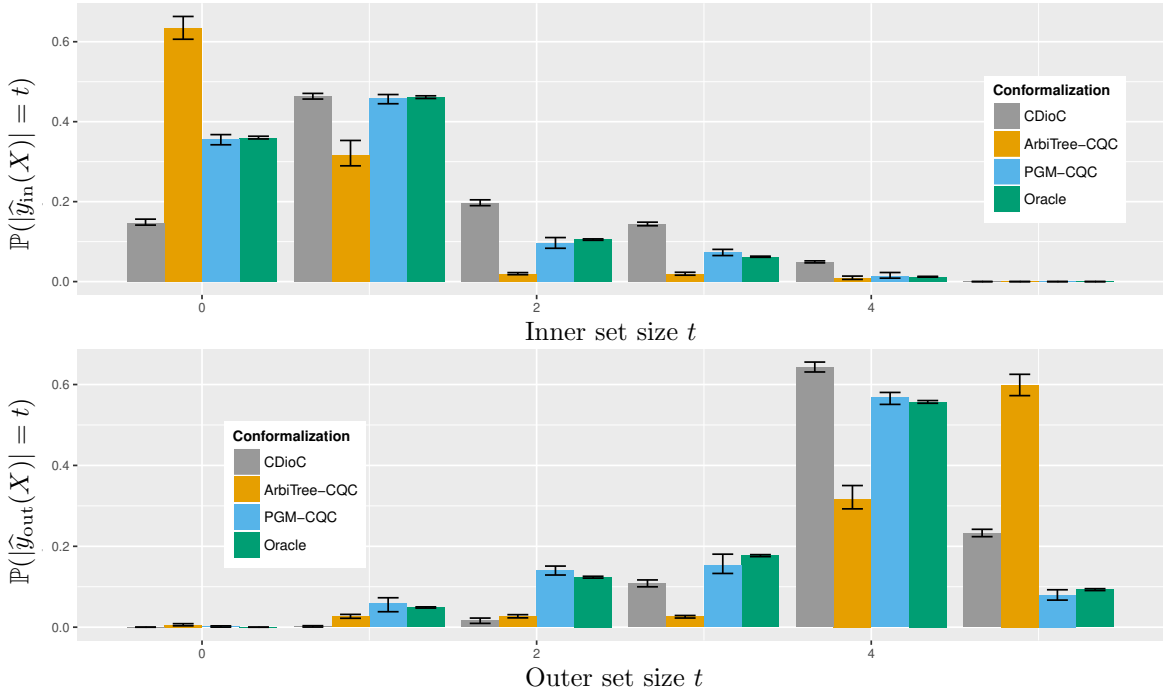

Figure 13. Simulated multilabel experiment with label distribution (27). Methods are the true oracle confidence set; the conformalized direct inner/outer method (CDioC), Alg. 3; and tree-based methods with implicit confidence sets $\widehat{C}_{\text{imp}}$ or explicit inner/outer sets $\widehat{C}_{\text{io}}$, labeled ArbiTree and PGM (see description in Sec. 4.2.2). Top: distribution of the inner sets $\widehat{y}_{\text{in}}$ sizes. Bottom: distribution of the outer sets $\widehat{y}_{\text{out}}$ sizes.

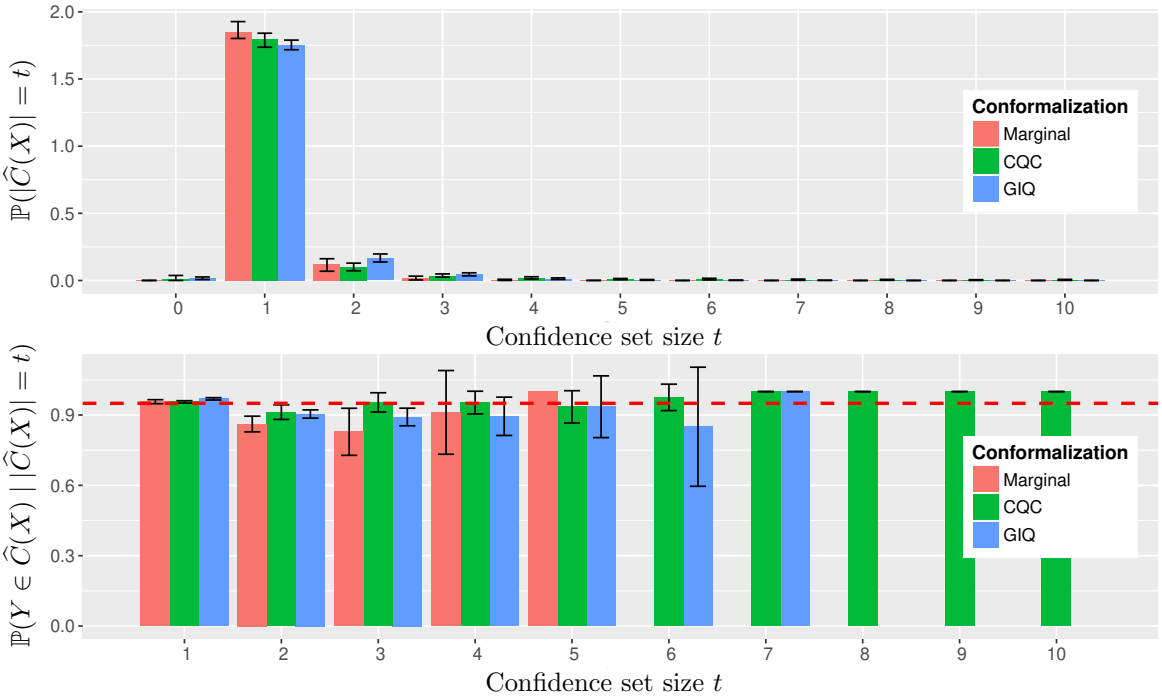

Figure 14. Results for CIFAR-10 data set over $M = 20$ trials. Methods are the marginal method (Marginal, procedure 3), the CQC method (Alg. 1), and the GIQ method (Alg. 1, Romano et al. (2020)). Top: Distribution of the confidence set size $|\widehat{C}(X)|$. Bottom: Probability of coverage conditioned on the size of the confidence set.

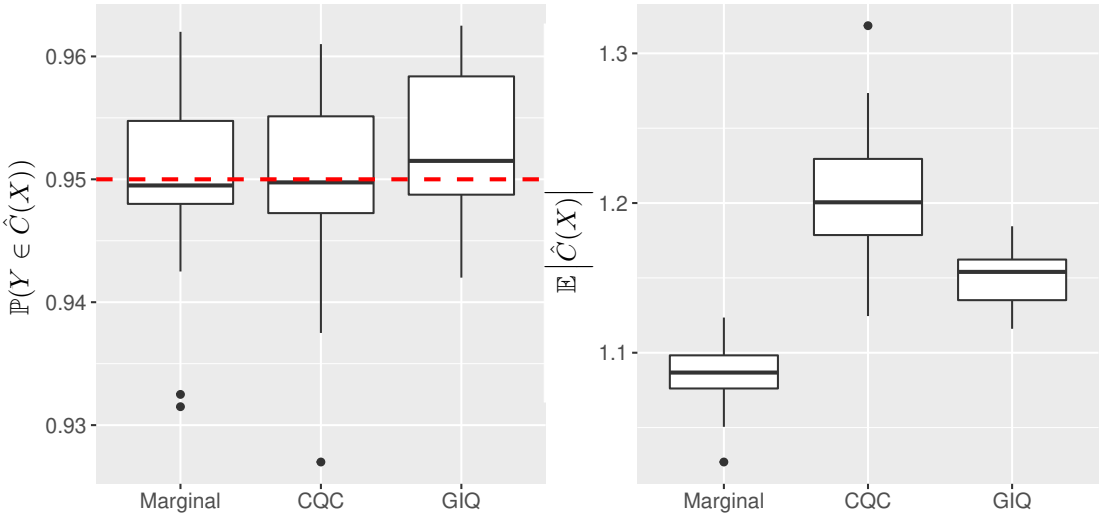

Figure 15. Marginal coverage and average confidence set size on CIFAR-10 over $M = 20$ trials.

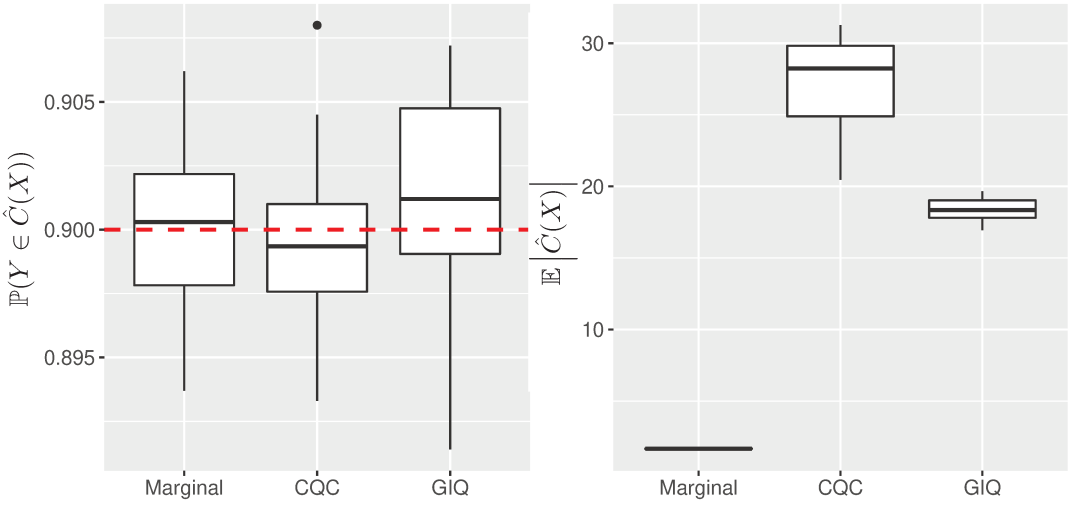

Figure 16. Marginal coverage and average confidence set size on ImageNet over $M = 20$ trials.

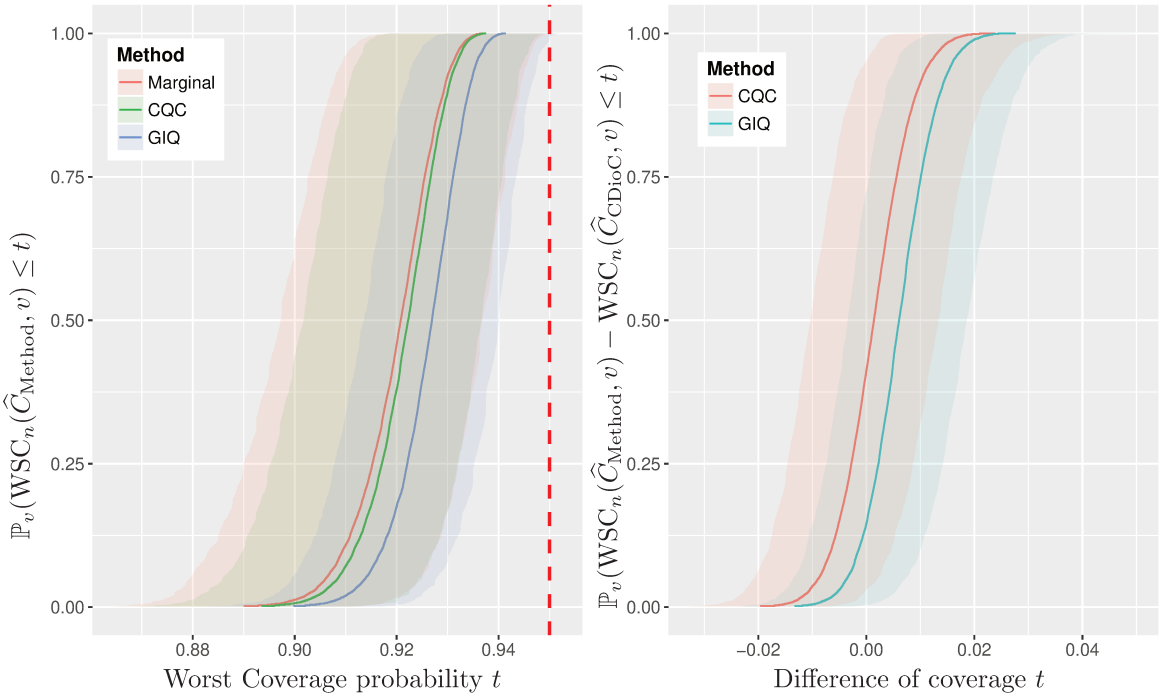

Figure 17. Worst-slab coverage for CIFAR-10 with $\delta = .2$ over $M = 1000$ draws $v \overset{\text{iid}}{\sim} \mathsf{Uni}(\mathbb{S}^{d-1})$. The dotted line is the desired (marginal) coverage. Left: distribution of worst-slab coverage. Right: distribution of the coverage difference $\mathrm{WSC}_n(\widehat{C}_{\mathrm{CQC}}, v) - \mathrm{WSC}_n(\widehat{C}_{\mathrm{Marginal}}, v)$.

References

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, to appear, 2019.

P. L. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

A. P. Dawid. Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2(1):25–36, 1992.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv:1810.08750 [stat.ML]*, 2018.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007. URL `http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html`.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012. URL `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Yotam Hechtlinger, Barnabás Póczos, and Larry Wasserman. Cautious deep learning. *arXiv:1805.09460 [stat.ML]*, 2019.

Francisco Herrera, Francisco Charte, Antonio J Rivera, and María J Del Jesus. Multilabel classification. In *Multilabel Classification*, pages 17–31. Springer, 2016.

Rafael Izbicki, Gilson Shimizu, and Rafael Stern. Flexible distribution-free conditional predictive bands using density estimators. In *Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics*, pages 3068–3077, 2020.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv:1607.01759 [cs.CL]*, 2016.

Nidhi Kalra and Susan M Paddock. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, 94:182–193, 2016.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1):33–50, 1978.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001.

Antonis Lambrou and Harris Papadopoulos. Binary relevance multi-label conformal predictor. In *Proceedings of the 5th International Symposium on Conformal and Probabilistic Prediction with Applications (COPA)*, pages 90–104. Springer, 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

Huma Lodhi, John Shawe-Taylor, Nello Cristianini, and Christopher J. C. H. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

Andrew McCallum and Kamal Nigam. Employing EM in pool-based active learning for text classification. In *Machine Learning: Proceedings of the Fifteenth International Conference*, 1998.

Kai min Chung and Hsueh-I Lu. An optimal algorithm for the maximum-density segment problem. In *Proceedings of the 11th Annual European Symposium on Algorithms*, 2003.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159, 2020.

Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. A deep neural network conformal predictor for multi-label text classification. In *Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 105 of *Proceedings of Machine Learning Research*, pages 228–245, 2019.

Harris Papadopoulos. A cross-conformal predictor for multi-label classification. In Lazaros Iliadis, Ilias Maglogiannis, Harris Papadopoulos, Spyros Sioutas, and Christos Makris, editors, *Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 241–250. Springer Berlin Heidelberg, 2014.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85:333–359, 2011.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32*, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems 33*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019.

Matteo Sesia and Emmanuel J. Candès. A comparison of some conformal quantile regression methods. *arXiv:1909.05433 [stat.ME]*, 2019.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 4278—-4284, 2017.

Mingxing Tan, Ruoming Pang, and Quoc V Le. EfficientDet: Scalable and efficient object detection. *arXiv:1911.09070 [cs.CV]*, 2019.

A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, 2012.

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

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Huazhen Wang, Xin Liu, Bing Lv, Fan Yang, and Yanzhu Hong. Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional chinese medicine. *PLOS ONE*, 9(6):1–14, 2014.

Huazhen Wang, Xin Liu, Ilia Nouretdinov, and Zhiyuan Luo. A comparison of three implementations of multi-label conformal prediction. In *Proceedings of the International Symposium on Statistical Learning and Data Sciences*, pages 241–250. Springer International Publishing, 2015.

Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.