Proceedings of Machine Learning Research 204:1–16, 2023 Conformal and Probabilistic Prediction with Applications

Conformal Prediction with Partially Labeled Data

Alireza Javanmardi ALIREZA.JAVANMARDI@IFI.LMU.DE
Yusuf Sale YUSUF.SALE@IFI.LMU.DE
Paul Hofman PAUL.HOFMAN@IFI.LMU.DE
Eyke Hüllermeier EYKE@IFI.LMU.DE
Institute of Informatics, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Germany

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

While the predictions produced by conformal prediction are set-valued, the data used for training and calibration is supposed to be precise. In the setting of superset learning or learning from partial labels, a variant of weakly supervised learning, it is exactly the other way around: training data is possibly imprecise (set-valued), but the model induced from this data yields precise predictions. In this paper, we combine the two settings by making conformal prediction amenable to set-valued training data. We propose a generalization of the conformal prediction procedure that can be applied to set-valued training and calibration data. We prove the validity of the proposed method and present experimental studies in which it compares favorably to natural baselines.

Keywords: conformal prediction, superset learning, partial label learning, imprecise data

1. Introduction

Conformal prediction (CP), a prominent uncertainty quantification technique, has drawn increasing attention in statistics and machine learning over the past decade. With its roots in classical frequentist statistics, this framework enables the construction of reliable prediction sets without the need for any distributional assumptions (Vovk et al., 2005). A key advantage of conformal prediction lies in its validity guarantees for the constructed prediction sets, which cover the true outcomes with high probability. This makes it appealing for applications in safety-critical domains, such as risk assessment in finance (Gammerman and Vovk, 2007), medical diagnosis and disease prediction (Papadopoulos et al., 2009), drug discovery and toxicity prediction (Svensson et al., 2018), among many others.

Another machine learning (ML) setting dealing with set-valued data is partial label learning (PLL), a specific type of weakly supervised learning (Grandvalet, 2002; Jin and Ghahramani, 2002; Nguyen and Caruana, 2008; Cour et al., 2011). In a sense, PLL is orthogonal to conformal prediction: While the predictions produced by CP are set-valued, the data used for training and calibration is supposed to be precise. In PLL, it is exactly the other way around: Although the training data might be imprecise (set-valued), the goal is to induce a unique model producing precise (point) predictions. This may strike as odd, as one may argue that if the training data is imprecise or ambiguous, it might be all the more important to reflect this imprecision or ambiguity in the induced model and the predictions produced by this model. For example, one may allow for a set of incomparable, undominated models, resulting, for instance, from the interval order induced by set-valued loss functions (Couso and Sánchez, 2016), or by the application of conservative, imprecise Bayesian updating rules (Zaffalon and Miranda, 2009).

As an alternative, we suggest the use of CP to capture (predictive) uncertainty in the setting of PLL. In other words, we propose to combine PLL and CP within a single framework. To this end, we propose a generalization of the CP procedure that can be applied to set-valued training and calibration data. For this approach, we establish theoretical validity guarantees. Moreover, through experimental studies, we showcase the enhanced accuracy of our method in weakly supervised learning settings compared to natural baselines.

2. Background

2.1. Partial Label Learning

As already said, partial label learning (PLL) is a specific type of learning from weak supervision, in which the outcome (response) associated with a training instance is only characterized in terms of a subset of possible candidates. Thus, PLL is somehow in-between supervised and semi-supervised learning, with the latter being a special case. Motivated by practical applications in which only partial information about outcomes is available, PLL has been studied under various names, such as *learning from ambiguously labeled examples* (Hüllermeier and Beringer, 2006), and under slightly different assumptions on the incomplete information being provided. Often, the only assumption made is that the set of candidates covers the actual (precise) outcome, which is also reflected by the name *superset learning* (Liu and Dietterich, 2012).

More formally, consider the standard setting of supervised learning with a data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the instance space and $\mathcal{Y}$ is the output space, respectively. As our focus is on the multi-class classification scenario, we refer to the output variable as "label" and assume $\mathcal{Y}$ to be finite (typically of small to moderate size). The learning task normally consists of choosing an optimal model (hypothesis) h^* from a given model space (hypothesis space) $\mathcal{H}$, based on a set of training data

$$\mathcal{D} = \big\{ \, (x_i, y_i) \, \big\}_{i=1}^{n} \in (\mathcal{X} \times \mathcal{Y})^n \ . \tag{1}$$

More specifically, optimality typically refers to prediction accuracy, i.e., a model is sought whose expected prediction loss or *risk*

$$\mathcal{R}(h) \, = \mathbb{E}_{(x,y)\sim P} \, L\big(y, h(x)\big) \, = \int L\big(y, h(x)\big) \, d\, P(x,y) \tag{2}$$

is minimal; here, $L : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$ is a loss function, and P is an (unknown) probability measure on $\mathcal{X} \times \mathcal{Y}$ modeling the underlying data generating process.

In PLL, the learning algorithm does not have direct access to the data (1) because the labels $y_i \in \mathcal{Y}$ are not observed precisely. Instead, only supersets $S_i \subseteq \mathcal{Y}$ are observed so that the training data consists of (imprecise, coarse, ambiguous) observations

$$\mathcal{O} = \big\{ (x_i, S_i) \big\}_{i=1}^{n} \in (\mathcal{X} \times 2^{\mathcal{Y}})^n \ . \tag{3}$$

There are various ways of learning from data of that kind, notably the idea of generalizing the principle of empirical risk minimization through the use of a generalized loss function.

For example, Hüllermeier (2014) introduces the *optimistic superset loss* as an extension of the loss L in (2):

$$L_O(S, \hat{y}) = \min \big\{ L(y, \hat{y}) \,|\, y \in S \big\} \ . \tag{4}$$

Learning is then accomplished by finding a model minimizing this loss (or maybe a regularized version thereof) on the training data:

$$h^* \in \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L_O\big(S_i, h(x_i)\big) \ . \tag{5}$$

A key motivation of this approach is the idea of *data disambiguation*, i.e., the idea of simultaneously inducing the true model and reconstructing the values of the underlying precise data. The same type of loss function has more recently been introduced under the notion of *infimum loss* (Cabannnes et al., 2020).

Obviously, depending on the underlying loss function L, the optimization problem (5) may become complex, especially since (4) could be non-convex. From a theoretical perspective, an important question concerns conditions under which successful learning (for example, in the sense of convergence toward a truly optimal model) is actually possible, despite the imprecision of the data. An analysis of this kind obviously requires assumptions about the process of "imprecisiation", i.e., the way in which precise outcomes are turned into imprecise observations. The first positive results, showing that successful learning is possible under specific assumptions, have been obtained by Liu and Dietterich (2014); Cabannnes et al. (2020, 2021).

2.2. Conformal Prediction

Suppose a training dataset (1) to be given, and denote by $(x_{new}, y_{new}) \in \mathcal{Z}$ a new test point. Assuming that x_{new} is observed, but y_{new} is not, CP aims to construct a *prediction set* of the form $\mathcal{T}(x_{new}) \subseteq \mathcal{Y}$ that is valid in the sense that $y_{new} \in \mathcal{T}(x_{new})$ with high probability. Informally speaking, the idea of CP is to test the hypothesis $y_{new} = y$ for all $y \in \mathcal{Y}$ and to exclude from the prediction set only those outcomes y for which this hypothesis can be rejected at the predefined level of confidence. Hypothesis testing is done in a nonparametric way: Consider any *nonconformity function* that assigns scores $\alpha(x, y)$ to input/output tuples; the latter can be interpreted as a measure of "strangeness" of the pair (x, y), i.e., the higher the score, the less the data point (x, y) conforms to what one would expect to observe. Assuming *exchangeability* of the data $\mathcal{D}$, CP then finds a critical value q for the degree of nonconformity so that those y with $\alpha(x, y) > q$ are excluded. Theoretically, the CP procedure is able to guarantee *marginal coverage*, meaning that, in an infinite sequence of predictions, the miscoverage rate (fraction of predictions $\mathcal{T}(x_{new})$ not covering y_{new}) does not exceed a prespecified value $\epsilon > 0$.

This guarantee holds true regardless of how the nonconformity function is defined. Yet, this function has a strong influence on the *efficiency* of predictions, i.e., the (average) size of the prediction sets. A common approach, which we will also assume in the following, is to train a probabilistic predictor $\hat{f}$ so that $\hat{f}(x)$ is a prediction of the conditional probability $p(\cdot \,|\, x)$ on $\mathcal{Y}$. Nonconformity scores are then naturally defined in terms of reciprocals of class probabilities, i.e., $\alpha(x, y) = 1 - \hat{f}(x)_y$.

CP has originally been developed in a transductive setting, which, however, comes with major computational challenges (e.g., one would need to retrain the predictor $\hat{f}$ after each new observation). Later on, inductive variants of CP have also been developed (Papadopoulos et al., 2002a,b). To construct prediction sets using inductive conformal prediction (ICP), the first step is to partition the training data $\mathcal{D}$ into two subsets, the *proper training set* $\mathcal{D}_{\text{train}}$ and the *calibration set* $\mathcal{D}_{\text{calib}}$:

$$\mathcal{D}_{\text{train}} = \{(x_i, y_i) : i \in \mathcal{I}_1\}$$
$$\mathcal{D}_{\text{calib}} = \{(x_i, y_i) : i \in \mathcal{I}_2\}$$

Then, a multi-class classification algorithm $\mathcal{A}$ is used to fit a (probabilistic) predictor to the proper training set:

$$\hat{f}(\cdot) \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}}) \tag{6}$$

The next step is called calibration, which involves computing the nonconformity score of each calibration data instance that determines how well it conforms to the established classifier $\hat{f}$. As already said, a natural choice for the nonconformity is one minus the predicted probability of the ground-truth class, giving rise to a score set

$$\mathcal{E} := \left\{ 1 - \hat{f}(x_j)_{y_j} : j \in \mathcal{I}_2 \right\} . \tag{7}$$

For any set of nonconformity scores $\mathcal{E}$, define the critical score $q(\mathcal{E}, \epsilon)$ in terms of its $\lceil (1 + |\mathcal{E}|)(1-\epsilon) \rceil$ smallest value, or equivalently, its $|\mathcal{E}|^{-1} \lceil (1 + |\mathcal{E}|)(1-\epsilon) \rceil$ empirical quantile. Furthermore, given an instance $x \in \mathcal{X}$, a classifier $\hat{f} : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$, a set of nonconformity scores $\mathcal{E}$, and an error rate ϵ, define the prediction set $\mathcal{T}(x, \hat{f}, \mathcal{E}, \epsilon)$ as

$$\mathcal{T}(x, \hat{f}, \mathcal{E}, \epsilon) := \left\{ y \in \mathcal{Y} : \hat{f}(x)_y \geq 1 - q(\mathcal{E}, \epsilon) \right\} . \tag{8}$$

ICP outputs $\mathcal{T}(x_{new}, \hat{f}, \mathcal{E}, \epsilon)$ as the prediction set $\hat{Y}_{new}$ for a new test point x_{new}, thereby satisfying the marginal coverage property (in expectation) if samples in $\mathcal{D}_{\text{calib}} \cup \{(x_{new}, y_{new})\}$ are drawn exchangeably from a joint probability distribution over the data space.

3. Conformal Prediction with Partially Labeled Data

Coming back to the idea of combining conformal prediction with partial label learning, we are now again interested in the case of set-valued training data (3), where each data instance x_i is associated with a set of potential labels $S_i \subseteq \mathcal{Y}$. We keep denoting the ground-truth label of instance x_i by y_i and assume it lies in its candidate set, i.e., $y_i \in S_i$.

Applying ICP to such data, we again start by partitioning the data $\mathcal{O}$ into proper training $\mathcal{O}_{\text{train}} = \{(x_i, S_i) : i \in \mathcal{I}_1\}$ and calibration subsets $\mathcal{O}_{\text{calib}} = \{(x_i, S_i) : i \in \mathcal{I}_2\}$. As mentioned in Section 2.1, the task of learning from partially labeled data has been well-studied in the literature, making the training step of ICP feasible for such data. All we need to do next is replace the algorithm $\mathcal{A}$ in (6) with a partial label learning algorithm $\mathcal{A}_{\text{PLL}}$ that allows us to fit a classifier on $\mathcal{O}_{\text{train}}$:

$$\hat{f}_{\text{PLL}}(\cdot) \leftarrow \mathcal{A}_{\text{PLL}}(\mathcal{O}_{\text{train}}) \tag{9}$$

Like before, we assume that the induced classifier $\hat{f}_{\text{PLL}}(\cdot)$ predicts probability distributions over the classes. However, each calibration instance x_j is now associated with (possibly) multiple labels S_j. This begs the general question of how to compute the nonconformity scores for set-valued data (x_i, S_i).

A relative straightforward approach is to consider each candidate label $y \in S_j$ separately, compute its nonconformity score $1 - \hat{f}_{\text{PLL}}(x_j)_y$, and then pessimistically pick the maximum one:

$$\mathcal{E}_{\max} := \left\{ 1 - \min_{y \in S_j} \hat{f}_{\text{PLL}}(x_j)_y : j \in \mathcal{I}_2 \right\} . \tag{10}$$

For a new test point x_{new} and any $\epsilon \in (0, 1]$, the prediction set is then given by

$$\hat{Y}_{new} = \mathcal{T}(x_{new}, \hat{f}_{\text{PLL}}, \mathcal{E}_{\max}, \epsilon) . \tag{11}$$

Let $\mathcal{O}'_{\text{calib}} = \{(x_i, y_i) : i \in \mathcal{I}_2\}$ be the precise counterpart of $\mathcal{O}_{\text{calib}}$, i.e., the underlying precise data that the PLL learner could not observe. Moreover, define

$$\mathcal{E}_1 := \left\{ 1 - \hat{f}_{\text{PLL}}(x_j)_{y_j} : j \in \mathcal{I}_2 \right\} .$$

The following theorem establishes the validity of the prediction sets made by this approach.

Theorem 1 *If the data points in $\mathcal{O}'_{calib} \cup (x_{new}, y_{new})$ are exchangeable, then the prediction set (11) with underlying score set (10) satisfies*

$$\mathbb{P}\Big(y_{new} \in \mathcal{T}(x_{new}, \hat{f}_{PLL}, \mathcal{E}_{max}, \epsilon) \Big) \geq 1 - \epsilon .$$

Proof The vanilla CP guarantees that the prediction set $\mathcal{T}(x_{new}, \hat{f}_{\text{PLL}}, \mathcal{E}_1, \epsilon)$ is valid. To conclude the proof, we show that $\mathcal{T}(x_{new}, \hat{f}_{\text{PLL}}, \mathcal{E}_1, \epsilon) \subseteq \mathcal{T}(x_{new}, \hat{f}_{\text{PLL}}, \mathcal{E}_{\max}, \epsilon)$. To this end, it is enough to show that $q(\mathcal{E}_{\max}, \epsilon) \geq q(\mathcal{E}_1, \epsilon)$, which immediately follows from

$$1 - \min_{y \in S_j} \hat{f}_{\text{PLL}}(x_j)_y \geq 1 - \hat{f}_{\text{PLL}}(x_j)_{y_j}$$

for all $j \in \mathcal{I}_2$. ■

Although the mentioned pessimistic approach preserves the validity of CP, it usually ends up in unnecessarily large prediction sets. Here we suggest another approach by incorporating the nonconformity scores of all candidate labels of each calibration instance. Indeed, consider the following set:

$$\mathcal{E}_{\text{all}} := \left\{ 1 - \hat{f}_{\text{PLL}}(x_j)_y : j \in \mathcal{I}_2 \text{ and } y \in S_j \right\} \tag{12}$$

Compared to the previous cases, $\mathcal{E}_{\text{all}}$ has a bigger cardinality. The following theorem shows that $\mathcal{T}(x_{new}, \hat{f}_{\text{PLL}}, \mathcal{E}_{\text{all}}, \epsilon)$ is also valid under certain assumptions.

Theorem 2 *For any $\epsilon \leq \min\left(\frac{1}{4}, \frac{|\mathcal{O}_{calib}| + |\mathcal{Y}|}{|\mathcal{Y}| \cdot (1 + |\mathcal{O}_{calib}|)}\right)$, if the points in $\mathcal{O}'_{calib} \cup \{(x_{new}, y_{new})\}$ are exchangeable and $q(\mathcal{E}_1, \epsilon) \leq 0.5$, then the prediction set $\mathcal{T}(x_{new}, \hat{f}_{PLL}, \mathcal{E}_{all}, \epsilon)$ is valid.*

Proof Again, we prove this result by showing $\mathcal{T}(x_{new}, \hat{f}_{\mathrm{PLL}}, \mathcal{E}_1, \epsilon) \subseteq \mathcal{T}(x_{new}, \hat{f}_{\mathrm{PLL}}, \mathcal{E}_{\mathrm{all}}, \epsilon)$. We start with the set $\mathcal{E}_1$ and show that by adding the nonconformity scores of the other (false) candidates $y \in S_j \setminus \{y_j\}$, its critical score can only get larger. This immediately follows from the observation that most of the nonconformity scores of false candidates are going to be added "to the right" of $q(\mathcal{E}_1, \epsilon)$, i.e., they exceed this value.

First, observe that for any instance $(x_j, S_j) \in \mathcal{O}_{\mathrm{calib}}$, the nonconformity score for any false label in a candidate set is at least as great as the predicted probability of the true class, i.e.,

$$1 - \hat{f}_{\mathrm{PLL}}(x_j)_y \geq 1 - \max_{y \in S_j \setminus \{y_j\}} \hat{f}_{\mathrm{PLL}}(x_j)_y \geq \hat{f}_{\mathrm{PLL}}(x_j)_{y_j}, \quad \forall y \in S_j \setminus \{y_j\}. \tag{13}$$

Let $t := \lceil (1+|\mathcal{E}_1|)(1-\epsilon) \rceil$. Take (x_l, y_l) as an instance that is among the t smallest elements of $\mathcal{E}_1$. By definition, we have $1 - \hat{f}_{\mathrm{PLL}}(x_l)_{y_l} \leq q(\mathcal{E}_1, \epsilon)$, which implies $\hat{f}_{\mathrm{PLL}}(x_l)_{y_l} \geq 1 - q(\mathcal{E}_1, \epsilon)$. Since $q(\mathcal{E}_1, \epsilon) \leq 0.5$, we have $\hat{f}_{\mathrm{PLL}}(x_l)_{y_l} \geq q(\mathcal{E}_1, \epsilon)$. This, together with (13), tells us that the nonconformity scores of all false labels will be located to the right of $q(\mathcal{E}_1, \epsilon)$. Hence for these t points, there will be at least t scores added to the right of $q(\mathcal{E}_1, \epsilon)$[1].

Now take (x_k, y_k) as an instance which is among the $|\mathcal{E}_1| - t$ greatest elements of $\mathcal{E}_1$. Again by definition, we have $\hat{f}_{\mathrm{PLL}}(x_k)_{y_k} < 1 - q(\mathcal{E}_1, \epsilon)$. For the sake of worst-case consideration, we can assume that $\hat{f}_{\mathrm{PLL}}(x_k)_{y_k} < q(\mathcal{E}_1, \epsilon)$ as well so that the nonconformity scores of all false labels of these $|\mathcal{E}_1| - t$ instances will be located on the left-hand side of $q(\mathcal{E}_1, \epsilon)$. Therefore, for these $|\mathcal{E}_1| - t$ points, there will be at most $(|\mathcal{E}_1| - t) \cdot (\mathcal{Y} - 1)$ scores added to the left-hand side of $q(\mathcal{E}_1, \epsilon)$.

Since $\epsilon \leq \frac{1}{4}$, all we need for $q(\mathcal{E}_{\mathrm{all}}, \epsilon)$ to be greater than or equal to $q(\mathcal{E}_1, \epsilon)$ is that the number of scores added to the right of $q(\mathcal{E}_1, \epsilon)$ to be greater than or equal to those added to its left-hand side (See Lemma B.1 for details), which is the case since

$$\begin{aligned}
\epsilon &\leq \frac{|\mathcal{O}_{\mathrm{calib}}| + |\mathcal{Y}|}{|\mathcal{Y}| \cdot (1 + |\mathcal{O}_{\mathrm{calib}}|)} \Rightarrow \\
1 - \epsilon &\geq \frac{|\mathcal{O}_{\mathrm{calib}}| \cdot (|\mathcal{Y}| - 1)}{|\mathcal{Y}| \cdot (1 + |\mathcal{O}_{\mathrm{calib}}|)} \Rightarrow \\
t &\geq \frac{|\mathcal{O}_{\mathrm{calib}}| \cdot (|\mathcal{Y}| - 1)}{|\mathcal{Y}|} \Rightarrow \\
t &\geq (|\mathcal{O}_{\mathrm{calib}}| - t) \cdot (|\mathcal{Y}| - 1)
\end{aligned}$$

■

The condition $q(\mathcal{E}_1, \epsilon) \leq 0.5$ implies that the error rate of the induced classifier on calibration data must not exceed $\epsilon \cdot 100\%$. Later on, in Section 4, we will see that the validity of prediction sets generated by this method holds even when this condition is violated, and also, compared to the pessimistic approach $\mathcal{E}_{\max}$, this method results in more efficient (i.e., smaller) prediction sets.

[1] It is obvious that the instance (x_l, S_l) will add $|S_l - 1|$ scores to the right of $q(\mathcal{E}_1, \epsilon)$.

Another possibility is to consider the average of the nonconformity scores per calibration instance. Hence, the set of nonconformity scores would be

$$\mathcal{E}_{\text{mean}} := \{1 - \frac{\sum_{y \in S_j} \hat{f}_{\text{PLL}}(x_j)_y}{|S_j|} : j \in \mathcal{I}_2\}. \tag{14}$$

The validity of the prediction set $\mathcal{T}(x_{new}, \hat{f}_{\text{PLL}}, \mathcal{E}_{\text{mean}}, \epsilon)$ is established by the following theorem.

Theorem 3 *If the data points in $\mathcal{O}'_{calib} \cup (x_{new}, y_{new})$ are exchangeable and $\hat{f}_{PLL}(x_j)_{y_j} \geq \frac{1}{|S_j|}$, $\forall j \in \mathcal{I}_2$, then the prediction set $\mathcal{T}(x_{new}, \hat{f}_{PLL}, \mathcal{E}_{mean}, \epsilon)$ is valid.*

Proof $\mathcal{T}(x_{new}, \hat{f}_{\text{PLL}}, \mathcal{E}_1, \epsilon) \subseteq \mathcal{T}(x_{new}, \hat{f}_{\text{PLL}}, \mathcal{E}_{\text{mean}}, \epsilon)$ holds because

$$\begin{aligned} \hat{f}_{\text{PLL}}(x_j)_{y_j} \geq \frac{1}{|S_j|} \geq \frac{\sum_{y \in S_j} \hat{f}_{\text{PLL}}(x_j)_y}{|S_j|}, \quad \forall j \in \mathcal{I}_2 \Rightarrow \\ 1 - \frac{\sum_{y \in S_j} \hat{f}_{\text{PLL}}(x_j)_y}{|S_j|} \geq 1 - \hat{f}_{\text{PLL}}(x_j)_{y_j}, \quad \forall j \in \mathcal{I}_2 \Rightarrow \\ q(\mathcal{E}_{\text{mean}}, \epsilon) \geq q(\mathcal{E}_1, \epsilon). \end{aligned}$$

■

The requirement of this theorem is demanding, especially for the case where most candidate sets consist of only two labels. However, it should be noted that it is likely to have valid prediction sets using this approach, even if, for some calibration instances, the average nonconformity score over the candidate set falls below the nonconformity score of the precise counterpart. Furthermore, as the cardinality of the candidate sets increases, this requirement becomes less burdensome.

4. Experiments

In this section, we evaluate the performance of the proposed frameworks numerically. Apart from the three approaches mentioned in previous sections, there are other approaches to form the set of nonconformity scores that are not necessarily coming with a coverage guarantee. Here, we bring two of them, which we consider in our comparisons as well:

- Taking minimum nonconformity score per calibration instance:

  $$\mathcal{E}_{\text{min}} := \left\{1 - \max_{y \in S_j} \hat{f}_{\text{PLL}}(x_j)_y : j \in \mathcal{I}_2\right\}. \tag{15}$$

  This optimistic approach is a natural baseline for the comparison. It can be seen as *calibration with the disambiguated data*, where the induced classifier is utilized to disambiguate the calibration data, and the calibration proceeds as in the vanilla conformal prediction.

Table 1: Description of the benchmark and real datasets.

		FashionMNIST	KMNIST	MNIST	BirdSong	Lost	MSRCv2	Soccer Player	Yahoo!News
	Num. of classes	10	10	10	13	16	23	171	219
Avg. CSS	Original	-	-	-	2.18	2.23	3.16	2.09	1.91
	Instance-dependent contamination	2.32	2.49	2.25	-	-	-	-	-
	Random contamination (p=0.1)	2.29	2.29	2.29	-	-	-	-	-
	Random contamination (p=0.7)	7.30	7.30	7.30	-	-	-	-	-

- Taking the weighted average of minimum and maximum nonconformity score per calibration instance:

$$\mathcal{E}_\mu := \left\{ \mu \cdot \left(1 - \max_{y \in S_j} \hat{f}_{\mathrm{PLL}}(x_j)_y\right) + (1-\mu) \cdot \left(1 - \min_{y \in S_j} \hat{f}_{\mathrm{PLL}}(x_j)_y\right) : j \in \mathcal{I}_2 \right\}, \qquad (16)$$

  with $\mu \in [0,1]$ being a hyperparameter. This approach lies between the pessimistic and optimistic ones. With a proper selection of μ, one might be able to achieve prediction sets that are both valid and efficient.

Our implementation code is publicly available on GitHub[2] to enable the reproducibility of the presented results.

4.1. Datasets

Experiments are performed using the benchmark datasets: MNIST (LeCun et al., 1998), Kuzushiji-MNIST (Clanuwat et al., 2018), and Fashion-MNIST (Xiao et al., 2017). However, these datasets are all precise and need to be synthetically contaminated. We use the following two methods to convert these datasets into partially labeled data:

- Random contamination: In this method, we create a candidate set for each instance in a random manner by including each non-ground-truth label with probability p. In cases where no label among the non-ground-truth labels is added to the set, a random label is added to ensure all data is partially labeled.

- Instance-dependent contamination: Similar to Xu et al. (2021), we train a simple classifier $\hat{f}_s$ on each benchmark dataset which we refer to as the supermodel. For each dataset, we exploit its supermodel to compute the probability of adding each non-ground-truth label $y \in \mathcal{Y} \setminus \{y_i\}$ to the candidate set of instance x_i as $p_y = \frac{\hat{f}_s(x_i)_y}{\max_{y \in \mathcal{Y} \setminus \{y_i\}} \hat{f}_s(x_i)_y}$. More details about the supermodels can be found in Appendix A.

In addition to the synthetic datasets, we adopt five commonly used real-world partial label datasets: Lost (Cour et al., 2011), MSRCv2 (Liu and Dietterich, 2012), BirdSong (Briggs et al., 2012), Soccer Player (Zeng et al., 2013) and Yahoo!News (Guillaumin et al., 2010). Table 1 gives an overview of the benchmark and real-world datasets, including the number of classes and the average candidate set sizes (CSS).

[2] https://github.com/pwhofman/conformal-partial-labels

Table 2: Performance comparison of different calibration approaches on benchmark datasets with random contamination ($p = 0.1$).

		FashionMNIST	KMNIST	MNIST
	Train acc.	96.64 ± 0.37	98.83 ± 0.02	99.50 ± 0.02
	Test acc.	88.53 ± 0.44	90.66 ± 0.20	98.12 ± 0.13
$\mathcal{E}_{\max}$	Efficiency	10.00 ± 0.00	9.42 ± 0.05	9.17 ± 0.05
	Coverage	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
$\mathcal{E}_{\text{all}}$	Efficiency	8.12 ± 0.17	8.83 ± 0.04	8.38 ± 0.05
	Coverage	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
$\mathcal{E}_{\text{mean}}$	Efficiency	1.07 ± 0.01	1.03 ± 0.00	1.01 ± 0.00
	Coverage	0.91 ± 0.01	0.92 ± 0.00	0.98 ± 0.00
$\mathcal{E}_{\min}$	Efficiency	0.98 ± 0.00	0.80 ± 0.00	0.90 ± 0.00
	Coverage	0.88 ± 0.00	0.79 ± 0.00	0.90 ± 0.00
$\mathcal{E}_{\mu=0.3}$	Efficiency	1.08 ± 0.01	1.04 ± 0.00	1.01 ± 0.00
	Coverage	0.92 ± 0.00	0.92 ± 0.00	0.99 ± 0.00
$\mathcal{E}_{\mu=0.5}$	Efficiency	1.04 ± 0.00	0.98 ± 0.00	0.99 ± 0.00
	Coverage	0.90 ± 0.00	0.90 ± 0.00	0.98 ± 0.00
$\mathcal{E}_{\mu=0.7}$	Efficiency	1.02 ± 0.00	0.93 ± 0.00	0.98 ± 0.00
	Coverage	0.89 ± 0.00	0.88 ± 0.00	0.97 ± 0.00

4.2. Models

We bring a well-known partial label learning algorithm, PRODEN (Lv et al., 2020), which progressively tries to find the ground-truth label and adjusts partial labels accordingly. For each benchmark dataset, we employ a multi-layer perceptron (MLP) of five layers with $784 - 300 - 300 - 300 - 300 - 10$ units and use ReLU as an activation function. This model is optimized using stochastic gradient descent (SGD) algorithm (Robbins and Monro, 1951) with a learning rate of 0.1, a momentum of 0.9, and a weight decay at 0.001, 0.0001, and 0.00001 for MNIST, Kuzushiji-MNIST, and Fashion-MNIST, respectively. The model is trained for 100 epochs and the learning rate is adjusted using cosine annealing (Loshchilov and Hutter, 2017). For real-world datasets, a softmax regression model is used. This model is optimized with the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 0.1, 0.1, 0.01, 0.01, and 0.01 and weight decay at 10^{-10}, 10^{-6}, 10^{-10}, 10^{-2}, and 10^{-6} for the Lost, MSCRCv2, BirdSong, Soccer Player, and Yahoo!News dataset, respectively. The model is trained for 200 epochs, and the learning rate is also adjusted using cosine annealing.

Table 3: Performance comparison of different calibration approaches on benchmark datasets with random contamination ($p = 0.7$).

		FashionMNIST	KMNIST	MNIST
	Train acc.	88.22 ± 0.82	93.55 ± 0.2	97.29 ± 0.03
	Test acc.	85.7 ± 0.76	82.25 ± 0.38	96.88 ± 0.22
$\mathcal{E}_{\text{max}}$	Efficiency	9.58 ± 0.06	9.82 ± 0.02	9.63 ± 0.03
	Coverage	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
$\mathcal{E}_{\text{all}}$	Efficiency	8.95 ± 0.03	9.40 ± 0.04	8.93 ± 0.03
	Coverage	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
$\mathcal{E}_{\text{mean}}$	Efficiency	1.31 ± 0.03	1.35 ± 0.00	1.08 ± 0.00
	Coverage	0.94 ± 0.00	0.90 ± 0.00	0.99 ± 0.00
$\mathcal{E}_{\text{min}}$	Efficiency	0.93 ± 0.01	0.81 ± 0.01	0.91 ± 0.00
	Coverage	0.83 ± 0.01	0.74 ± 0.01	0.90 ± 0.00
$\mathcal{E}_{\mu=0.3}$	Efficiency	1.20 ± 0.02	1.15 ± 0.01	1.02 ± 0.00
	Coverage	0.92 ± 0.00	0.86 ± 0.00	0.98 ± 0.00
$\mathcal{E}_{\mu=0.5}$	Efficiency	1.09 ± 0.01	1.01 ± 0.01	0.99 ± 0.00
	Coverage	0.89 ± 0.01	0.83 ± 0.00	0.97 ± 0.00
$\mathcal{E}_{\mu=0.7}$	Efficiency	1.01 ± 0.01	0.92 ± 0.01	0.97 ± 0.00
	Coverage	0.87 ± 0.01	0.80 ± 0.00	0.95 ± 0.00

4.3. Results

For the real-world datasets, since there are no separate test datasets, we apply the $80\%-20\%$ train-test split. For synthetic datasets, 10% of the available training data is selected at random for calibration, while for the real-world datasets, a larger subset of 20% is selected. We fix the miscoverage rate at $\epsilon = 0.1$. The experiments are repeated five times using different random seeds, and the means and standard deviations of the results are reported.

Table 2 and Table 3 present the results for benchmark datasets, where random contamination is applied with p being set to 0.1 and 0.7, respectively. It can be seen that $\mathcal{E}_{\text{max}}$ and $\mathcal{E}_{\text{all}}$ results in highly inefficient prediction sets. In fact, when the accuracy of the underlying classifier is high, and the candidate sets are generated in a completely random fashion, then the nonconformity scores of the non-ground-truth labels are so high, resulting in large critical scores and, accordingly, such conservative prediction sets.

Table 4 provides the results for the benchmark datasets with instance-dependent contamination. Compared to the random contamination case, the results of $\mathcal{E}_{\text{max}}$ and $\mathcal{E}_{\text{all}}$ are less inefficient, while they are the only cases that satisfy the coverage guarantee for all datasets. Finally, the results for the real-world datasets are provided in Table 5. Once again, $\mathcal{E}_{\text{max}}$ and $\mathcal{E}_{\text{all}}$ result in inefficient large prediction sets. While $\mathcal{E}_{\text{mean}}$ provides the best

Table 4: Performance comparison of different calibration approaches on benchmark datasets with instance-dependent contamination.

		FashionMNIST	KMNIST	MNIST
	Train acc.	83.82 ± 0.63	93.86 ± 0.16	97.23 ± 0.11
	Test acc.	82.48 ± 0.69	83.9 ± 0.31	96.93 ± 0.17
$\mathcal{E}_{\max}$	Efficiency	6.42 ± 0.76	6.54 ± 0.12	5.97 ± 0.14
	Coverage	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
$\mathcal{E}_{\text{all}}$	Efficiency	3.46 ± 0.54	5.14 ± 0.06	4.42 ± 0.07
	Coverage	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
$\mathcal{E}_{\text{mean}}$	Efficiency	1.14 ± 0.03	1.11 ± 0.00	1.02 ± 0.00
	Coverage	0.87 ± 0.01	0.87 ± 0.00	0.98 ± 0.00
$\mathcal{E}_{\min}$	Efficiency	0.91 ± 0.01	0.84 ± 0.01	0.90 ± 0.00
	Coverage	0.77 ± 0.01	0.76 ± 0.01	0.89 ± 0.00
$\mathcal{E}_{\mu=0.3}$	Efficiency	1.23 ± 0.04	1.09 ± 0.00	1.02 ± 0.00
	Coverage	0.89 ± 0.01	0.87 ± 0.00	0.98 ± 0.00
$\mathcal{E}_{\mu=0.5}$	Efficiency	1.06 ± 0.03	0.99 ± 0.00	1.00 ± 0.00
	Coverage	0.85 ± 0.01	0.84 ± 0.00	0.97 ± 0.00
$\mathcal{E}_{\mu=0.7}$	Efficiency	0.98 ± 0.00	0.93 ± 0.00	0.97 ± 0.00
	Coverage	0.82 ± 0.01	0.81 ± 0.00	0.96 ± 0.00

results for the synthetic data with random contamination case, the coverage property is not always satisfied for this approach in the two other cases.

5. Conclusion

This paper bridges two popular machine learning frameworks, namely conformal prediction and partial label learning. We propose an extension to conformal prediction, which allows it to handle training and calibration data that are only partially labeled. This is an essential extension as such data arises in many real-world applications, such as web mining, image annotation, text classification, etc., where obtaining complete label information may be difficult or expensive, and equipping predictions with a notion of uncertainty is of utmost importance.

Since this is the first paper dealing with partial label data for conformal prediction, there are various open problems and research directions to pursue in future work. For example, it is worth exploring whether there is room for enhancing the computation of the nonconformity scores in the calibration step. Indeed, while we theoretically show that the prediction sets constructed by the proposed approaches inherit the validity of the conformal predic-

Table 5: Performance comparison of different calibration approaches on real-world datasets.

		BirdSong	Lost	MSRCv2	Soccer Player	Yahoo!News
	Train acc.	74.45 ± 0.69	86.54 ± 3.29	54.61 ± 0.73	52.61 ± 0.48	69.86 ± 0.84
	Test acc.	72.04 ± 0.89	73.96 ± 2.43	48.58 ± 0.72	50.48 ± 0.39	61.13 ± 0.77
$\mathcal{E}_{\text{max}}$	Efficiency	13.00 ± 0.00	13.27 ± 1.05	21.4 ± 1.51	150.31 ± 1.09	122.84 ± 4.11
	Coverage	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.99 ± 0.00	0.99 ± 0.00
$\mathcal{E}_{\text{all}}$	Efficiency	9.80 ± 1.80	10.7 ± 0.83	18.16 ± 0.69	137.81 ± 1.30	48.07 ± 1.87
	Coverage	0.99 ± 0.00	0.99 ± 0.01	0.97 ± 0.01	0.98 ± 0.00	0.99 ± 0.00
$\mathcal{E}_{\text{mean}}$	Efficiency	2.06 ± 0.14	1.92 ± 0.22	2.85 ± 0.41	20.71 ± 4.64	2.91 ± 0.08
	Coverage	0.89 ± 0.01	0.88 ± 0.02	0.66 ± 0.02	0.75 ± 0.02	0.87 ± 0.01
$\mathcal{E}_{\text{min}}$	Efficiency	1.62 ± 0.15	1.56 ± 0.15	2.15 ± 0.24	16.12 ± 3.61	2.48 ± 0.10
	Coverage	0.84 ± 0.02	0.86 ± 0.02	0.63 ± 0.03	0.73 ± 0.02	0.84 ± 0.01
$\mathcal{E}_{\mu=0.3}$	Efficiency	2.23 ± 0.12	2.15 ± 0.26	3.07 ± 0.28	23.7 ± 4.35	3.17 ± 0.11
	Coverage	0.90 ± 0.01	0.89 ± 0.02	0.67 ± 0.02	0.76 ± 0.02	0.88 ± 0.01
$\mathcal{E}_{\mu=0.5}$	Efficiency	1.97 ± 0.13	1.90 ± 0.20	2.58 ± 0.29	20.17 ± 4.14	2.87 ± 0.08
	Coverage	0.88 ± 0.01	0.88 ± 0.02	0.65 ± 0.02	0.75 ± 0.02	0.87 ± 0.01
$\mathcal{E}_{\mu=0.7}$	Efficiency	1.81 ± 0.15	1.74 ± 0.16	2.40 ± 0.21	18.53 ± 3.72	2.67 ± 0.09
	Coverage	0.86 ± 0.01	0.87 ± 0.02	0.65 ± 0.02	0.74 ± 0.02	0.85 ± 0.01

tion under certain assumptions, their efficiency still needs to be improved. Moreover, as our theoretical validity results rely on the specific properties of nonconformity scores derived from probabilistic classifiers, another interesting contribution would be a generalization of these results to other types of nonconformity measures.

Acknowledgments

Alireza Javanmardi was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Project number 451737409. Yusuf Sale was supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

Forrest Briggs, Xiaoli Z. Fern, and Raviv Raich. Rank-loss support instance machines for miml instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 534–542. Association for Computing Machinery, 2012. URL https://doi.org/10.1145/2339530.2339616.

Vivien Cabannnes, Alessandro Rudi, and Francis Bach. Structured prediction with partial labelling through the infimum loss. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1230–1239. PMLR, 2020. URL https://proceedings.mlr.press/v119/cabannnes20a.html.

Vivien A Cabannnes, Francis Bach, and Alessandro Rudi. Disambiguation of weak supervision leading to exponential convergence rates. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 1147–1157. PMLR, 2021. URL https://proceedings.mlr.press/v139/cabannnes21a.html.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature, 2018.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12(42):1501–1536, 2011. URL http://jmlr.org/papers/v12/cour11a.html.

Inés Couso and Luciano Sánchez. Machine learning models, epistemic set-valued data and generalized loss functions: an encompassing approach. *Information Sciences*, 358:129–150, 2016.

Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007.

Yves Grandvalet. Logistic regression for partial labels. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 1935–1941, 2002.

Matthieu Guillaumin, Jakob J. Verbeek, and Cordelia Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *European Conference on Computer Vision*, pages 634–647, 2010.

Eyke Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.

Eyke Hüllermeier and Jürgen Beringer. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439, 2006.

Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/file/653ac11ca60b3e021a8c609c7198acfc-Paper.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Liping Liu and Thomas Dietterich. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/aaebdb8bb6b0e73f6c3c54a0ab0c6415-Paper.pdf.

Liping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1629–1637. PMLR, 2014. URL https://proceedings.mlr.press/v32/liug14.html.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6500–6510. PMLR, 2020. URL https://proceedings.mlr.press/v119/lv20a.html.

Nam Nguyen and Rich Caruana. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–559, 2008.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002a.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Qualified prediction for large data sets in the case of pattern recognition. In *International Conference on Machine Learning and Applications*, pages 159–163, 2002b.

Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Reliable diagnosis of acute abdominal pain with conformal prediction. *Engineering Intelligent Systems*, 17(2):127–137, 2009.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. URL http://www.jstor.org/stable/2236626.

Fredrik Svensson, Natalia Aniceto, Ulf Norinder, Isidro Cortes-Ciriano, Ola Spjuth, Lars Carlsson, and Andreas Bender. Conformal regression for quantitative structure–activity relationship modeling—quantifying prediction uncertainty. *Journal of Chemical Information and Modeling*, 58(5):1132–1140, 2018.

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

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Ning Xu, Congyu Qiao, Xin Geng, and Min-Ling Zhang. Instance-dependent partial label learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 27119–27130. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e38e37a99f7de1f45d169efcdb288dd1-Paper.pdf.

Marco Zaffalon and Enrique Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34:757–821, 2009.

Zinan Zeng, Shijie Xiao, Kui Jia, Tsung-Han Chan, Shenghua Gao, Dong Xu, and Yi Ma. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Appendix A.

An MLP with $784 - 100 - 10$ units with ReLU activation functions is used as a supermodel for the MNIST, Kuzushiji-MINST, and Fashion-MNIST datasets. Table 6 reports the train and test accuracies of these supermodels.

Table 6: Accuracies of the supermodels used for instance-dependent contamination.

	FashionMNIST	KMNIST	MNIST
Train accuracy	85.13	93.70	96.56
Test accuracy	82.88	78.35	94.77

Note that the only purpose of the supermodels is to convert precise datasets into partially labeled data. Moreover, supermodel training is independent of partial label learning and calibration steps. Indeed, the supermodel for each benchmark dataset is trained using the training set of that data. Subsequently, the resulting supermodel is utilized to generate partial labels for the same set. The resulting contaminated set will, later on, be divided into proper training and calibration subsets. These subsets will be employed in the partial label learning and calibration steps, respectively.

Appendix B.

Lemma B.1 *Consider a set $\mathcal{E}_1$ and its $\lceil (1+|\mathcal{E}_1|)(1-\epsilon)\rceil$ smallest value, $q(\mathcal{E}_1,\epsilon)$. Suppose we add $t_l \geq 2$ elements that are less than or equal to $q(\mathcal{E}_1,\epsilon)$ and $t_r \geq t_l$ elements that are greater than $q(\mathcal{E}_1,\epsilon)$ to form a new set $\mathcal{E}_2$. If $\epsilon \leq \frac{1}{4}$, then it is guaranteed that $q(\mathcal{E}_2,\epsilon) \geq q(\mathcal{E}_1,\epsilon)$.*

Proof Let d be the difference between t_r and t_l, i.e., $t_r = t_l + d$. For $q(\mathcal{E}_2,\epsilon) \geq q(\mathcal{E}_1,\epsilon)$ to be true, we need the following to hold:

$$\begin{aligned}
\lceil (1+|\mathcal{E}_2|)(1-\epsilon)\rceil &\geq \lceil (1+|\mathcal{E}_1|)(1-\epsilon)\rceil + t_l \Rightarrow \\
(1+|\mathcal{E}_1|+t_l+t_r)(1-\epsilon) &\geq (1+|\mathcal{E}_1|)(1-\epsilon) + t_l + 1 \Rightarrow \\
(t_l+t_r)(1-\epsilon) &\geq t_l + 1 \Rightarrow \\
\epsilon \leq \frac{t_l+d-1}{2t_l+d} &= \frac{1}{2} + \frac{d/2-1}{2t_l+d} \geq \frac{1}{4},
\end{aligned}$$

where the last inequality comes from the fact that $d \geq 0$ and $t_l \geq 2$. ■