Conformal prediction under ambiguous ground truth

Anonymous authors
Paper under double-blind review

Abstract

In safety-critical classification tasks, conformal prediction allows to perform rigorous uncertainty quantification by providing confidence sets including the true class with a user-specified probability. This generally assumes the availability of a held-out calibration set with access to ground truth labels. Unfortunately, in many domains, such labels are difficult to obtain and usually approximated by aggregating expert opinions. In fact, this holds true for almost all datasets, including well-known ones such as CIFAR and ImageNet. Applying conformal prediction using such labels underestimates uncertainty. Indeed, when expert opinions are not resolvable, there is inherent ambiguity present in the labels. That is, we do not have "crisp", definitive ground truth labels and this uncertainty should be taken into account during calibration. In this paper, we develop a conformal prediction framework for such ambiguous ground truth settings which relies on an approximation of the underlying posterior distribution of labels given inputs. We demonstrate our methodology on synthetic and real datasets, including a case study of skin condition classification in dermatology.

1 Introduction

In safety-critical applications it is insufficient to demonstrate good empirical predictive performance on a held-out test set. Many such application domains, such as medical diagnostics, require reasonable uncertainty estimates for decision making and benefit from statistical performance guarantees. Conformal prediction is a statistical framework allowing to quantify rigorously uncertainty by providing non-asymptotic guarantees without making any distributional and model assumptions. First introduced by Vovk et al. (2005), it has become very popular in machine learning as it is widely applicable and makes minimal assumptions, see e.g. (Romano et al., 2019; Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2021; Stutz et al., 2021; Fisch et al., 2022).

Here, we consider a classification task with K classes and a classifier $\pi : \mathcal{X} \to \Delta^K$ outputting the class probabilities where Δ^K is the K-simplex and $[K] := \{1, ..., K\}$. Based only on a held-out set of n calibration examples $(X_i, Y_i) \sim \mathbb{P}$, conformal prediction allows us to return for a test point (X, Y) (Y being unobserved) a confidence set $C(X) \subseteq [K]$ dependent on the calibration data such that

$$\mathbb{P}(Y \in C(X)) \geq 1 - \alpha, \tag{1}$$

whatever being $\mathbb{P}$ and π as long as the joint distribution of $((X_1, Y_1), ..., (X_n, Y_n), (X, Y))$ is exchangeable. Here $\alpha \in [0, 1]$ is a user-specified parameter and the probability in Equation (1) is not only over (X, Y) but also over the calibration set. The size of such confidence sets $|C(X)|$, also called *inefficiency*, is a good indicator of the uncertainty for X.

In order to apply conformal prediction, we need access to the label Y_i for each X_i in the calibration set. However, in most practical applications, such a label is obtained by aggregating several expert opinions. In medical applications, for example, it is generally not possible to identify the patient's actual condition as this would require invasive and expensive tests. Instead, labels are derived from expert annotations, e.g., doctor ratings (Liu et al., 2020). Although disagreements are common, researchers generally aggregate the annotations deterministically and declare, e.g., a top-1 aggregated or majority voted label to be the "true" ground truth. Formally, we consider $Y_i \sim \mathbb{P}_{\text{vote}}(\cdot|X = x_i)$ where $\mathbb{P}_{\text{vote}}(Y = y|X = x)$ is a one-hot distribution. This

is reasonable for some tasks, including popular benchmarks such as CIFAR10 (Krizhevsky, 2009; Peterson et al., 2019) and ImageNet (Russakovsky et al., 2015), because most examples are fairly unambiguous so that $\mathbb{P}(Y = y|X = x)$ is typically one-hot and a simple aggregation strategy can unambiguously determine the true class. However, in other tasks, there will be many examples where expert opinions are "irresolvable" (Schaekermann et al., 2016) and $\mathbb{P}_{\text{vote}}(Y = y|X = x)$ will differ very significantly from the true conditional distribution $\mathbb{P}(Y = y|X = x)$.

When using conformal prediction methods for such data, one does not obtain guarantees of the form in Equation (1) w.r.t. the underlying true distribution $\mathbb{P} = \mathbb{P}^X \otimes \mathbb{P}^{Y|X}$ but w.r.t. $\mathbb{P}_{\text{vote}} = \mathbb{P}^X \otimes \mathbb{P}^{Y|X}_{\text{vote}}$, i.e. a distribution dependent on the aggregation/voting strategy. As discussed further in Section 2.1 and illustrated in Figure 1, such a strategy can severely underestimate the true uncertainty. This is somewhat problematic as, conformal prediction being meant for uncertainty estimation in the first place, we would expect it to thrive in settings with ambiguous ground truth. Instead of summarizing the expert annotations by a single one-hot distribution, we use here a statistical model to obtain the *plausibilities* λ_y approximating $\mathbb{P}(Y = y|X = x)$. These plausibilities are then used to propose novel conformal prediction procedures.

First, we propose a procedure that does not yield confidence sets (i.e., sets of labels) but so-called *plausibility regions*, i.e., confidence intervals of plausibilities with associated coverage guarantees. We propose various strategies which allow to reduce these regions to sets of labels. Second, we consider a sampling-based approach, coined Monte Carlo conformal prediction, which allows us to use standard conformal prediction techniques to calibrate against sampled "pseudo ground truth" labels. We show how this approach can guarantee coverage despite not having access to exchangeable calibration examples. We discuss the relationship between both approaches and highlight advantages and limitations on a toy dataset with full information, i.e., known posterior distribution $\mathbb{P}(Y = y|X = x)$, and a real dataset for skin condition classification. Finally, we see how similar ideas can be extended to do conformal prediction in conjunction with data augmentation as well as multi-label classification.

Outline: The rest of this paper is structured as follows: In Section 2, we illustrate the problem on a toy dataset and introduce required background on conformal prediction. Then, Section 3 introduces our two original conformal approaches: confidence regions for plausibilities (Section 3.2) and Monte Carlo conformal prediction (Section 3.4). Section 3.5 highlights the applicability of our approaches to related problem and discusses related work. In Section 4, we present an application to skin condition classification following (Liu et al., 2020), multi-label classification and data augmentation before concluding in Section 5.

2 Motivation and Background

2.1 Motivational Example

In this section, we first consider a toy dataset to illustrate the impact a voting strategy can have on confidence sets produced by conformal prediction. The conditional density of X given $Y = y$ is assumed to be given by

$$p(x|y) = \mathcal{N}(x; \mu_y, \text{diag}(\sigma_y^2)), \tag{2}$$

with $\mu_y \in \mathbb{R}^d$ and $\text{diag}(\sigma_y^2) \in \mathbb{R}^{d\times d}$ being mean and variance. Further, one has $\mathbb{P}(Y = y) = w_y$ with $\sum_{y=1}^K w_y = 1$. To generate examples $(X, Y) \sim \mathbb{P}$, we first sample Y from $\mathbb{P}(Y = y) = w_y$ and then X from $p(x|y)$ so we have ground truth labels for each example. Furthermore, by Bayes' rule we have access to the true posterior $\mathbb{P}(Y = y|X = x)$. If the Gaussians are well-separated, $\mathbb{P}(Y = y|X = x)$ will be crisp, i.e., close to one-hot with low entropy. However, encouraging significant overlap between these Gaussians, e.g., by moving the means $(\mu_k)_{k\in[K]}$ close together, will result in highly ambiguous $\mathbb{P}(Y = y|X = x)$. We sample examples $Z_i = (X_i, Y_i)$ as outlined above and indicate classes by color in Figure 1 (left). The true posteriors $\mathbb{P}(Y_i = y|X = X_i)$ are also displayed, cf. Figure 1 (middle). As can be seen, these distributions can be very ambiguous in between all three classes. Let us assume a large set of annotators that allow us by majority vote to recover $\hat{Y}_i = \arg\max_{y\in[K]} \mathbb{P}(Y = y|X = X_i)$, i.e., the top-1 label, as shown on Figure 1 (right). Clearly, the top-1 labels ignore the underlying ambiguity.

We perform standard split conformal prediction using the top-1 labels (see Section 2.2 for details), randomly splitting the examples in two halves for calibration and testing. In the bottom, we plot the realized coverage,

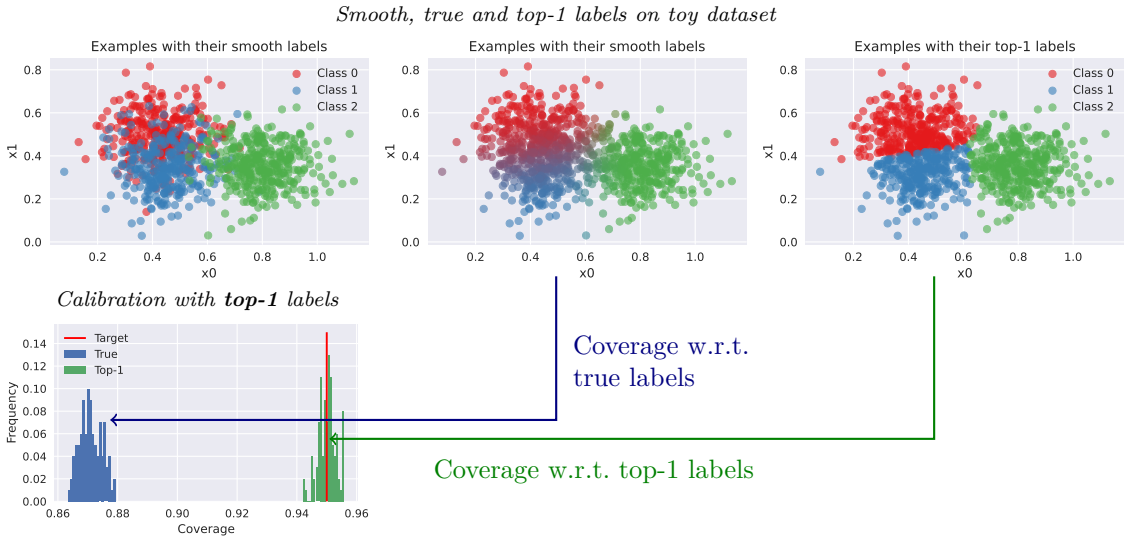

Figure 1: Illustration of an ambiguous problem with $K = 3$ classes, in two dimensions, using our toy dataset. Left: Examples colored by their *true* but usually unknown ground truth label. Note the high ambiguity between the classes. Middle: We color all examples according to their true posteriors $\mathbb{P}(Y = y|X = x)$ where faded colors represent higher ambiguity in $\mathbb{P}(Y = y|X = x)$. Right: Examples colored by their top-1 label $\hat{Y} = \arg\max_{y \in [K]} \mathbb{P}(Y = y|X = x)$. This leads to separable classes but clearly ignores the ambiguity of the problem. Bottom: Realized coverage over random calibration/test splits, i.e. the proportion of true or top-1 labels included in the constructed confidence sets $C(x)$, when calibrating against the top-1 labels following Section 2.2. Clearly, calibrating against top-1 labels produces confidence sets that *undercover* w.r.t. true labels.

i.e., the fraction of test examples for which (a) the true label (blue) or (b) the top-1 label (green) is included in the predicted confidence set. In this case, conformal prediction guarantees the latter, (b), to be 95% (on average across calibration/test splits). Strikingly, however, coverage against the (usually unknown) true labels is significantly worse. Of course, this gap depends on the ambiguity of the problem (see Appendix A for an illustration).

2.2 Conformal prediction

In the following, we briefly review standard (split) conformal prediction (Vovk et al., 2005; Papadopoulos et al., 2002). To this end, we assume a classifier $\pi_y(x) \approx \mathbb{P}(Y = y|X = x)$ approximating the posterior class probabilities is available. This model will be typically based on learned parameters using a training set. Then, given a set of calibration examples $(X_i, Y_i)_{i \in [n]}$ from $\mathbb{P}$, we want to construct a confidence set $C(X) \subseteq [K]$ for the test point (X, Y) such that the coverage guarantee from Equation (1) holds. As mentioned earlier, this requires the calibration examples and the test example to be exchangeable but does not make any further assumption on the data distribution or on the underlying model π.

A popular conformal predictor proceeds as follows: given a real-valued conformity score $E(X, k)$ based on the model predictions $\pi(x) \in \mathbb{R}^K$, we define

$$C(X) = \{k \in [K] : E(X, k) \geq \tau\} \tag{3}$$

where τ is the $\lfloor \alpha(n + 1) \rfloor$ smallest element of $\{E(X_i, Y_i)\}_{i \in [n]}$, equivalently τ is obtained by computing the $\lfloor \alpha(n + 1) \rfloor / n$ quantile of the distribution of the conformity scores of the calibration examples

$$\tau = Q\left(\{E(X_i, Y_i)\}_{i \in [n]}; \frac{\lfloor \alpha(n + 1) \rfloor}{n}\right). \tag{4}$$

Here, $Q(\cdot; q)$ denotes the q-quantile. By construction of the quantile, see e.g. (Romano et al., 2019; Vovk et al., 2005; Angelopoulos & Bates, 2021), this ensures that the lower bound on coverage in Equation (1) is

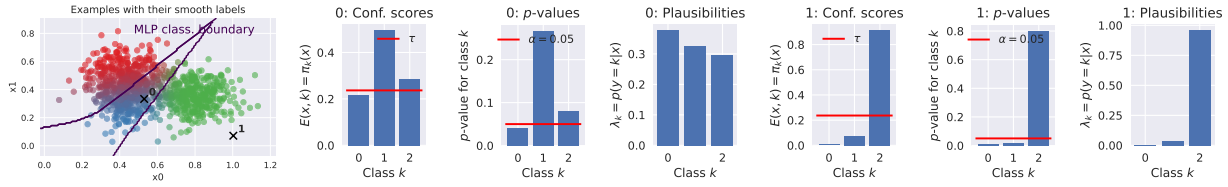

Figure 2: Illustration of conformal calibration on the toy dataset from Section 2.1 for two examples X, indexed 0 and 1. Here, the conformity scores $E(X, k)$, $k \in [K]$, are taken to be the predicted probabilities $\pi_k(x)$ of the MLP whose decision boundary is shown. To construct the confidence sets $C(X)$, these scores are thresholded using the threshold τ, cf. Equations (3) and (4). Alternatively, we can use the p-value associated formulation and threshold $(\rho_y)_{y\in[K]}$ at confidence level α. Calibrating against true labels, both approaches obtain coverage $1 - \alpha$ (w.r.t. the true labels). Figure 1, in contrast, shows that this is not the case when calibrating against the top-1 labels.

satisfied. Additionally, if the conformity scores are almost surely distinct, then we have the following upper bound $\mathbb{P}(Y \in C(X)) \leq 1 - \alpha + \frac{1}{n+1}$. The conformity score is a design choice. A standard choice, which we will use throughout the paper, is $E(x, k) = \pi_k(x)$ (Sadinle et al., 2019) but many alternative scores have been proposed in the literature (Romano et al., 2020; Angelopoulos et al., 2021).

An alternative view on conformal prediction can be obtained through a p-value formulation; see e.g. (Sadinle et al., 2019). The conformity scores of the calibration examples and test example (X, Y) are exchangeable so if the distribution of $E(X, Y)$ is continuous then

$$\rho_Y = \frac{|\{i \in [n] : E(X_i, Y_i) \leq E(X, Y)\}| + 1}{n + 1} = \frac{\sum_{i=1}^{n} \mathbb{I}[E(X_i, Y_i) \leq E(X, Y)] + 1}{n + 1} \tag{5}$$

is uniformly distributed over $\{1/(n + 1), 2/(n + 1)..., 1\}$ and thus ρ_Y is a p-value in the sense it satisfies $\mathbb{P}(\rho_Y \leq \alpha) \leq \alpha$, equivalently $\mathbb{P}(\rho_Y > \alpha) \geq 1-\alpha$. If the distribution is not continuous, then it can be checked that $\mathbb{P}(\rho_Y \leq \alpha) \leq \alpha$ still holds (see e.g. (Bates et al., 2023)). It follows directly that $C(X) := \{y \in [K] : \rho_y > \alpha\}$ satisfies $\mathbb{P}(Y \in C(X)) \geq 1 - \alpha$ and the confidence set obtained this way is identical to the one obtained using Equations (3) and (4). This mechanism is illustrated in Figure 2. In contrast to calibrating the threshold τ, this requires computing $(\rho_k)_{k\in[K]}$ for each test example X and is thus computationally more expensive. The p-value formulation will be useful in our context as p-values can easily be combined to obtain a new p-value; see e.g. (Vovk & Wang, 2020). For completeness, the equivalence between both formulations is detailed in Section B.1 of Appendix B.

3 Formulation and methods

This paper is about using plausibilities, i.e., approximations of the true posterior $\mathbb{P}(Y = y|X = x)$ obtained from expert annotations, to perform conformal prediction. The central goal is to obtain a coverage guarantee against the unknown labels by only using the plausibilities. We first present models used to obtain the plausibilities, the corresponding approximation of the class conditional probabilities $\mathbb{P}_{\text{agg}}(Y = y|X = x)$ and resulting coverage. This motivates the introduction of a novel "expected" conformity score. Then, we will focus on a scalable Monte Carlo, i.e., sampling-based conformal prediction approach.

3.1 From expert annotations to plausibilities and "*expected*" coverage

We assume that we have calibration data $X_i, B_i \sim \mathbb{P}$ where B_i corresponds to a set of expert annotations, the space of annotations being dependent on the application. For example, on CIFAR10-H (Peterson et al., 2019), each annotator provides a single label. In contrast, in our dermatology application, using data from (Liu et al., 2020), B_i represents a set of partial rankings whose cardinality depends on i. We then infer the plausibilities λ_i using a statistical model from (X_i, B_i)[1].

[1]We could in principle develop a conformal prediction method returning confidence sets for B, i.e., confidence sets of rankings, satisfying a coverage guarantee. However, modeling $\mathbb{P}(B = b|X = x)$ can be rather difficult given the format of B.

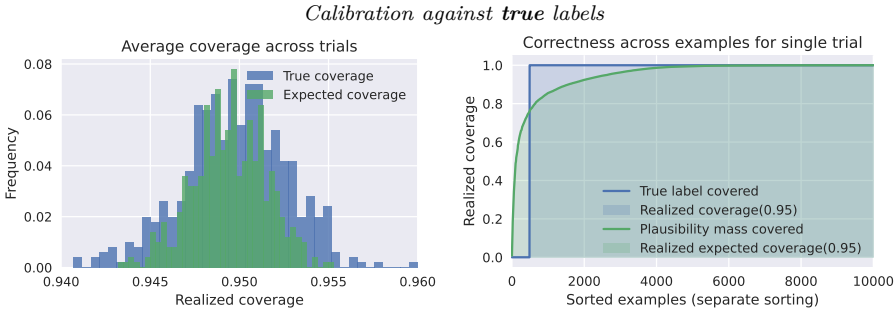

Figure 3: For the toy dataset from Figure 1, when calibrating against the true but unknown labels, we illustrate the definition of expected coverage w.r.t. the true posteriors in comparison to standard coverage w.r.t. the true ground truth labels. In practice, the true posterior is usually approximated by the plausibilities λ, while here we use $\lambda_k = p(Y = k|X)$. Left: In expectation across random calibration/test splits, both types of coverage coincide and achieve the target of $1 - \alpha = 95\%$. Right: However, on a per-example basis, expected coverage assigns values $\sum_{k\in[K]} \mathbb{P}_{\text{agg}}(Y = k|X)\mathbb{I}[k \in C(X)] \in [0, 1]$ while true coverage is binary. The area under these curves does usually not coincide exactly, but is close to $1-\alpha$ in this case and matches across calibration/test splits as shown on the left.

Assume the distribution of (X_i, B_i) admits a distribution $p(x, b)$. Then, the plausibilities λ are inferred from the annotations b and we will assume access to sample (X_i, λ_i) of density

$$p(x, \lambda) = \int p(x, \lambda, b)\mathrm{d}b = \int p(\lambda|b, x)p(x, b)\mathrm{d}b. \tag{6}$$

Then, our aggregation model for label Y can be written as

$$\mathbb{P}_{\text{agg}}(Y = y|X = x) = \int p(y|\lambda)p(\lambda|x)\mathrm{d}\lambda = \int\int \lambda_y p(\lambda|b, x)p(b|x)\mathrm{d}\lambda\mathrm{d}b. \tag{7}$$

We give examples of how $p(\lambda|b, x)$[2] can be modeled in Appendix C. The coverage guarantee we will provide for confidence sets for Y will thus be of the form $\mathbb{P}_{\text{agg}}(Y \in C(X)) \geq 1 - \alpha$ where $\mathbb{P}_{\text{agg}} = \mathbb{P}^X \otimes \mathbb{P}^{Y|X}_{\text{agg}}$ and

$$\mathbb{P}_{\text{agg}}(Y \in C(X)) = \mathbb{E}_{X\sim\mathbb{P}^X}\left[\mathbb{E}_{Y\sim\mathbb{P}_{\text{agg}}(\cdot|X)}[\mathbb{I}[Y \in C(X)]]\right]. \tag{8}$$

We will refer abusively to Equation (8) as "expected" coverage to emphasize that we average over the distribution $\mathbb{P}_{\text{agg}}(Y = y|X = x)$. Again, it is important to realize that the coverage in Equation (1) – what all conformal methods guarantee – is in most practical scenarios with respect to $\mathbb{P}_{\text{vote}} = \mathbb{P}^X \otimes \mathbb{P}^{Y|X}_{\text{vote}}$ where $\mathbb{P}^{Y|X}_{\text{vote}}$ is a one-hot distribution determined using a voting strategy given X, B. Thus, these methods ignore any ambiguity in $\mathbb{P}^{Y|X}$ as highlighted in Figure 1. Calibrating with $\mathbb{P}_{\text{agg}}$ holds the promise to take this ambiguity into account. We will assume from now on that the user has access to a set of exchangeable calibration data $(X_i, \lambda_i)_{i\in[n]}$ of examples and corresponding plausibilities from $p(x, \lambda)$.

3.2 Expected conformity scores and plausibility regions

Given the "expected" coverage from Equation (8), it appears natural to extend the basic conformity score of $E(x, k) = \pi_k(x)$ analogously. Specifically, we consider the "expected" conformity score

$$e(x, \lambda) := \sum_{k=1}^{K} \lambda_k E(x, k) = \mathbb{E}_{\mathbb{P}(k|\lambda)}[E(x, k)], \tag{11}$$

as $\mathbb{P}(Y = k|\lambda) = \lambda_k$. For the examples in Figure 2, the expected conformity score can be calculated using the dot product of predicted probabilities $\pi(x)$ and the plausibilities λ (where $\lambda_y = \mathbb{P}(Y = y|X)$ on the

[2]We consider simplified models where $p(\lambda|b, x) = p(\lambda|b)$.

Algorithm 1 Conformal prediction for plausibilities with $1-\alpha$ coverage guarantee on plausibility regions and a coverage guarantee on the confidence sets C depending on the reduction Ψ employed, see text.

Input: Calibration examples $(X_i, \lambda_i)_{i\in[n]}$; test example X; confidence level α; plausibility region to confidence set reduction Ψ; conformity score E
Output: Confidence plausibility region $c(X)$ and confidence set $C(X)$

1. Compute the "expected" conformity scores for calibration examples:

$$e(X_i, \lambda_i) = \sum_{k=1}^{K} \lambda_{ik} E(X_i, k). \tag{9}$$

2. Calibrate the threshold:

$$\tau = Q(\{e(X_i, \lambda_i\}_{i\in[n]}; \lfloor \alpha(n+1) \rfloor / n). \tag{10}$$

3. Define s grid points $\lambda^j \in \Delta^K, j \in [s]$ on the K-simplex and compute $e(X, \lambda^j)$, $j \in [s]$.

4. Compute $c(X) = \{j \in [s] : e(X, \lambda^j) \geq \tau\}$ and return $C(X) = \Psi(c(X))$.

toy example; $e(x, \lambda) \approx 0.33, 0.87$ for examples 0, 1). In Figure 4, we display the distributions of scores e and E, which are clearly different, on our toy dataset. Here, we know both the true posteriors and true labels that we can use to compute e and E, respectively. Using this expected conformity score, we can easily calibrate a threshold $\tau = Q(e(X_i, \lambda_i)_{i=1}^n; \lfloor \alpha(n+1) \rfloor / n)$. However, at test time, for example X with unknown plausibilities λ, conformal prediction returns a *plausibility region*

$$c(X) = \{\lambda : e(X, \lambda) \geq \tau\} \subseteq \Delta^K. \tag{12}$$

Thus, the coverage guarantee can be stated as follows:

$$\mathbb{P}(\lambda \in c(X)) \geq 1 - \alpha, \tag{13}$$

where the probability is w.r.t. Equation (6) and thus dependent on the statistical model relating expert opinions to plausibilities. The resulting algorithm is described in Algorithm 1.

This essentially resembles the confidence intervals for conformal regression and inherits the problem of actually enumerating all possible λ to compute $c(X)$ (Chen et al., 2018). In conformal regression, this is often approximated using a multi-dimensional grid. This can be shown to provide finite-sample guarantees depending on the grid size (Chen et al., 2018). Figure 5 shows the obtained plausibility regions $c(x)$ on the toy dataset for four selected examples, illustrated using samples from the plausibility regions in the 3-simplex. As can be seen, these plausibility regions capture the underlying ambiguity rather well. We emphasize the generality of this approach: any conformity score E for classification works and the grid points can be replaced with any other sampling strategy.

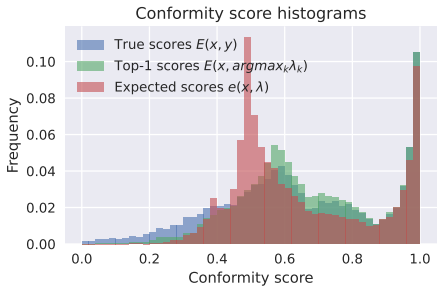

Figure 4: Histograms of conformity scores on the toy dataset, illustrating the difference between expected conformity scores $e(x, \lambda)$ from Equation (11), conformity scores w.r.t. the true but unknown labels $E(x, y)$ and w.r.t. the top-1 labels $E(x, \arg\max_k \lambda_k)$. Here, we use $E(x, k) = \pi_k(x)$ and assume that $\lambda_y = \mathbb{P}(Y = y|X)$. Clearly, the distribution of expected conformity scores is radically different from both the true and top-1 conformity scores.

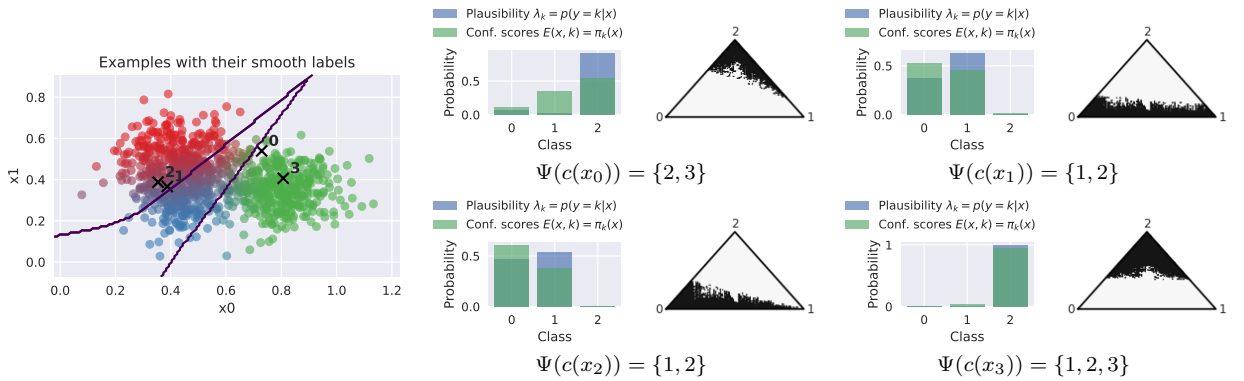

Figure 5: For four examples from our toy dataset, we show the plausibility regions with coverage $1-\alpha = 95\%$ obtained from the expected conformity scores in Figure 4. For ambiguous cases, the plausibility regions capture the underlying ambiguity. For less ambiguous cases, the plausibility regions are more concentrated in one corner of the 3-simplex. As outlined in Section 3.3, we can reduce the plausibility regions to confidence sets by including the top-1 labels $\arg\max_y \lambda_y$ for each $\lambda \in c(x)$. However, this often results in larger confidence sets, even in cases where the plausibility region is concentrated in one corner.

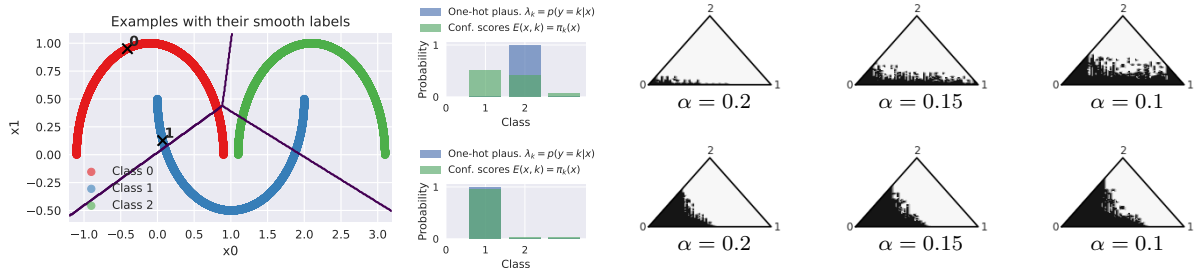

Figure 6: On a "three moons" dataset without ambiguous ground truth, we show that the conformity scores E and e are equal if $\mathbb{P}_{\text{agg}}(Y = y|X = x) = \mathbb{P}(Y = y|X = x)$ and this distribution is crisp for all examples x. However, this does not mean that plausibility regions, depending on the confidence level α are concentrated in the corners of the 3-simplex. This is because the plausibility regions mainly depend on the classifier output $\pi(x)$ through the conformity scores which, in this case, has difficulties separating the three moons linearly (violet). We show examples of plausibility regions for two points at three confidence levels α to illustrate this fact.

The question remains how these plausibility regions can be *reduced* to useful confidence sets, ideally with some type of coverage guarantees. To see why these reductions are generally non-trivial, we consider a simple case with crisp, one-hot posteriors; i.e. $\mathbb{P}(Y = y|X = x)$ is a one-hot distribution for all x. Then, as shown in Figure 6, the expected conformity scores e and standard conformity scores E coincide for $\lambda_y = \mathbb{P}_{\text{agg}}(Y = y|X = x) = \mathbb{P}(Y = y|X = x)$. As such, the calibrated thresholds τ (or equivalently any p-values) are also equal for both conformity scores e and E. This means that we can predict confidence sets C as usual. However, importantly, the plausibility regions c do not just contain one-hot plausibilities. Instead, c contains richer uncertainty estimates than C.

3.3 *Reducing* plausibility regions to confidence sets

While designing reductions from $c(x)$ to $C(x)$ is non-trivial, it also offers the opportunity to obtain application-tailored guarantees on $C(x)$ besides coverage. For example, if we are particularly interested in the top-1 labels of the plausibilities λ. In this case, we define

$$C(x) = \Psi(c(x)) = \left\{ y \in [K] : \left(\exists \lambda \in c(x) : y = \arg\max_{k \in [K]} \lambda_k \right) \right\}. \tag{14}$$

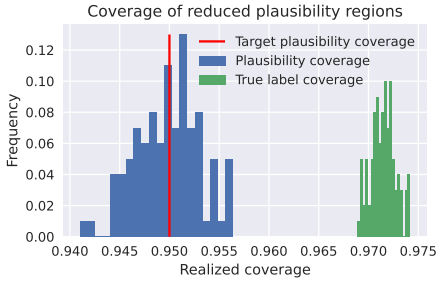

Figure 7: Coverage of the plausibility regions, i.e., whether $\lambda \in c(x)$, alongside coverage of the reduced confidence sets $\Psi(c(x))$ from Equation (14) w.r.t. to the true but unknown labels. As can be seen, the plausibility regions obtain the expected coverage of $1-\alpha$ w.r.t. the plausibilities, but the reduced plausibility regions over-estimate true coverage significantly.

Algorithm 2 Monte Carlo conformal prediction with $1-\alpha$ coverage guarantee for $m = 1$ and $1-2\alpha$ for $m \geq 2$.

Input: Calibration examples $(X_i, \lambda_i)_{i\in[n]}$; test example X; confidence level α; number of samples m
Output: Confidence set $C(X)$ for test example

1. Sample m labels $(Y_i^j)_{j\in[m]}$ per calibration example $(X_i)_{i\in[n]}$ where $\mathbb{P}(Y_i^j = k) = \lambda_{ik}$.
2. Calibrate the threshold τ using
$$\tau = Q\left(\{E(X_i, Y_i^j)\}_{i\in[n],j\in[m]}; \frac{\lfloor \alpha m(n+1) \rfloor - m + 1}{mn}\right). \tag{16}$$
3. Return $C(X) = \{k \in [K] : E(X, k) \geq \tau\}$.

In words, we include every class $y \in [K]$ in the confidence set $C(x)$ for which a corresponding plausibility vector $\lambda \in c(x)$ exists such that y is the arg max label, i.e., $y = \arg\max_{k\in[K]} \lambda_k$. As this only considers the top-1 labels of the plausibilities, we do likely not obtain an expected coverage guarantee as in Equation (1). However, we can use the coverage guarantee on plausibility regions from Equation (13) to derive a guarantee on the reduced top-1 labels of the following form:

$$\mathbb{P}\left((\arg\max_{k\in[K]} \lambda_k) \in C(X)\right) \geq \mathbb{P}(\lambda \in c(X)) \geq 1-\alpha. \tag{15}$$

Note that, if the plausibilities are crisp, we have $\mathbb{P}_{\text{agg}}(Y \in C(X)) = \mathbb{P}((\arg\max_{k\in[K]} \lambda_k) \in C(X))$ because $Y = \arg\max_{k\in[K]} \lambda_k$ in all cases. Examples for this simple reduction are given in Figure 5. Note that even for unambiguous examples where the plausibility region is concentrated in one corner of the 3-simplex, the obtained reduction $\Psi(c(x))$ may contain multiple classes. In terms of coverage, Figure 7 shows that the plausibility regions obtain the target coverage of 95% fairly closely. The reduced confidence sets $C(X) = \Psi(c(x))$, in contrast, over-estimate coverage (of the true label).

This is only a single example of how to construct such reductions in order to obtain specific coverage guarantees. As we have total freedom of how we define Ψ, Appendix C includes several additional examples. For example, we can extend Equation (14) to a top-k reduction, providing coverage guarantees on top-k label sets, or even consider *ordered* top-k lists, providing guarantees on rankings of labels. Such approaches might be particularly interesting in applications with ambiguous ground truth if very specific guarantees are needed.

3.4 Monte Carlo conformal prediction

To avoid working with plausibility regions as in Algorithm 1, we propose Monte Carlo conformal prediction, a sampling-based approach to conformal prediction under ambiguous ground truth. Given the calibration examples $(X_i, \lambda_i)_{i\in[n]}$, we sample m labels Y_i^j for each X_i according to plausibilities λ_i, i.e. $\mathbb{P}(Y_i^j = k) = \lambda_{ik}$ and duplicate the corresponding inputs. That is, we obtain $m \cdot n$ new calibration examples $(X_i, Y_i^j)_{i\in[n],j\in[m]}$

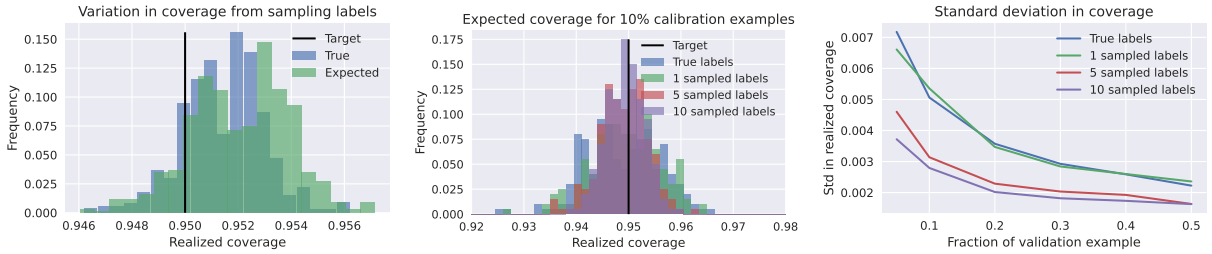

Figure 8: Left: Coverage w.r.t. the true labels as well as "expected" coverage for Monte Carlo conformal prediction with a fixed calibration/test split but different randomly sampled labels Y_i^j, cf. Algorithm 2. The approach obtains empirically expected coverage $1-\alpha$ across calibration/test splits but overestimates coverage slightly for this particular split. However, coverage is marginal across not only test and calibration examples but also the sampled labels during calibration. Middle and right: Fixing the sampled labels for different m, instead, and plotting variation in coverage across random calibration/test split shows that larger standard conformal prediction against the true unknown labels as well as Monte Carlo conformal prediction with $m \geq 1$ sampled labels obtain empirically expected coverage $1-\alpha$ on average. However Monte Carlo conformal prediction reduces the variation in *expected* coverage observed across calibration/test splits as m increases. We consider 5% to 50% calibration examples.

and then apply a conformal calibration outlined in Algorithm 2. This approach has the major advantage that we directly obtain confidence sets $C(x) \subseteq [K]$ rather than plausibility regions $c(x) \subseteq \Delta^K$.

For $m = 1$, it is clear that this approach boils down to standard conformal prediction and guarantees $\mathbb{P}_{\text{agg}}(Y \in C(X)) \geq 1-\alpha$. This "expected" coverage guarantee is not only marginal across test example (X, Y) and calibration examples $(X_i)_{i\in[n]}$, but also marginal across the sampled labels $(Y_i^j)_{i\in[n],j\in[m]}$. For $m \geq 2$, calibration examples include m repetitions of X_i so they are not exchangeable with the test example X which invalidates the standard argument used by conformal prediction to obtain a coverage guarantee. We ignore this point for the time being and illustrate the procedure in Figure 8 (left) where we show (expected) coverage w.r.t. the true (unknown) labels and the plausibilities for a *fixed* calibration/test split but across random label samples. That is, the variation in realized coverage comes purely from sampling different Y_i^j, while the mode of the distribution is influenced by the fixed calibration/test split (over-estimate coverage slightly in this particular case). Across calibration/test splits, we observe that Monte Carlo conformal prediction obtains expected coverage $1-\alpha$ (middle), but we found this often comes at the expense of larger confidence sets compared to calibration against top-1 labels. This is the "cost" of taking ambiguous ground truth into account. We also observe that the variability of the coverage across different calibration/test splits and sampled labels Y_i^j (right) decreases as m increases.

The following section focuses on establishing rigorous coverage guarantees for Monte Carlo conformal prediction when $m \geq 2$, showing that we can establish $\mathbb{P}_{\text{agg}}(Y \in C(X)) \geq 1-2\alpha$ without modifications. However, as pointed out above, we empirically observe a $1-\alpha$ coverage reduction for $m \gg 2$. In fact, we were unable to produce a statistically significant coverage gap in all our experiments, even when trying to construct the calibration examples in a particularly non-exchangeable way. This is akin to the observation in (Barber et al., 2021) that jackknife+, cross-conformal or out-of-bag conformal prediction consistently obtain empirical coverage $1-\alpha$ while they only guarantee $1-2\alpha$. However, if rigorous coverage guarantees better than $1-2\alpha$ are necessary, we show that this can be achieved by utilizing an additional split of the calibration examples.

3.4.1 Monte Carlo conformal prediction as averaging p-values

In order to establish a more conservative coverage guarantee of $\mathbb{P}_{\text{agg}}(Y \in C(X)) \geq 1-2\alpha$, let us introduce for $j \in [m]$

$$\rho_Y^j = \frac{\sum_{i=1}^n \mathbb{I}[E(X_i, Y_i^j) \leq E(X, Y)] + 1}{n+1}. \tag{17}$$

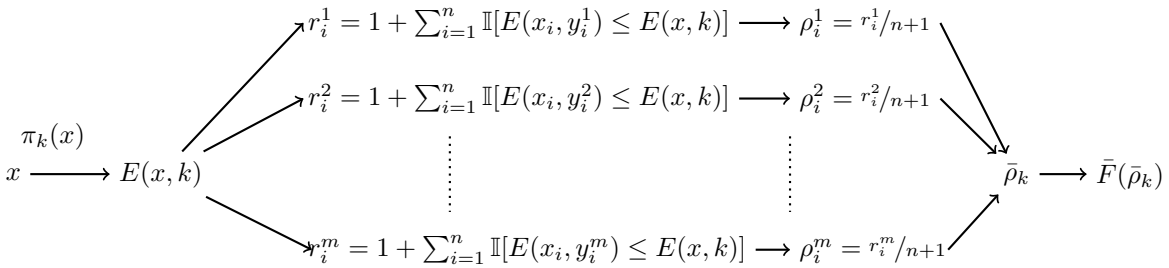

Figure 9: The m label samples result in $(\rho_k^j)_{j\in[m]}$ for the test example X, $(\rho_Y^j)_{j\in[m]}$ being p-values. We can average ρ_k^j over j and threshold the resulting average $\bar{\rho}_k$ at level α to obtain the confidence set. Alternatively we can combine the p-values ρ_Y^j using the ECDF approach described in Section 3.4.2; see Algorithm 3 for a detailed computational description.

The random variables ρ_Y^j are p-values, i.e. $\mathbb{P}(\rho_Y^j \leq \alpha) \leq \alpha$, since the scores $\{E(X_i, Y_i^j)\}_{i\in[n]}$ and $\{E(X,Y)\}$ are exchangeable for fixed $j \in [m]$. We can now average these quantities over $j \in [m]$ to obtain

$$\bar{\rho}_Y = \frac{1}{m}\sum_{j=1}^m \rho_Y^j = \frac{\sum_{j=1}^m \left(\sum_{i=1}^n \mathbb{I}[E(X_i, Y_i^j) \leq E(X,Y)] + 1\right)}{m(n+1)}. \tag{18}$$

As $\bar{\rho}_Y$ is an average of (dependent) p-values, it follows from standard results (Rüschendorf, 1982; Meng, 1994) that $\mathbb{P}(2\bar{\rho}_Y \leq 2\alpha) \leq 2\alpha$, equivalently $\mathbb{P}(\bar{\rho}_Y > \alpha) \geq 1 - 2\alpha$. Hence we have $\mathbb{P}(Y \in C(X)) \geq 1 - 2\alpha$ for $C(X) = \{y \in [K] : \bar{\rho}_y > \alpha\}$. Nevertheless, as discussed previously, we obtain empirical coverage close to $1 - \alpha$, cf. Figure 8, see also the discussion in (Barber et al., 2021, Section 4). This approach is illustrated in Figure 9. Note that in the context of aggregating m different models, (Linusson et al., 2017) also considered similar types of averaging. While it is generally not possible to improve the $1-2\alpha$ guarantee without making additional assumptions on the p-values, Section 3.4.2 shows that we can improve this guarantee at the cost of an additional data split.

Practically, we do not have to compute and actually average ρ_y^j for $j \in [m]$ and $y \in [K]$ to determine $C(X) = \{y \in [K] : \bar{\rho}_y > \alpha\}$. As in Section 2.2, we can reformulate this confidence set as $C(X) = \{k \in [K] : E(X,k) \geq \tau\}$ as in Equation (3) where τ is the $\lfloor \alpha m(n+1) \rfloor - m + 1$ smallest element of $\{E(X_i, Y_i^j)\}_{i\in[n], j\in[m]}$, equivalently τ is obtained by computing the $\frac{\lfloor \alpha m(n+1) \rfloor - m + 1}{mn}$ quantile of the distribution of these scores, i.e. $C(X) = \{k \in [K] : E(X,k) \geq \tau\}$ for

$$\tau = Q\left(\{E(X_i, Y_i^j)\}_{i\in[n], j\in[m]}; \frac{\lfloor \alpha m(n+1) \rfloor - m + 1}{mn}\right). \tag{19}$$

This is established in Section B.2 of Appendix B.

3.4.2 Beyond $1 - 2\alpha$ coverage guarantees

The $1 - 2\alpha$ coverage guarantee from Monte Carlo conformal prediction arises from the fact that we use a standard result about average of p-values (Rüschendorf, 1982; Meng, 1994). When the p-values are independent, a few techniques have been proposed to get back to a $1 - \alpha$ coverage; see e.g. (Cinar & Viechtbauer, 2022) for a comprehensive review. However, in our case, the p-values $(\rho_Y^j)_{j\in[m]}$ we want to combine are strongly dependent as they use the same calibration examples X_i and then rely on different pseudo-labels Y_i^j from the same distribution (given by λ_i). In this setting, many standard methods yield overly conservative results.

We follow here a method that directly estimates the cumulative distribution function (CDF) of the combined, e.g., averaged p-values (Balasubramanian et al., 2015; Toccaceli & Gammerman, 2019; Toccaceli, 2019). For example, let $\bar{\rho}_Y = {}^1/_m \sum_{j=1}^m \rho_Y^j$ be the averaged p-values. As shown in Figure 10 (left), these averaged

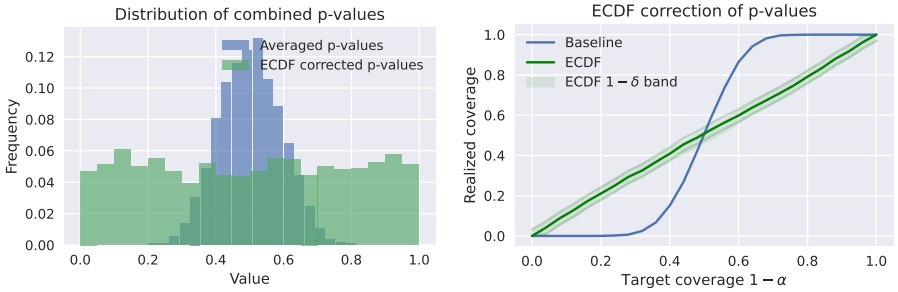

Figure 10: For illustration, we sample 10k p-values for $m = 10$ independent tests and duplicate them to obtain $m = 20$ dependent tests. Left: Histograms for averaged and ECDF-corrected p-values. The ECDF correction is able to ensure that p-values are distributed approximately uniformly. Right: Target confidence level α plotted against the empirical confidence level when calibrating with averaged p-values (blue) and the ECDF-corrected ones (green). As the p-values are one-dimensional, the Dvoretzky–Kiefer–Wolfowitz theorem provides tight finite-sample guarantees on the ECDF correction, using $\delta = 0.0001 = 0.01\%$.

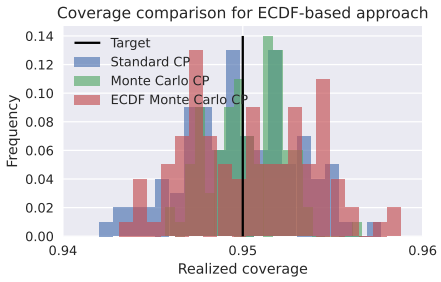

Figure 11: On our toy dataset, we plot coverage obtained using standard calibration with true labels compared to Monte Carlo and ECDF Monte Carlo conformal prediction for $m = 10$ and $l = \lfloor n/2 \rfloor$. As expected, Monte Carlo conformal prediction, even for $m > 1$, does not result in reduced coverage. Thus, ECDF correction does not yield meaningfully different results.

p-values will not be uniformly distributed. However, $\rho_Y = F(\bar{\rho}_Y)$ is a p-value[3] with F being the CDF of $\bar{\rho}_Y$ and thus the confidence set $C(X) = \{y \in [K] : \rho_y > \alpha\}$ will obtain coverage $1 - \alpha$. The true CDF is unknown but we can obtain an empirical estimate $\bar{F}$ by considering a split $X_1, \ldots, X_l$ and $X_{l+1}, \ldots, X_n$ of the original calibration examples. The procedure is described in Algorithm 3.

As $\bar{F}$ is not the true CDF F but an empirical CDF (ECDF) estimate, we only obtain an approximate coverage guarantee at level $1 - \alpha$. However, if the original calibration examples $X_{l+1}, \ldots, X_n$ are not only exchangeable, but also i.i.d., then the Dvoretzky–Kiefer–Wolfowitz theorem (see e.g. (Wasserman, 2006)) provides rigorous finite sample guarantees for the ECDF, i.e.

$$\mathbb{P}(\sup_{f \in [0,1]} |\bar{F}(f) - F(f)| \geq \epsilon) \leq 2 \exp(-2(n-l)\epsilon^2). \tag{20}$$

Basic manipulation results in a confidence band for $\bar{F}$

$$\mathbb{P}(\forall f \in [0,1], \bar{F}_-(f) \leq F(f) \leq \bar{F}_+(f)) \geq 1 - \delta \quad \text{with} \quad \bar{F}_\pm(f) = \frac{\min}{\max} \left\{ \bar{F}(f) \pm \sqrt{\frac{1}{2(n-l)} \log \frac{2}{\delta}}, \frac{1}{0} \right\}. \tag{21}$$

This confidence band is illustrated shown in Figure 10 (right). We can now define the confidence set $C(X) = \{y \in [K] : \rho_y > \alpha\}$ for $\rho_Y = \bar{F}_+(\bar{\rho}_Y)$ and show that it satisfies $\mathbb{P}_{\text{agg}}(Y \in C(X)) \geq (1 - \alpha)(1 - \delta)$. Indeed, writing $\bar{F}_+ \geq F$ to abbreviate $\bar{F}_+(f) \geq F(f)\ \forall f$, we have $\mathbb{P}(\bar{F}_+(\bar{\rho}_Y) > \alpha) \geq \mathbb{P}(\bar{F}_+(\bar{\rho}_Y) > \alpha | \bar{F}_+ \geq F)\mathbb{P}(\bar{F}_+ \geq F) \geq \mathbb{P}(F(\bar{\rho}_Y) > \alpha | \bar{F}_+ \geq F)\mathbb{P}(\bar{F}_+ \geq F)$. Now we have from Equation (20) that $\mathbb{P}(\bar{F}_+ \geq F) \geq 1 - \delta$ and $\mathbb{P}(F(\bar{\rho}_Y) > \alpha | \bar{F}_+ \geq F) = \mathbb{P}(F(\bar{\rho}_Y) > \alpha)$ as the probability in Equation (20) is over $(X_i, Y_i)_{i \in \{l+1, \ldots, n\}}$ while $\bar{\rho}_Y$ is only a function of $(X_i, Y_i^j)_{i \in [l], j \in [m]}$ and (X, Y). Using the fact that $\mathbb{P}(F(\bar{\rho}_Y) > \alpha) \geq 1 - \alpha$, we obtain the desired coverage guarantee.

[3] We have $\mathbb{P}(F(\bar{\rho}_Y) \leq f) = \mathbb{P}(F(F^{-1}(U) \leq f)$ for U a uniform random variable on $[0, 1]$ for $F^{-1}(f) = \inf\{\rho \in \mathbb{R} : F(\rho) \geq f\}$. However $F(F^{-1}(U)) \geq U$ so $\mathbb{P}(F(F^{-1}(U) \leq f) \leq \mathbb{P}(U \leq f) = f$. Hence, we have $\mathbb{P}(\rho_Y \leq \alpha) \leq \alpha$.

Algorithm 3 ECDF Monte Carlo conformal prediction with $(1-\alpha)(1-\delta)$ coverage guarantee.

Input: Calibration examples $(X_i, \lambda_i)_{i\in[n]}$; test example X; confidence levels α, δ; data split $1 \leq l \leq n-1$; number of samples m

Output: Confidence set $C(X)$ for test example X

1. Sample m labels $(Y_i^j)_{j\in[m]}$ per calibration example $(X_i)_{i\in[l]}$ where $\mathbb{P}(Y_i^j = k) = \lambda_{ik}$.
2. Sample one label Y_i per calibration example $(X_i)_{i\in\{l+1,\dots,n\}}$ where $\mathbb{P}(Y_i = k) = \lambda_{ik}$.
3. Compute $(\bar{\rho}^i)_{i\in\{l+1,\dots,n\}}$ where
$$\bar{\rho}^i = \frac{\sum_{j=1}^{m}\left(\sum_{p=1}^{l}\mathbb{I}[E(X_p, Y_p^j) \leq E(X_i, Y_i)] + 1\right)}{m(l+1)}. \tag{22}$$
4. Build the ECDF $\bar{F}(f) = \frac{1}{n-l}\sum_{i=l+1}^{n}\mathbb{I}[\bar{\rho}^i \leq f]$ and its upper bound $\bar{F}_+(f)$ using (21).
5. For test example X, compute for $k \in [K]$
$$\bar{\rho}_k = \frac{\sum_{j=1}^{m}\left(\sum_{p=1}^{l}\mathbb{I}[E(X_p, Y_p^j) \leq E(X, k)] + 1\right)}{m(l+1)}, \quad \bar{\rho}_k^{\text{corr}} = \bar{F}_+(\bar{\rho}_k). \tag{23}$$
6. Return $C(X) = \{k \in [K] : \bar{\rho}_k^{\text{corr}} > \alpha\}$.

Figure 11 compares the empirical coverage obtained using standard conformal prediction with true labels to both approaches of Monte Carlo conformal prediction (cf. Algorithms 2 and 3) to verify that this approach indeed obtains coverage $1-\alpha \approx (1-\alpha)(1-\delta)$ on average. As discussed, without ECDF correction from above, this is only empirical as we can only guarantee $1-2\alpha$. We also did not observe a significant difference in confidence set size when using or not using the ECDF correction.

3.5 Discussion

We presented two approaches to perform conformal prediction under ambiguous ground truth: estimating plausibility regions and reducing them to confidence sets and Monte Carlo conformal prediction. While the former allows to construct very specific guarantees, it also requires introducing a grid on the plausibility space to determine the corresponding confidence regions. Monte Carlo conformal prediction, instead, offers a $1-2\alpha$ coverage guarantee out-of-the-box (and $1-\alpha$ for $m=1$) which can be improved to $1-\alpha$ with an additional calibration split. Monte Carlo conformal prediction is, however, dependent on the sampled calibration labels and the coverage guarantee has to be understood marginally across calibration examples *and* their sampled labels. We now show that these developments are relevant beyond the setting of ambiguous ground truth and discuss related work.

3.5.1 Applications to related problems

Multi-label classification: Let $\mathcal{Y}_i \subseteq [K]$ be the unknown, ground truth multi-label *set* for each calibration example X_i. For simplicity, we express the multi-label setting using plausibilities λ_i that divide probability mass equally across the $L_i = |\mathcal{Y}_i| \geq 1$ labels. We can then apply Monte Carlo conformal prediction to this setup. When $m \gg 1$, we obtain calibration examples where the proportion of $(Y_i^j)_{j\in[m]}$ equal to a given class is approximately $1/L_i$. This is related to an empirical existing method to perform multi-label conformal prediction (Tsoumakas & Katakis, 2007; Wang et al., 2014; 2015) where for each calibration example $(X_i, \mathcal{Y}_i)$ we perform conformal prediction using the calibration examples $(X_i, Y_i^1), \dots, (X_i, Y_i^{L_i})$.

Monte Carlo conformal prediction provides coverage guarantees on $\mathbb{P}_{\text{agg}}(Y \in C(X))$ for

$$\mathbb{P}_{\text{agg}}(Y = y|X = x) = \sum_{\mathcal{Y}} \lambda(\mathcal{Y})_y \; p(\mathcal{Y}|x) \tag{24}$$

with $p(\mathcal{Y}|x)$ being the conditional probability of the set $\mathcal{Y}$ given $X = x$ and $\lambda(\mathcal{Y})_y = \mathbb{I}(y \in \mathcal{Y})/|\mathcal{Y}|$.

An alternative approach to multi-label conformal prediction is operating on the power-sets (Papadopoulos, 2014; Cauchois et al., 2021). This is similar, in spirit, to our plausibility region based approach. Instead of calibrating using conformity scores $E(X_i, Y_i^j)$, a conformity score $e(X_i, \mathcal{Y}_i)$ defined on the set of labels in the set $\mathcal{Y}_i$ is used. Defining the plausibilities λ as above, we could use the expected score $e(X_i, \mathcal{Y}_i) = \sum_{y \in \mathcal{Y}_i} E(X_i, y)$. This results in the prediction set $c(X)$ including sets of labels instead of individual labels (or plausibilities). Then, we can apply reductions Ψ that map a set of sets of labels to a confidence set of labels. Again, Ψ can be used to obtain very specific guarantees, similar to (Fisch et al., 2022).

Data augmentation and robustness: Consider a scenario where we have exchangeable calibration data $(X_i, Y_i)_{i \in [n]}$ for $Y_i \in [K]$. We want to augment the set of calibration data by using data augmentation, i.e. for each $X_i^1 := X_i$ we sample additional $X_i^2, ..., X_i^m \sim p(\cdot|X_i)$. For example, these could correspond to versions of X_i which are (adversarially or randomly) perturbed or corrupted, rotated, flipped, etc. As we usually train with data augmentation and frequently want to improve robustness against distribution shifts or specific types of perturbations, considering these augmentations for calibration is desirable. However, performing conformal prediction using the $m \cdot n$ calibration data $(X_i^j, Y_i)_{i \in [n], j \in [m]}$ breaks exchangeability. To address this, a procedure very similar to Monte Carlo conformal prediction allows us to use these augmented inputs for calibration. It proceeds by averaging the following m p-values

$$\rho_Y^j = \frac{\sum_{i=1}^n \mathbb{I}[E(X_i^j, Y_i) \leq E(X, Y)] + 1}{n + 1} \tag{25}$$

which yields, following arguments similar to the derivation of (16), a confidence set of the form

$$\tau = Q\left(\{E(X_i^j, Y_i)\}_{i \in [n], j \in [m]}; \frac{\lfloor \alpha m(n+1) \rfloor - m + 1}{mn}\right) \tag{26}$$

and provides coverage of at least $1 - 2\alpha$. Moreover, in this setting, adapting the plausibility region approach actually also provides confidence sets: we can use the expected conformity score $e(X_i, k) = \sum_{i=1}^M E(X_i^j, k)$ both for calibration and at test time (assuming we can sample from process $p(x'|x)$ is known). However, in both cases, the obtained expected coverage guarantee also implies that coverage is now also marginal across the augmentations $(X_i^j)_{j \in \{2, ...m\}}$ for $i \in [n]$.

3.5.2 Related work

Conformal prediction (Vovk et al., 2005) has recently found numerous applications in machine learning, see e.g. (Romano et al., 2019; Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2021; Stutz et al., 2021; Fisch et al., 2022). In this paper, we particularly focus on split conformal prediction (Papadopoulos et al., 2002). However, there are also transductive and cross-validation/bagging-inspired variants being studied (Vovk et al., 2005; Vovk, 2015; Steinberger & Leeb, 2016; Barber et al., 2021; Linusson et al., 2020). Our work is related to these approaches in that many of them guarantee coverage $1 - 2\alpha$ while empirically obtaining coverage close to $1 - \alpha$. For example, cross-conformal prediction (Vovk, 2015) was recently shown to satisfy a $1 - 2\alpha$ guarantee in (Vovk et al., 2018; Kim et al., 2020). Moreover, this guarantee is also based on combining p-values without making any assumption about their dependence structure.

As outlined before, our work is also related to conformal prediction for multi-label classification which faces similar challenges as conformal prediction for ambiguous ground truth (Wang et al., 2014; 2015; Lambrou & Papadopoulos, 2016; Papadopoulos, 2014; Cauchois et al., 2021). Finally, our work has similarities to work on adversarially robust conformal prediction (Gendler et al., 2022), especially in terms of our ideal coverage guarantee.

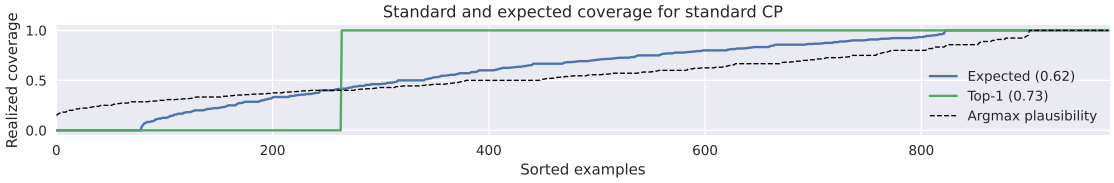

Figure 12: Realized standard $\mathbb{I}[\arg\max_k \lambda_{ik} \in C(X_i)]$ (green) and "expected" coverage $\sum_{y \in [K]} \lambda_{ik} \mathbb{I}[y \in C(X_i)]$ (blue) for standard conformal calibration using top-1 labels from plausibilities, i.e., $\arg\max_k \lambda_{ik}$. Additionally, we plot the maximum plausibility per example (dashed black) as proxy of ambiguity. We sort examples on the x-axis according to each plot individually. Clearly, many cases are very ambiguous and expected coverage is underestimated (62% vs. the target of 73%).

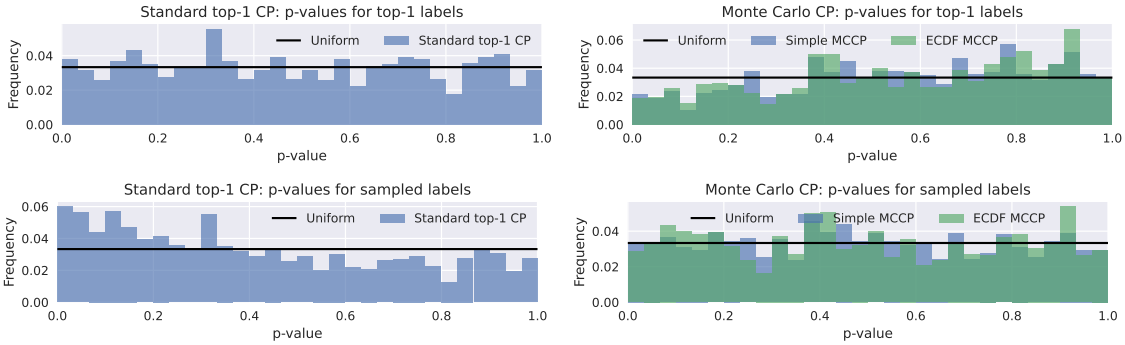

Figure 13: p-value histograms for standard (left) and Monte Carlo (right) conformal prediction. In both cases, we show the p-values for the top-1 labels (arg max of plausibilities, top) and labels samples from the plausibilities (bottom). Calibrating against top-1 labels guarantees a uniform distribution of the p-values shown on top, Monte Carlo calibration guarantees the same for the bottom histograms. As a result, the distribution of p-values w.r.t. sampled labels is skewed for standard conformal prediction while p-values w.r.t. to top-1 labels are skewed for Monte Carlo conformal prediction.

There is also a long history of work on combining dependent or independent p-values (Fisher, 1925; Rüschendorf, 1982; Meng, 1994; Heard & Rubin-Delanchy, 2017). Key work has been done in (Vovk et al., 2018), showing results without dependence assumption and thereby establishing guarantees for, e.g., cross-conformal prediction. Similar to us, (Balasubramanian et al., 2015; Linusson et al., 2017; Toccaceli & Gammerman, 2019; Toccaceli, 2019) use the ECDF to combine p-values but they do not provide rigorous coverage guarantees for this procedure.

4 Applications

4.1 Main case study: skin condition classification

In the main case study of this paper, we follow (Liu et al., 2020) and (Stutz et al., 2023) and consider a very ambiguous as well as safety-critical application in dermatology: skin condition classification from multiple images. We use the dataset of Liu et al. (2020) consisting of 1949 test examples and 419 classes with up to 6 color images resized to 448×448 pixels. The classes, i.e., conditions, were annotated by various dermatologists and aggregated deterministically (i.e., the p_{agg} model) to obtain the plausibilities λ. We followed (Roy et al., 2022; Stutz et al., 2023) to train a classifier that achieves 72.6% top-3 accuracy against the top-1 label from the plausibilities[4]. We chose a coverage level of $1 - \alpha = 73\%$ for our experiments (with results for $\alpha = 0.1$ in the appendix) to stay comparable to the base model.

[4]To be precise, in 72.6% of the cases, the top-1 label from the plausibilities is included in the top-3 prediction set derived from the predicted softmax output $\pi(x)$ *without* any conformal calibration.

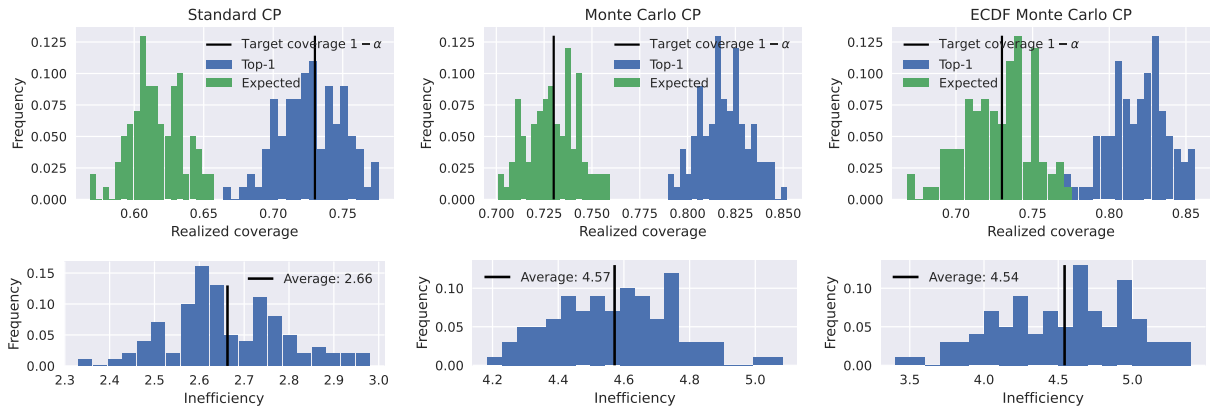

Figure 14: Comparison between standard (left), Monte Carlo (middle) and ECDF Monte Carlo (right) conformal prediction in terms of (standard top-1) coverage (blue) and expected coverage (green) across 100 random calibration/test splits. Clearly, standard conformal prediction exhibits a significant gap between expected coverage and target coverage of 73% (black). Monte Carlo conformal prediction overcomes this gap at the expense of higher inefficiency (bottom). Using ECDF-corrected p-values increases observed variation slightly due to the additional data split.

In Figure 12, we highlight how ambiguous the plausibilities for skin condition classification are: in dotted black, we plot the largest plausibility against (sorted) examples. As baseline, we performed standard conformal prediction using the classifier's softmax output as conformity scores and calibrating against the top-1 plausibility labels $\arg\max_k \lambda_{ik}$ per calibration example (X_i, λ_i). In blue, we plot the achieved standard coverage by evaluating $\mathbb{I}[\arg\max_k \lambda_{ik} \in C(X_i)]$ per example. Unsurprising, this is a step function and roughly 27% of the examples on the x-axis are covered. In green, we plot expected coverage by evaluating $\sum_{y \in [K]} \lambda_{ik} \mathbb{I}[y \in C(X_i)]$ per example. For many examples, expected coverage lies somewhere in between $[0, 1]$ showing that the obtained confidence sets only cover part of the plausibility mass. More importantly, expected coverage is, with 62% on average, significantly under-estimated by calibrating against top-1 labels. This is the core problem we intend to address in this paper.

The different between standard and Monte Carlo conformal prediction can also be observed in the obtained p-values. Figure 13 shows the p-values of standard (left) and Monte Carlo (right) conformal prediction corresponding to the top-1 labels (top) or 10 labels randomly samples from the plausibilities (per example, bottom). Standard calibration against top-1 labels results in the corresponding p-values being uniformly distributed. However, the distribution of p-values corresponding to sampled labels is slightly skewed towards 0. With Monte Carlo calibration, we observe the opposite: the p-values corresponding to top-1 labels are not entirely uniformly distributed while those corresponding to sampled labels are. This is to be expected as for many cases we cannot identify a single, unique true label and the top-1 label ignores this ambiguity.

While the above results consider a fixed calibration/test split, Figure 14 shows our overall results across 100 random splits. Left, we show how standard conformal prediction does not achieve the target 73% for the "expected" coverage $\mathbb{P}_{\text{agg}}(Y \in C(X))$ (green). However, coverage of 73% is achieved w.r.t. the top-1 labels; additionally the average inefficiency (confidence set size, bottom) is, with 2.66 smaller than the baseline model which always predicts 3 conditions to achieve the same performance. In the middle and on the right, Monte Carlo conformal prediction achieves (in expectation) expected coverage of 73% and thereby also increases coverage w.r.t. the top-1 label (blue). Of course, this comes at the cost of larger confidence sets: The inefficiency increases to 4.57. However, these larger confidence sets are required in order to cover the conditions annotated by dermatologists as shown in Figure 15. Our ECDF Monte Carlo conformal prediction generally performs similarly; the slight difference in average inefficiency stems from randomness in the additional calibration split. This also explains the slightly larger variation in coverage.

Finally, we also ran experiments using reduced plausibility regions using the reduction Ψ to top-1 labels from Equation (14). As it is extremely expensive to sample a fine enough regular grid on the simplex

Figure 15: Comparison of standard and Monte Carlo conformal prediction on two concrete examples. Both are ambiguous cases as shown by the high-entropy plausibilities. Monte Carlo conformal prediction clearly covers more plausibility mass, potentially improving patient outcome. Appendix G includes more qualitative results.

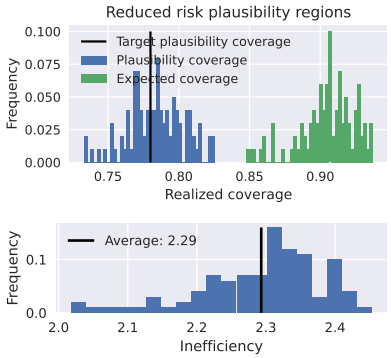

Figure 16: Reduced plausibility regions using Ψ_1 from Equation (14) on 3 risk categories, low, medium and high. These are obtained through a fixed condition to risk mapping as reliably estimating plausibility regions on 419 conditions is prohibitively expensive. Top: Clearly, the plausibility regions realize, on average, the target coverage of 78%. However, the top-1 reduced confidence sets obtain significantly higher expected coverage. Bottom: The reduced confidence sets are, however, relatively large with on average 2.29 out of 3 risk categories.

of 419 conditions, we perform experiments on 3 risk categories instead. These are obtained using a fixed mapping of conditions to low, medium and high risk. Plausibilities for this setting are obtained by summing the plausibility for each risk level and the same approach is applied on the classifier's softmax output to obtain conformity scores. Our model achieves 77.7% top-1 accuracy against the top-1 risk labels from the plausibilities; we set $\alpha = 0.22$. Figure 16 shows that the predicted plausibility regions $c(x)$ yield, on average, the target coverage, but expected coverage of the reduced confidence sets $\Psi(c(x))$ is significantly higher, similar to our observations in Figure 7.

4.2 Case study: multi-label classification

In Figure 17 we consider a simple MNIST-based multi-label classification problem with up to two, differently colored digits per image. We trained 10 multi-layer perceptrons with 100 hidden units for each digit to determine if the digit is present in the image. This simple classifier achieves 58.8% expected coverage when thresholding the 10 individual sigmoids at 0.5. As discussed in Section 2, a common strategy (Wang et al., 2015; 2014; Tsoumakas & Katakis, 2007) of performing multi-label conformal prediction consists in repeating each example according to its number of labels (here, at most 2). We can achieve something similar with expected coverage guarantee by uniformly sampling labels to perform Monte Carlo conformal prediction. As expected, this works well in practice, achieving the 90% coverage target in Figure 17 (top left). Furthermore, our discussion in Section 3.4 establishes the corresponding guarantee of $1 - 2\alpha$ without ECDF correction. However, it is important to understand what expected coverage means for multi-label classification: we allow the conformal predictor to decide how many of the labels it intends to cover per example in order to achieve the marginal coverage guarantee. This is highlighted in Figure 17 (bottom) showing the corresponding p-

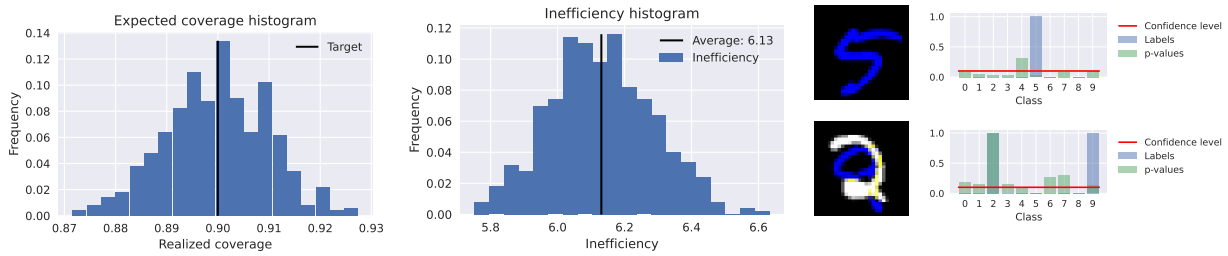

Figure 17: The multi-label conformal prediction strategy of Wang et al. (2015; 2014); Tsoumakas & Katakis (2007) is a slight variant of our Monte Carlo approach. As shown on this example of up to two overlaid, differently colored digits, Monte Carlo conformal prediction achieves target coverage of 90% (top left). However, it is free to decide how many labels to cover per examples (top middle) and, due to the poor performance of the base model (58.8% expected coverage), inefficiency is rather high. On the bottom, we show two examples of our dataset with the corresponding ground truth label set $\mathcal{Y}$ (blue) and the obtained Monte Carlo p-values (green).

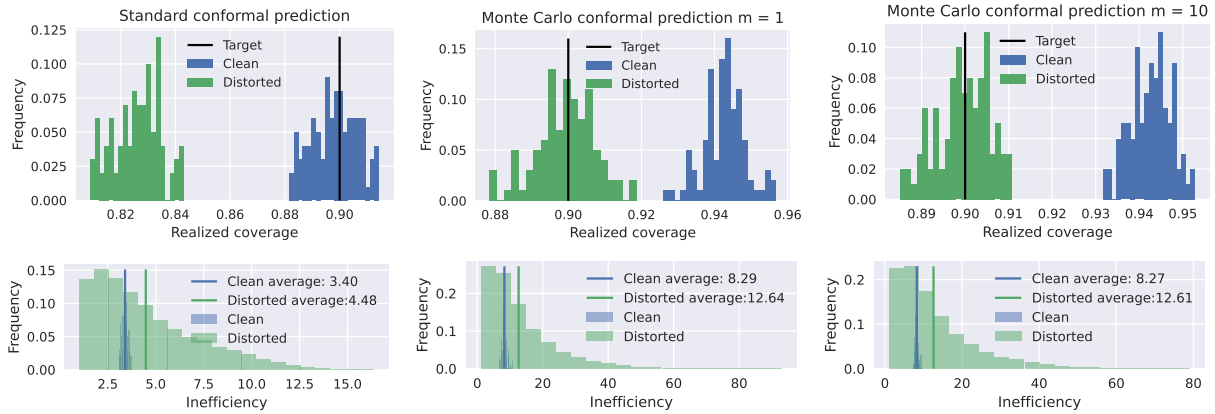

Figure 18: Standard (left) and Monte Carlo (middle and right) conformal prediction results on augmented images from ImageNet for $m = 1$ (middle) and $m = 10$ (left). We show coverage on clean (blue) and augmented images (green, average across 25 AutoAugment augmentations per image). Monte Carlo conformal prediction is able to overcome the coverage gap between augmented and clean images on average (green vs. blue) at the expense on higher inefficiency (on both clean and augmented images). Using more augmentations during calibration generally reduces the observed variation across random calibration/test splits.

values for two examples. In the first example, only one out of two examples is covered by the confidence set.

4.3 Case study: data augmentation and robustness

We also apply Monte Carlo conformal prediction in a data augmentation and robustness setting. Specifically, we took a pre-trained MobileNet V2 (Howard et al., 2017) achieving 71.3% (top-1) accuracy on the first 5k test examples of ImageNet (Russakovsky et al., 2015) and additionally evaluated it on augmented images using AutoAugment (Cubuk et al., 2018). We generated 25 random augmentations per test example and the model achieves 60.2% accuracy on average, significantly lower than on clean images. Similarly, Figure 18 shows a significant gap in coverage when calibrating only on clean images. While coverage on clean examples is around the target of 90%, depending on the random calibration/test split, coverage of perturbed images is only slightly above 80%. With Monte Carlo conformal prediction we can address this problem by calibrating on multiple augmentations per image, without loosing the coverage guarantee but at the cost of higher inefficiency. Moreover, using more augmented images during calibration significantly reduces the observed variation in coverage. We interpret these results in two ways: First, there is no reason anymore to train state-of-the-art models with data augmentation but discard augmented images during calibration. Second,

our Monte Carlo approach is effective in improving robustness of conformal predictors against augmentations or other corruptions.

5 Conclusion

In this paper, we presented a conformal prediction framework for obtaining valid confidence sets in settings with ambiguous ground truth. To take this ambiguity into account, we rely on an approximation of the underlying posterior distribution of labels given inputs, referred to as plausibilities. To overcome the requirement of "crisp", definitive labels for conformal calibration, we discussed two possible approaches based on reducing plausibility regions or Monte Carlo conformal prediction. The former essentially performs conformal regression in plausibility space and its coverage guarantee depends on how we extract confidence sets of labels from the obtained plausibility regions. The latter performs conformal prediction with labels sampled from the plausibilities. If done naively, this violates the assumption of exchangeable calibration and test examples. However, we show that Monte Carlo conformal prediction still guarantees coverage $1 - 2\alpha$ out-of-the-box which can be improved to $1 - \alpha$ using an additional calibration split. We demonstrates the applicability of these approaches in the safety-critical and particularly ambiguous setting of skin condition classification. We also highlighted how our methodology allows conformal calibration with augmented examples and provides, for the first time, a coverage guarantee for multi-label conformal prediction.

Data availability

The de-identified dermatology data used in this paper is not publicly available due to restrictions in the data-sharing agreements.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv.org*, abs/2107.07511, 2021.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

Vineeth Nallure Balasubramanian, Shayok Chakraborty, and Sethuraman Panchanathan. Conformal predictions for information fusion - A comparative study of p-value combination methods. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):45–65, 2015.

Rina Foygel Barber, J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486 – 507, 2021.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1), 2023.

Maxime Cauchois, Suyash Gupta, and John C. Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research (JMLR)*, 22:81:1–81:42, 2021.

Wenyu Chen, Kelli-Jean Chun, and Rina Foygel Barber. Discretized conformal prediction for efficient distribution-free inference. *Stat*, 7(1), 2018.

Ozan Cinar and Wolfgang Viechtbauer. The poolr package for combining independent and dependent p values. *Journal of Statistical Software*, 101(1), 2022.

Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *arXiv.org*, abs/1805.09501, 2018.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Conformal prediction sets with limited false positives. In *Proc. of the International Conference on Machine Learning (ICML)*, 2022.

R.A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2022.

Nicholas A. Heard and Patrick Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105:239–246, 2017.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv.org*, abs/1704.04861, 2017.

Byol Kim, Chen Xu, and Rina Foygel Barber. Predictive inference is free with the jackknife+-after-bootstrap. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Keith Knight and GW Bassett. Second order improvements of sample quantiles using subsamples. Technical report, University of Toronto and University of Illinois, Chicago, 2002.

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

Antonis Lambrou and Harris Papadopoulos. Binary relevance multi-label conformal predictor. In *Proc. of the Symposium on Conformal and Probabilistic Prediction and Applications (COPA)*, 2016.

Henrik Linusson, Ulf Norinder, Henrik Boström, Ulf Johansson, and Tuve Löfström. On the calibration of aggregated conformal predictors. In *Proc. of the Symposium on Conformal and Probabilistic Prediction and Applications (COPA)*, 2017.

Henrik Linusson, Ulf Johansson, and Henrik Boström. Efficient conformal predictor ensembles. *Neurocomputing*, 397:266–278, 2020.

Yuan Liu, Ayush Jain, Clara Eng, David H. Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, Vishakha Gupta, Nalini Singh, Vivek Natarajan, Rainer Hofmann-Wellenhof, Gregory S. Corrado, Lily H. Peng, Dale R. Webster, Dennis Ai, Susan Huang, Yun Liu, R. Carter Dunn, and David Coz. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26:900–908, 2020.

Xiao-Li Meng. Posterior Predictive p-Values. *The Annals of Statistics*, 22(3):1142 – 1160, 1994.

Harris Papadopoulos. A cross-conformal predictor for multi-label classification. In *Proc. of the Symposium on Conformal and Probabilistic Prediction and Applications (COPA)*, IFIP Advances in Information and Communication Technology, 2014.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2002.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

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

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

Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zachary Beaver, Nam Vo, Peggy Bui, Samantha Winter, Patricia MacWilliams, Gregory S. Corrado, Umesh Telang, Yun Liu, A. Taylan Cemgil, Alan Karthikesalingam, Balaji Lakshminarayanan, and Jim Winkens. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75:102274, 2022.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

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

Mike Schaekermann, Edith Law, Alex C Williams, and William Callaghan. Resolvable vs. irresolvable ambiguity: A new hybrid framework for dealing with uncertain ground truth. In *SIGCHI Workshop on Human-Centered Machine Learning*, volume 2016, 2016.

Lukas Steinberger and Hannes Leeb. Leave-one-out prediction intervals in linear regression models with many variables. *arXiv.org*, abs/1602.05801, 2016.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

David Stutz, Ali Taylan Cemgil, Abhijit Guha Roy, Tatiana Matejovicova, Melih Barsbey, Patricia MacWilliams, Mike Schaekermann, Jan Freyberg, Rajeev Rikhye, Beverly Freeman, Javier Perez Matos, Umesh Telang, Dale Webster, Yuan Liu, Greg Corrado, Yossi Matias, Pushmeet Kohli, Yun Liu, Arnaud Doucet, and Alan Karthikesalingam. Evaluating ai systems under uncertain ground truth: a case study in dermatology. *arXiv.org*, abs/2307.02191, 2023.

Paolo Toccaceli. Conformal predictor combination using neyman-pearson lemma. In *Proc. of the Symposium on Conformal and Probabilistic Prediction and Applications (COPA)*, 2019.

Paolo Toccaceli and Alexander Gammerman. Combination of inductive mondrian conformal predictors. *Machine Learning*, 108(3):489–510, 2019.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *Internatinal Journal on Data Warehousing and Mining*, 3(3):1–13, 2007.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):9–28, 2015.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 2020.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005.

Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. Cross-conformal predictive distributions. In *Proc. of the Symposium on Conformal and Probabilistic Prediction and Applications (COPA)*, 2018.

Hua-zhen Wang, Xin Liu, Ilia Nouretdinov, and Zhiyuan Luo. A comparison of three implementations of multi-label conformal prediction. In *Proc. of the International Symposium on Statistical Learning and Data Sciences*, 2015.

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.

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

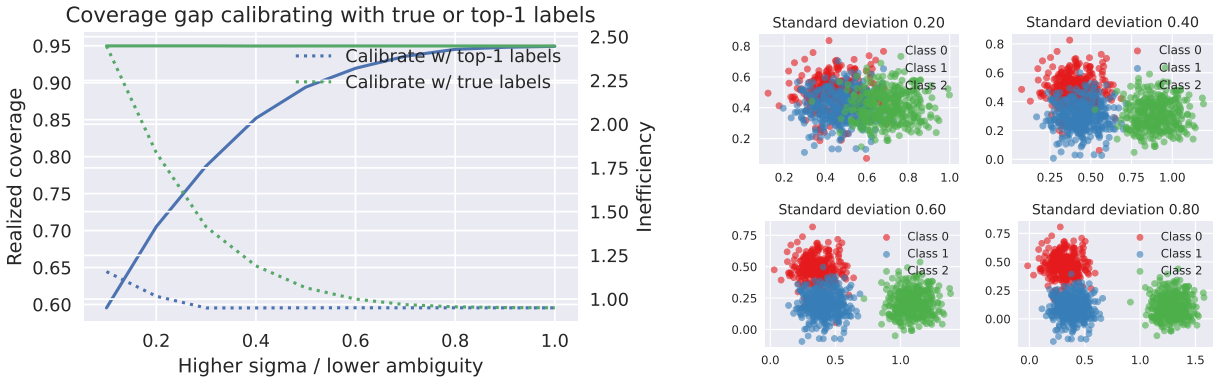

Figure 19: Left: Coverage against true labels (left y-axis), cf. Figure 1, when performing conformal calibration with true or top-1 labels for different levels of ambiguity as measured using the standard deviation of the mean class distributions (x-axis). We also plot inefficiency (right y-axis). We see how the gap in true coverage between both approaches reduces the smaller the ambiguity (i.e., higher standard variation on x-axis). To compensate for the ambiguity, calibrating with true labels will generally also result in higher inefficiency. Right: Examples from the toy dataset for different levels of ambiguity.

A True and top-1 coverage gap by ambiguity

Figure 19 highlights how the gap between calibrating with true or top-1 labels in terms of true coverage depends on the ambiguity of the problem. Specifically, the means μ_k of our toy example in Equation (2) are generated by sampling them from a Gaussian with lower or larger standard deviation. Clearly, overlap between the class distributions increases the closer the means μ_k are together and we expect a larger gap. At the same time, if ambiguity is high, calibrating against the true labels will result in high inefficiency (size of $C(x)$). Pushing the class distributions outwards then reduces the gap alongside inefficiency. If the classes are perfectly separable, true and top-1 coverage coincide. This corresponds to a case where the true posteriors $\mathbb{P}(Y = y|X = x)$ are crisp.

B Calibration threshold and p-values

B.1 Single p-value

We include for completeness a proof of the results presented in Section 2.2. These are standard results. The p-value formulation detailed here will be then extended to establish the validity of some of the procedures proposed in this work.

We first establish that the confidence set defined by Equations (3) and (4) satisfies

$$1 - \alpha \leq \mathbb{P}(Y \in C(X)) \leq 1 - \alpha + \frac{1}{n+1}, \tag{27}$$

the upper bound requiring the additional assumption that the conformity scores are almost surely distinct.

Let us write $(X_{n+1}, Y_{n+1}) = (X, Y)$ and $S_i = E(X_i, Y_i)$. As $(S_i)_{i\in[n+1]}$ are exchangeable, then a direct application of (Romano et al., 2019, Lemma 1) shows that for any $\alpha \in (0, 1)$ and $\tau_1 = Q\left(\{-S_i\}_{i\in[n]}; (1-\alpha)(1+1/n)\right)$

$$\mathbb{P}(-S_{n+1} \leq \tau_1) \geq 1 - \alpha \tag{28}$$

and, if the random variables $(S_i)_{i\in[n+1]}$ are almost surely distinct, then

$$\mathbb{P}(-S_{n+1} \leq \tau_1) \leq 1 - \alpha + \frac{1}{n}. \tag{29}$$

Note that we have additionally

$$\mathbb{P}(-S_{n+1} \leq \tau_1) = \mathbb{P}(-S_{n+1} \leq \tau_2) \tag{30}$$

for $\tau_2 = Q\left(\{-S_i\}_{i\in[n]}; \lceil (1-\alpha)(n+1)\rceil / n\right)$. So $C(X_{n+1})$ includes all the values y such that $-S(y, X_{n+1})$ is smaller or equal than the $\lceil (1-\alpha)(n+1)\rceil$ smallest values of $(-S_i)_{i\in[n]}$. This is equivalent to considering all the values y such that $S(y, X_{n+1})$ is larger or equal than the $\lfloor \alpha(n+1)\rfloor$ smallest values of $(S_i)_{i\in[n]}$. Hence this corresponds to the confidence set defined by Equations (3) and (4).

We now show that this confidence set can also be obtained by thresholding p-values. Indeed Equation (5) can be rewritten as

$$\rho_{Y_{n+1}} = \frac{\sum_{i=1}^{n+1} \mathbb{I}(S_i \leq S_{n+1})}{n+1}. \tag{31}$$

As $(S_i)_{i\in[n+1]}$ are exchangeable, then $\rho_{Y_{n+1}}$ is uniformly distributed on $\{\frac{1}{n+1}, \frac{2}{n+1}, ..., 1\}$ if the distribution of the scores is continuous and thus $\rho_{Y_{n+1}}$ is a p-value, i.e. $\mathbb{P}(\rho_{Y_{n+1}} \leq \alpha) \leq \alpha \iff \mathbb{P}(\rho_{Y_{n+1}} > \alpha) \geq 1-\alpha$. It can be checked that this property still holds if the distribution of the scores is not continuous (see e.g. (Bates et al., 2023)). So if we define the confidence set as

$$C(X_{n+1}) = \{y : \rho_y > \alpha\} \tag{32}$$

then by construction it follows that

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1-\alpha. \tag{33}$$

We show here that Equation (32) does indeed coincide with the set defined by Equations (3) and (4). We have

$$\rho_y > \alpha \iff \sum_{i=1}^{n} \mathbb{I}(S_i \leq y) > \alpha(n+1) - 1. \tag{34}$$

If $\alpha(n+1)$ is an integer then $\sum_{i=1}^{n} \mathbb{I}(S_i \leq y) \geq \alpha(n+1)$, i.e. y is larger or equal than the $\alpha(n+1)$ smallest values of $(S_i)_{i\in[n]}$. If $\alpha(n+1)$ is not an integer then it means that $\sum_{i=1}^{n} \mathbb{I}(S_i \leq y) \geq \lceil \alpha(n+1)\rceil - 1$, i.e. y is larger than the $\lceil \alpha(n+1)\rceil - 1$ smallest values of $(S_i)_{i\in[n]}$. However, we have $\lceil \alpha(n+1)\rceil - 1 = \lfloor \alpha(n+1)\rfloor$ as $\alpha(n+1)$ is not an integer. So overall, we have $\rho_y > \alpha$ corresponds to y being larger or equal than the $\lfloor \alpha(n+1)\rfloor$ smallest values $(S_i)_{i\in[n]}$. So the confidence set Equation (32) does coincide with the set defined by Equations (3) and (4).

Finally note that when the distributions of the scores is continuous, i.e. the scores are almost surely distinct, then we also have $\mathbb{P}(\rho_{Y_{n+1}} \leq \alpha) \geq \alpha - 1/n+1$ so $\mathbb{P}(\rho_{Y_{n+1}} > \alpha) \leq 1 - \alpha + 1/n+1$.

B.2 Average of p-values

We establish here the expression of the confidence set obtained by thresholding the following average p-value

$$\rho_{Y_{n+1}} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n+1} \mathbb{I}(S_i^j \leq S_{n+1}^j)}{m(n+1)} \tag{35}$$

where $S_{n+1}^j = S_{n+1}$ for all $j \in [m]$. We know that by construction the set $C(X_{n+1}) = \{y : \rho_y > \alpha\}$ is such that

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - 2\alpha. \tag{36}$$

We have

$$\rho_y > \alpha \iff \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{I}(S_i^j \leq y) > \alpha m(n+1) - m. \tag{37}$$

If $\alpha m(n+1)$ is an integer then y needs to be larger or equal than the $\alpha m(n+1) - m + 1$ smallest values of $(S_i^j)_{i\in[n],j\in[m]}$. If $\alpha m(n+1)$ is not an integer then $\sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{I}(S_i^j \leq y) \geq \lceil \alpha m(n+1)\rceil - m$, i.e. y is larger

Figure 20: We show two examples from CIFAR10-H with their corresponding label frequencies. These frequencies could be used as plausibility estimates or within Bayesian inference, see text for details.

or equal than the $\lceil \alpha m(n+1) \rceil - m$ smallest values of S_i. However, we have $\lceil \alpha m(n+1) \rceil - 1 = \lfloor \alpha m(n+1) \rfloor$ as $\alpha m(n+1)$ is not an integer. So $\lceil \alpha m(n+1) \rceil - m = \lfloor \alpha m(n+1) \rfloor - m + 1$. So overall we need y larger or equal than the $\lfloor \alpha m(n+1) \rfloor - m + 1$ smallest values of $(S_i^j)_{i \in [n], j \in [m]}$, equivalently larger that the quantile of $(S_i^j)_{i \in [n], j \in [m]}$ at level $\frac{\lfloor \alpha m(n+1) \rfloor - m + 1}{mn}$, i.e. $C(X_{n+1}) = \{k \in [K] : E(X, k) \geq \tau\}$ for τ defined in Equation (19).

C Examples for aggregation model $\mathbb{P}_{\text{agg}}(Y = y|X = x)$

We consider our toy example from Section 2.1 and assume annotators provide samples from $\mathbb{P}_{\text{agg}}(Y = y|\lambda) = \lambda_y$ which is supposed to approximate the true $\mathbb{P}(Y = y|X = x)$. We can then infer the plausibilities λ using frequencies from annotations $B_i = b_i$ for example i. This would essentially provide a maximum likelihood estimate, approximating $p(\lambda|b)$ by a delta-mass at this estimate. Instead, we can also put a prior on the plausibilities λ, for example a Dirichlet distribution. Then, the posterior $p(\lambda|b)$ is also a Dirichlet distribution specified by the frequencies. This gives us a closed-form solution for $p(\lambda|b)$ which then defines the model as in Equation (8):

$$\mathbb{P}_{\text{agg}}(Y = y|X = x) = \int \mathbb{P}(Y = y|\lambda) p(\lambda|x) \mathrm{d}\lambda = \int \int \lambda_y p(\lambda|b) p(b|x) \mathrm{d}\lambda \mathrm{d}b \tag{38}$$

This setting is directly applicable to many popular datasets such as CIFAR10 (Krizhevsky, 2009) which is a popular multiclass classification benchmark of 32×32 pixel color images. Here, recent work (Peterson et al., 2019) gathered additional human annotations for the test set where each annotator provides a single label. Frequencies for two very ambiguous examples are shown in Figure 20.

D More examples for reducing plausibility regions

In the main paper, we introduced the top-1 reduction Ψ in Equation (14). However, especially for more ambiguous cases where $\mathbb{P}_{\text{agg}}(Y = y|X = x)$ is not one-hot, we might be interested in a generalization of Ψ_1 to the top-k labels:

$$\Psi_2(c(x)) = \{y \in [K] : (\exists \lambda \in c(x) : y \in \text{top\_3}(\lambda))\} \tag{39}$$

includes all top-3 labels for the plausibilities included in $c(x)$. It then follows directly from the coverage guarantee on the plausibility region $c(X)$ that

$$\mathbb{P}(\text{top\_3}(\lambda) \subseteq C(X)) \geq 1 - \alpha. \tag{40}$$

This does not necessarily have to imply a tight coverage guarantee and also includes the arg max, i.e., top-1, case. We also emphasize that this cannot be translated to a guarantee on (8). Even though the top-3 classes are included, we do not know how much probability mass these cover $\mathbb{P}_{\text{agg}}(Y = y|X = x)$.

To generalize Ψ_2 and avoid fixing the number of classes included in the guarantee, e.g., the top-3, we could use

$$\Psi_3(c(x)) = \left\{ y \in [K] : \left(\exists \lambda \in c(x), l \in [K] \text{ s.t. } \sum_{i=1}^{l} \lambda_{\sigma_i} \geq 1 - \delta \text{ and } \exists i \leq l \text{ s.t. } \sigma_i = y \right) \right\}. \tag{41}$$

Here, we sort the plausibilities λ included in the plausibility region $c(x)$ such that $\lambda_{\sigma_1} \geq \ldots \geq \lambda_{\sigma_K}$ and pick an l such that the l-largest plausibilities, i.e., $\sum_{i=1}^{l} \lambda_{\sigma_i}$, exceed $1-\delta$. This means shifting from fixing, e.g., a top-3 set, to always considering the top classes constituting at least $1-\delta$ of the plausibility mass. By construction of Ψ_3, we have

$$\lambda \in c(x) \Rightarrow \mathbb{E}_{Y \sim \mathbb{P}(\cdot|\lambda)}[\mathbb{I}[Y \in C(x)]] = \sum_{k=1}^{K} \lambda_k \mathbb{I}[k \in C(x)] \geq 1-\delta. \tag{42}$$

In the ideal scenario where $\lambda_y = \mathbb{P}(Y=y|X=x)$ then $\mathbb{E}_{Y \sim \mathbb{P}(\cdot|\lambda)}[\mathbb{I}[Y \in C(x)]] = \mathbb{E}_{Y \sim \mathbb{P}(\cdot|X=x)}[\mathbb{I}[Y \in C(x)]]$ and, as $\mathbb{P}(\lambda \in c(X)) \geq 1-\alpha$, this implies the following coverage guarantee:

$$\mathbb{P}(Y \in c(X)) = \mathbb{E}_{X \sim \mathbb{P}^X}[\mathbb{E}_{Y \sim \mathbb{P}^{Y|X}(\cdot|x)}[\mathbb{I}[Y \in C(x)]]] \geq (1-\alpha)(1-\delta). \tag{43}$$

When choosing δ very small, in order to roughly obtain expected coverage $(1-\alpha)$, however, Ψ_3 will likely include many classes in the confidence sets.

While the above examples provide guarantees in the context of classification problems, we can also design Ψ to provide more structured guarantees. For example, considering the toy dataset from Section 2.1, we can provide coverage guarantees not only on the top-k labels but also their ranking. To this end, we construct

$$\Psi_4(c(x)) = \{\mu = (\mu_1, \ldots, \mu_k) : (\exists \lambda \in c(x) \text{ s.t. } (\sigma_1, \ldots, \sigma_L) = \mu)\}. \tag{44}$$

Again, we consider the sorted plausibilities $\lambda_{\sigma_1} \geq \ldots \geq \lambda_{\sigma_K}$ and let μ denote one of the (partial) rankings to be included in the prediction set. Informally, we include all such rankings μ that can be derived from the plausibilities in $c(x)$. For simplicity, we assume rankings of fixed length $1 \leq k \leq K$. By construction, we can guarantee that the top-k ranking of the true plausibilities is included in $\Psi_4(c(x))$ with probability at least $(1-\alpha)$. This is essentially a variant of Ψ_2 that also considers the ordering of labels.

E Relationship between Monte Carlo conformal prediction and expected conformity scores

It is also important to realize that Monte Carlo conformal prediction is not equivalent to using the expected conformity score e of Equation (11). This is because

$$F(f) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}[E(X_i, Y_i^j) \leq f] \neq \frac{1}{n} \sum_{i} \mathbb{I}[e(X_i, \lambda_i) \leq f]. \tag{45}$$

Consider a simple case with four examples with corresponding conformity scores $E_1 = 0.1$, $E_2 = 0.2$ and $E_3 = 0.3$. The fourth example has $E_4 = 0$ or $E_4 = 0.5$ with equal probability as this is the only example with non-one-hot plausibilities λ. Then, $e_4 = 0.25$. If we now evaluate the cumulative distribution for $E = 0.2$, the left side of Equation (45) is $5/8$ in expectation, *i.e.*, performing m trials and always having an even number of $E_4 = 0$ and $E_4 = 0.5$. The right side, in contrast, is always $1/2$. This suggests that the computed quantiles using Monte Carlo conformal prediction are smaller than those using the expected conformity score, indicating why using a threshold calibrated on top of expected conformity scores cannot be used to directly predict confidence sets C.

F Interpretation of Monte Carlo conformal prediction

While our approach is clearly a sampling-based approach, we want to highlight more explicitly why we refer to it as Monte Carlo conformal prediction. Specifically, consider that standard conformal prediction estimates the $\alpha(1+1/n)$-quantile of the empirical distribution

$$F(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[E(X_i, Y_i) \leq f]. \tag{46}$$

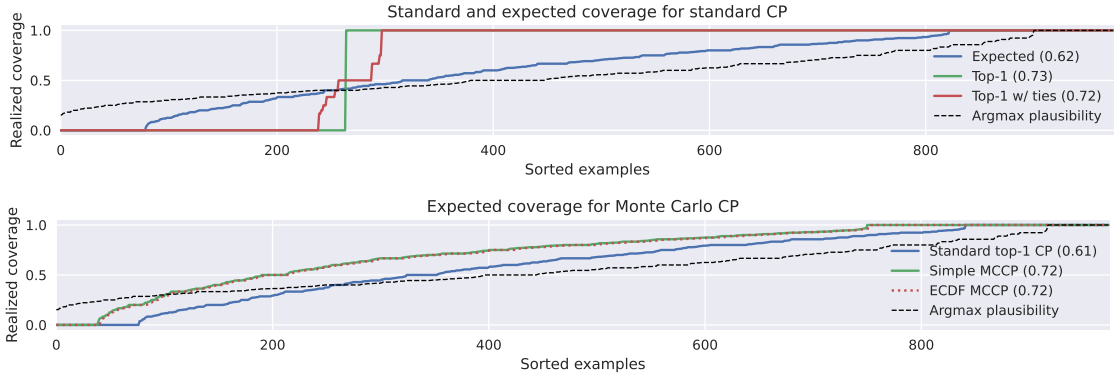

Figure 21: Top: Complementary to Figure 12, we include evaluation against top-1 labels while considering ties (red, $1/L_i \sum_{j=1}^{L_i} \mathbb{I}[Y_i^j \in C(X_i)]$ with Y_i^j being one of L_i tied labels for case i) which hints towards several ambiguous cases. Looking at expected coverage (i.e., $\sum_{y \in [K]} \lambda_{ik} \mathbb{I}[y \in C(X_i)]$) in blue, however, shows that a significant portion of examples are ambiguous and standard evaluation against the top-1 plausibility labels (i.e., $\mathbb{I}[\arg\max_k \lambda_{ik} \in C(X_i)]$)is unreasonable. Bottom: Expected coverage plot for our approach, comparing standard, Monte Carlo and ECDF Monte Carlo conformal prediction.

As we do not have access to the required Y_i, we approximate this as

$$F(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}[E(X_i, Y_i^j) \leq f]. \tag{47}$$

Clearly, the inner sum corresponds to a Monte Carlo estimate of $\mathbb{E}_{Y \sim \mathbb{P}_{\text{agg}}(\cdot | X = x)}[\mathbb{I}[E(X, Y) \leq f]]$ for $m \to \infty$. Note that performing calibration for very large m is usually not a problem in practice, i.e., Monte Carlo conformal prediction scales very well. Alternatively, we can also view Monte Carlo conformal prediction as conducting m separate calibrations. Specifically, letting τ_j be the threshold computed as $Q(\{E(X_i, Y_i^j\}_{i=1}^n; \alpha(1 + 1/n))$ for arbitrary but fixed j, we obtain m different thresholds. These can be understood as estimates of the true threshold obtained when calibrating against the true but unknown labels Y_i[5]. We also note that Monte Carlo conformal prediction is clearly not equivalent to conformal prediction on the expected conformity score as described before, see Appendix E.

G Additional results for skin condition classification

G.1 Skin condition classification

Figure 21 presents complementary results to Figure 12 in the main paper. Specifically, on top, we additionally consider ties among the top-1 labels (i.e., there is no unique arg max in the plausibilities λ). In the calibration examples, we break these ties randomly. At test time, however, we can decide to evaluate coverage proportional to the tied labels. Formally, we assume that $Y_i^1, \ldots, Y_i^{L_i}$ are the tied labels and compute $1/L_i \sum_{j=1}^{L_i} \mathbb{I}[Y_i^j \in C(X_i)]$ instead of the binary indicator $\mathbb{I}[Y_i \in C(X_i)]$ to evaluate coverage. In red, we see that quite a few examples exhibit ties and the predicted confidence sets often cover only a part of the tied labels. This is an early indicator for high ambiguity in the ground truth of this dataset. On the bottom, we additionally plot expected coverage for (ECDF) Monte Carlo conformal prediction in comparison to standard calibration against top-1 labels. The gap between Monte Carlo and standard conformal prediction is clearly visible and once again demonstrates how significantly standard conformal prediction under-estimates expected coverage on this dataset. While both Monte Carlo approaches (with and without ECDF correction) look identical in this example, they do not have to be due to randomness in how the calibration set is further split (cf. Algorithm 3). In expectation across splits, however, we found that they

[5]This is because quantiles can be estimated by averaging quantiles of random subsamples (Knight & Bassett, 2002, Theorem 1).

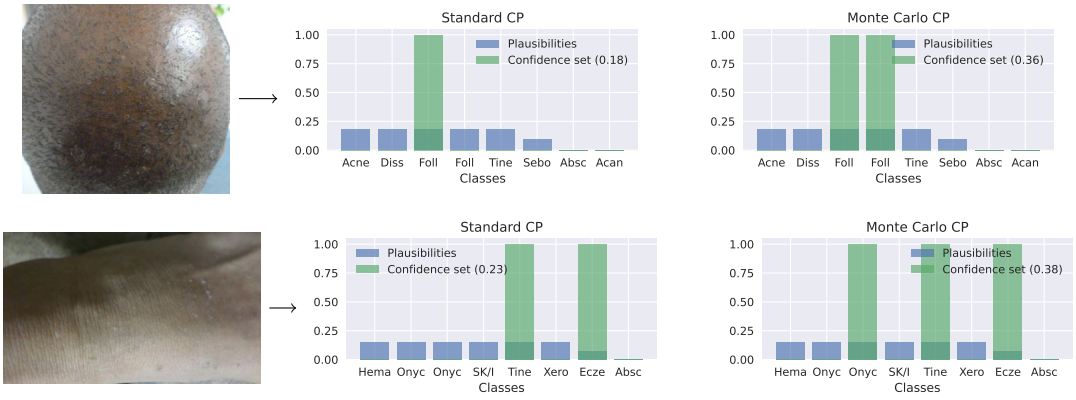

Figure 22: Additional qualitative results corresponding to Figure 15.

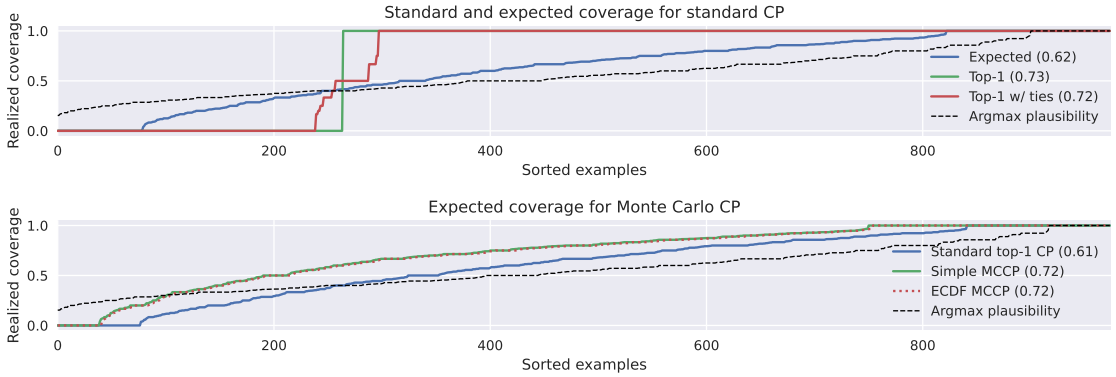

Figure 23: Results corresponding to Figure 12 with $\alpha = 0.1$.

coincide. This also highlights that the simple variant empirically achieves expected coverage $1 - \alpha$ despite only guaranteeing $1 - 2\alpha$.

Figure 22 shows two additional qualitative examples where Monte Carlo conformal prediction improves results, i.e., more conditions with significant plausibility (blue) are covered (green) but important conditions are still not covered. This can be addressed using a lower confidence level such as $\alpha = 0.1$ in Figure 25.

Results of our main experiments in dermatology for $\alpha = 0.1$ can be found in Figures 23 to 25.

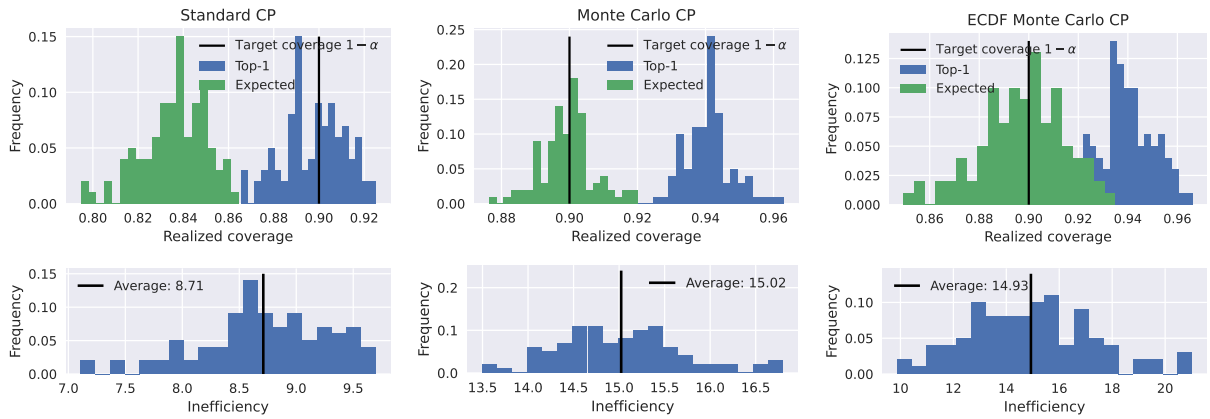

Figure 24: Results corresponding to Figure 14 with $\alpha = 0.1$.

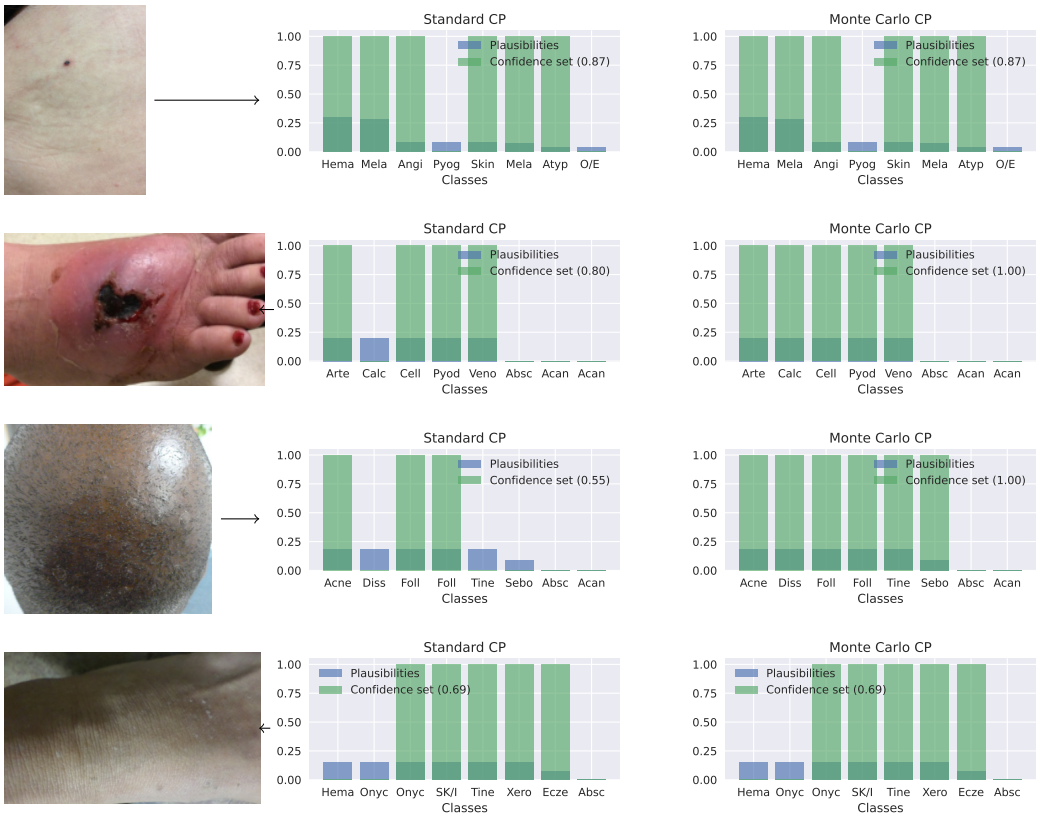

Figure 25: Additional qualitative results for $\alpha = 0.1$ corresponding to Figures 15 and 22.