An Information Theoretic Perspective on Conformal Prediction

Alvaro H.C. Correia **Fabio Valerio Massoli** **Christos Louizos**[†] **Arash Behboodi**[†]
Qualcomm AI Research[*]
Amsterdam, The Netherlands
{acorreia, fmassoli, clouizos, behboodi}@qti.qualcomm.com

Abstract

Conformal Prediction (CP) is a distribution-free uncertainty estimation framework that constructs prediction sets guaranteed to contain the true answer with a user-specified probability. Intuitively, the size of the prediction set encodes a general notion of uncertainty, with larger sets associated with higher degrees of uncertainty. In this work, we leverage information theory to connect conformal prediction to other notions of uncertainty. More precisely, we prove three different ways to upper bound the intrinsic uncertainty, as described by the conditional entropy of the target variable given the inputs, by combining CP with information theoretical inequalities. Moreover, we demonstrate two direct and useful applications of such connection between conformal prediction and information theory: (i) more principled and effective *conformal training* objectives that generalize previous approaches and enable end-to-end training of machine learning models from scratch, and (ii) a natural mechanism to incorporate side information into conformal prediction. We empirically validate both applications in centralized and federated learning settings, showing our theoretical results translate to lower *inefficiency* (average prediction set size) for popular CP methods.

1 Introduction

Machine learning (ML) models have rapidly grown in popularity and reach, having now found use in many safety-critical domains like healthcare [1] and autonomous driving [28]. In these areas, predictions must be accompanied by reliable measures of uncertainty to ensure safe decision-making. However, most ML models are designed and trained to produce only point estimates, which capture only crude notions of uncertainty with no statistical guarantees. Conformal prediction (CP) [64], in particular its split variant (SCP) [47], has recently gained in popularity as a principled and scalable solution to equip *any*, potentially black-box, model with proper uncertainty estimates in the form of prediction sets; in loose terms, larger sets are associated with higher degrees of uncertainty.

In this work, we take a closer look at conformal prediction through the lens of information theory (IT), establishing a connection between conformal prediction and the underlying intrinsic uncertainty of the data-generating process, as captured by the conditional entropy $H(Y|X) = -E_{P_{XY}}[\log P_{Y|X}]$ of the target variable Y given the inputs X. We prove conformal prediction can be used to bound $H(Y|X)$ from above in three different ways: one derived from the data processing inequality, which we dub *DPI bound*, and two coming from a variation of Fano's inequality [19], a model agnostic one, which we call (simple) *Fano bound*, and another informed by the predictive model itself, to which we refer as *model-based Fano bound*. To the best of our knowledge, these bounds represent the first bridge connecting information theory and conformal prediction, which we hope will bring new tools to both fields. We already present two such tools in this paper: (i) we show our upper bounds serve as

[*]Qualcomm AI Research, Qualcomm Technologies Netherlands B.V (Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.). [†]Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

principled training objectives to learn classifiers that are more amenable to SCP, and (ii) we advance a systematic way to incorporate side information into the construction of prediction sets. In a number of classification tasks, we empirically validate that both these applications of our theoretical results lead to better predictive efficiency, i.e., narrower and, consequently, more informative prediction sets.

The rest of the paper is organized as follows. In Section 2, we first introduce the necessary background to guide the reader through our main theoretical results. We introduce our three new upper bounds to the intrinsic uncertainty in Section 3, and their applications to conformal training and side information in Sections 4 and 5, respectively. Thereafter, we explore the related work in CP and IT in Section 6, present and analyze our experimental results in Section 7, and finally conclude in Section 8.

2 Background

In this section, we present the needed background on conformal prediction and list decoding [18, 22, 68], an area of information theory that, as we show, is closely related to CP and especially useful in deriving our main results. We start with the necessary notation. As usual, we denote random variables in uppercase letters and their realization in lowercase, e.g., $X = x$. We reserve calligraphic letters, e.g. $\mathcal{X}$, for sets and use $P_X, Q_X, \dots$ to denote probability measures on the space $\mathcal{X}$. To simplify the notation, we use $P, Q, \dots$ when the underlying space is clear. For example, given a probability measure Q_{XY}, the probability of the event $\{(X, Y) : Y \in \mathcal{C}(X)\}$ is denoted as $Q(Y \in \mathcal{C}(X))$.

2.1 Conformal Prediction

Conformal prediction (CP) is a theoretically grounded framework that provides *prediction sets* with finite-sample guarantees under minimal distribution-free assumptions. Concretely, given a set of n data points $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \dots, n$ drawn from some (unknown) joint distribution P_{XY}, CP allows us to construct sets $\mathcal{C}(X) \in \mathcal{Y}$, such that for a new data point from the same distribution (X_{test}, Y_{test}) we have the following guarantee for a target error rate $\alpha \in (0, 1)$

$$\mathbb{P}(Y_{test} \in \mathcal{C}(X_{test})) \geq 1 - \alpha, \tag{1}$$

where the probability is over the randomness in the sample $\{(X_i, Y_i)\}_{i=1}^n \cup \{(X_{test}, Y_{test})\}$. To make this more tangible, the reader can picture $\mathcal{C}(X)$ as a subset of the possible labels in a classification problem, or as a confidence interval around the point estimate of a regressor in a regression setting.

In this work, we focus on a variant called split conformal prediction (SCP) [47] that gained popularity in the ML community, since it can leverage *any* pre-trained model $f : \mathcal{X} \to \mathcal{Y}$ in the construction of prediction sets. In this setting, the aforesaid n data points constitute a *calibration data set* $\mathcal{D}_{cal}$, which must be disjoint from the training data set used to fit the predictive model f. This separation between training and calibration data is what gives the name *split* to the method.

The first step in SCP is to define a *nonconformity score function* $s_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which is itself a function of model f and captures the magnitude of the prediction error at a given data point; the higher the score $s_f(x, y)$, the higher the disagreement between input x and prediction y. At calibration time, we evaluate the score function at every $(X_i, Y_i) \in \mathcal{D}_{cal}$ to get a collection of scores $\{S_i = s_f(X_i, Y_i)\}_{i=1}^n$, and at test time, we construct prediction set $\mathcal{C}(X_{test})$ as

$$\mathcal{C}(X_{test}) = \{y \in \mathcal{Y} : s(X_{test}, y) \leq \text{Quantile}(1 - \alpha; \{S_i\}_{i=1}^n \cup \{\infty\})\}, \tag{2}$$

where $\text{Quantile}(1 - \alpha; \{S_i\}_{i=1}^n)$ is the level $1 - \alpha$ quantile of the empirical distribution defined by $\{S_i\}_{i=1}^n$. The central result in conformal prediction, which we restate below for completeness, proves that prediction sets thus constructed achieve *marginal valid coverage*, i.e., satisfy (1).

Theorem 2.1 ([31, 64]). *If $\{(X_i, Y_i)\}_i^n$ are i.i.d. (or only exchangeable), then for a new i.i.d. draw (X_{test}, Y_{test}), and for any $\alpha \in (0, 1)$ and for any score function s such that $\{S_i\}_{i=1}^n$ are almost surely distinct, then $\mathcal{C}(X_{test})$ as defined above satisfies*

$$1 - \alpha \leq \mathbb{P}(Y_{test} \in \mathcal{C}(X_{test})) \leq 1 - \alpha_n, \quad \textit{where} \quad \alpha_n = \alpha - {}^1\!/_{n+1}.$$

See Appendix B for the proof and a more thorough introduction to CP. It is worth noting that valid coverage is not sufficient; the uninformative set predictor that always outputs $\mathcal{C}(X_{test}) = \mathcal{Y}$ trivially satisfies (1). We would also like our prediction sets to be as narrow as possible, and that is why CP methods are often compared in terms of their (empirical) *inefficiency*, i.e., the average prediction set size ${}^1\!/_{|D_{test}|} \sum_{x \in \mathcal{D}_{test}} |\mathcal{C}(x)|$ for some test data set $\mathcal{D}_{test}$. This is, in fact, not the only type of inefficiency criterion, but we use it as our main performance metric since it is the most common [65].

We depict the split conformal prediction procedure in Figure 1, where we include two extra variables that will be useful later in the text: the model prediction $\hat{Y} = f(X)$, and the event of valid coverage $E = \{Y \in \mathcal{C}(X)\}$, i.e., the event of the prediction set containing the correct class.

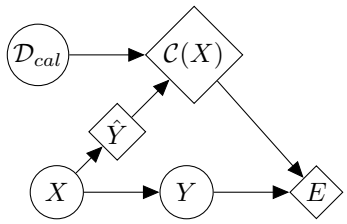


Figure 1: Graphical model of SCP. $\mathcal{D}_{cal}$ is a calibration set, $\mathcal{C}(X)$ the prediction set, $\hat{Y} = f(X)$ the model prediction, and E the event $\{Y \in \mathcal{C}(X)\}$. Square and round nodes are, respectively, deterministic and stochastic functions of their parents.

2.2 Conformal Prediction as List Decoding

In a nutshell, we can see conformal prediction as defining a map from an input x to a set of candidates in the target set $\mathcal{Y}$. It turns out that the conformal prediction framework is equivalent to (variable-size) list decoding, an error-recovery model going back to the works of Elias [18] and Wozencraft [68] in communication theory—we review some of these results in Appendix C.2. In particular, consider the mapping from the true label y to the input x as a noisy communication channel $p(x|y)$. The goal of an error-correcting code is then to decode the input x and recover the one true label y. List decoding generalizes this idea, allowing the decoder to return a set of outcomes (a list) instead of a pointwise prediction. If the correct *solution* is not part of the set output by the decoder, an error is declared. Although conformal prediction and list decoding were developed for different purposes, namely uncertainty quantification and error correction, it is easy to see that, if we allow for variable-size lists, the list decoding problem for the channel $p(x|y)$ as described above is equivalent to the conformal prediction problem.

To our knowledge, this link between conformal prediction and information theory (and list decoding in particular) has gone unnoticed in the literature, and in this paper we leverage it in two directions. First, we apply information-theoretic inequalities for list decoding to upper bound the conditional entropy $H(Y|X)$ of the data-generating process. This leads to new objectives for conformal training (see Section 4) and new bounds on the inefficiency of a given model (see Appendices E and G.3). Second, the information-theoretic interpretation of CP gives us an effective and theoretically grounded way of incorporating side-information into CP to improve predictive efficiency (see Section 5).

3 Information Theory Applied to Conformal Prediction

In this section, we develop our main results, which link information theory and conformal prediction. Concretely, we provide three novel upper bounds on the conditional entropy $H(Y|X)$: one coming from the data processing inequality and two derived in a similar way to Fano's inequality. We defer the proofs to Appendix D for the sake of conciseness, but in broad strokes, our results come from relating the bounds on the error probability provided by these information-theoretic inequalities (typically in the context of error-correcting codes) to the guarantees provided by CP in Theorem 2.1.

3.1 Data Processing Inequality for Conformal Prediction

We start by using the classical data processing inequality (DPI) in the context of conformal prediction. Specifically, we focus on the DPI for f-divergences, which we discuss thoroughly in Appendix C.1. In brief, for a convex function f with $f(1) = 0$, the f-divergence between two probability measures P and Q is defined as (see [58])

$$D_f(P||Q) := \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right].$$

In particular, with $f(x) = x \log x$ we recover the familiar notion of KL-divergence [27]. The DPI for f-divergences states that for any two probability measures P_X and Q_X defined on a space $\mathcal{X}$, and any map $W_{Y|X}$, which maps (P_X, Q_X) to (P_Y, Q_Y), we have

$$D_f(P_X||Q_X) \geq D_f\left(P_Y||Q_Y\right).$$

We can apply the DPI for f-divergence above in the context of conformal prediction by considering the probability of the event of valid coverage $\{Y \in \mathcal{C}(X)\}$ under two different probability measures P and Q. Taking P as the data-generating distribution $P := P_X P_{Y|X}$ and constructing $Q := P_X Q_{Y|X}$ for an arbitrary $Q_{Y|X}$ (e.g., a machine learning model), we get the following proposition.

Proposition 3.1 (DPI Bound)**.** Consider any conformal prediction method satisfying the upper and lower bounds of Theorem 2.1 for $\alpha \in (0, 0.5)$. For any arbitrary conditional distribution $Q_{Y|X}$, the true conditional distribution $P_{Y|X}$ and the input measure P_X, define the following two measures $Q := P_X Q_{Y|X}$ and $P := P_X P_{Y|X}$. Then, we have

$$H(Y|X) \leq h_b(\alpha) + (1-\alpha)\log Q(Y \in \mathcal{C}(X)) + \alpha_n \log Q(Y \notin \mathcal{C}(X)) - \mathbb{E}_P\left[\log Q_{Y|X}\right], \quad (3)$$

with $\alpha_n = \alpha - 1/n+1$ and $h_b(\cdot)$ the binary entropy function $h_b(\alpha) = -\alpha\log(\alpha) - (1-\alpha)\log(1-\alpha)$.

Note that the entropy term $H(Y|X)$ is computed using the measure P. We relegate the proof to Appendix D.1. In the bound in (3), the term $Q(Y \in \mathcal{C}(X))$ appears inside a log, so an empirical estimate $\hat{Q}(Y \in \mathcal{C}(X))$ would result in a lower bound and would be biased. We can provide an upper confidence bound on this estimate using the empirical Bernstein inequality [41] and use that instead. Based on the empirical Bernstein inequality, with probability $1-\delta$, we have

$$\Delta_\delta(\mathbf{Z}, n) := \sqrt{\frac{2V_n(\mathbf{Z})\log(2/\delta)}{n}} + \frac{7\log(2/\delta)}{3(n-1)}$$

$$Q(Y \in \mathcal{C}(X)) \leq \hat{Q}(Y \in \mathcal{C}(X)) + \Delta_\delta(\mathbf{Z}, n) := \tilde{Q}(Y \in \mathcal{C}(X)),$$

$$Q(Y \notin \mathcal{C}(X)) \leq \hat{Q}(Y \notin \mathcal{C}(X)) + \Delta_\delta(\mathbf{Z}, n) := \tilde{Q}(Y \notin \mathcal{C}(X)),$$

with $V_n(\boldsymbol{Z})$ the empirical variance of $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, $Z_i = Q(y_i \in \mathcal{C}(x_i))$. Using these bounds, we get the following inequality with probability $1-\delta$:

$$H(Y|X) \leq h_b(\alpha) + (1-\alpha)\log \tilde{Q}(Y \in \mathcal{C}(X)) + \alpha_n \log \tilde{Q}(Y \notin \mathcal{C}(X)) - \mathbb{E}_P\left[\log Q_{Y|X}\right]. \quad (4)$$

This upper bound is one of our main results, which we dub the *DPI bound*. Note that for the last expectation, we can use the empirical estimate, as it is an unbiased approximation.

3.2 Model-Based Fano's Inequality and Variations

Next, we present an inequality which is a variation of Fano's inequality [19], a classical result that, among other things, is used to prove Shannon's classical theorem on channel capacity. See Appendix C.4 for more details. In our context, we can use Fano's inequality to relate the conditional entropy $H(Y|X)$ to the probability of error, i.e., $\mathbb{P}(Y \notin \mathcal{C}(X))$. From that insight, we obtain Proposition 3.2 by modifying the classical proof of Fano's inequality, which can be found in [14], and applying the conformal guarantees from Theorem 2.1 to bound the probability of error.

Proposition 3.2 (Model-Based Fano Bound)**.** Consider any conformal prediction method satisfying the upper and lower bounds of Theorem 2.1 for $\alpha \in (0, 0.5)$. Then, for the true distribution P, and for any probability distribution Q, we have

$$H(Y|X) \leq h_b(\alpha) + \alpha \mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\notin\mathcal{C}(X)}}\left[-\log Q_{Y|X,\mathcal{C}(X),Y\notin\mathcal{C}(X)}\right] + (1-\alpha_n)\,\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\in\mathcal{C}(X)}}\left[-\log Q_{Y|X,\mathcal{C}(X),Y\in\mathcal{C}(X)}\right]. \quad (5)$$

Note that we have one term conditioned on the event of valid coverage, $\{Y \in \mathcal{C}(X)\}$, and another conditioned on $\{Y \notin \mathcal{C}(X)\}$. We provide the proof in Appendix D.2. A good choice for Q is the predictive model itself, and that is why we refer to the bound above as *Model-Based (MB) Fano bound*. Another natural choice for Q is the uniform distribution, which gives us the following result.

Corollary 3.1 (Simple Fano Bound)**.** Consider any conformal prediction method satisfying the upper and lower bounds of Theorem 2.1 for $\alpha \in (0, 0.5)$. Then, for the true distribution P we have

$$H(Y|X) \leq h_b(\alpha) + \alpha \mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\notin\mathcal{C}(X)}}\left[\log(|\mathcal{Y}| - |C(X)|)\right] + (1-\alpha_n)\,\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\in\mathcal{C}(X)}}\left[\log|C(X)|\right]. \quad (6)$$

The proof follows directly from Proposition 3.2 by replacing Q with the uniform distribution. We refer to the bound in (6) as (simple) *Fano bound*, since it is model agnostic and can be approximated directly with only empirical estimates of the prediction set size. This last bound explicitly relates the central notion of uncertainty in conformal prediction, the prediction set size, to an information-theoretic concept of uncertainty in $H(Y|X)$. This reinterpretation of conformal prediction as a form of list decoding introduces various information-theoretic tools, potentially useful for various applications. In Appendix E we derive some new inequalities for conformal prediction, in particular offering new lower bounds on the inefficiency of the conformal prediction. In the next section, we discuss how these inequalities can be used as conformal training schemes.

4 Conformal Training

Although split conformal prediction is applicable to any pretrained ML model as a post-processing step, the overall performance of any CP method (commonly measured by its inefficiency) is highly dependent on the underlying model itself. Therefore, previous works have proposed to take CP into account already during model training and directly optimize for low predictive inefficiency [8, 13, 61]. We use the term *conformal training* to refer to such approaches in general (see Appendix F for an overview of the topic). In particular, we focus on ConfTr [61] since it generalizes and outperforms [8]. In ConfTr [61] each training batch $\mathcal{B}$ is split into calibration $\mathcal{B}_{cal}$ and test $\mathcal{B}_{test}$ halves to simulate the SCP process (see Section 2) for each gradient update of model f and minimize the following *size loss*

$$\log \mathbb{E}[|\mathcal{C}_f(X)|] \approx \log \left({}^{1}/_{|\mathcal{B}_{test}|} \sum_{x \in \mathcal{B}_{test}} |\mathcal{C}_f(x)| \right), \tag{7}$$

where $\mathcal{C}_f(x)$ is constructed using the statistics of the nonconformity scores computed on $\mathcal{B}_{cal}$. We use the notation $\mathcal{C}_f(x)$ to emphasize the dependence of the prediction set on model f. Still, SCP involves step functions and Stutz et al. [61] introduce two relaxations to recover a differentiable objective: (i) the computation of quantiles is relaxed via differentiable sorting operators [9, 17, 50]; (ii) the thresholding operation in the construction of prediction sets in (2) is replaced by smooth assignments via the logistic sigmoid. The latter relaxation gives "soft" prediction sets $\hat{\mathcal{C}}_f(x)$, which contain each of the labels $y \in \mathcal{Y}$ with a certain probability. See Algorithm 1 for a depiction of conformal training.

Our upper bounds on $H(Y|X)$, namely DPI, MB Fano and simple Fano, presented in the previous section can be made differentiable in the same way, and thus can also serve as proper loss functions for conformal training. The motivation for doing so is twofold. First, the conditional entropy $H(Y|X)$ captures the underlying uncertainty of the task, or equivalently, the uncertainty under the true labelling distribution $P_{Y|X}$. Thus, by minimizing these upper bounds, we can hope to push the model f closer to the true distribution, which is known to achieve minimal inefficiency [65]. Interestingly, the cross-entropy loss also bounds $H(Y|X)$ from above and thus can be motivated as a conformal training objective from the same angle. In that regard, the DPI bound from Proposition 3.1 is particularly advantageous as it is provably tighter than the cross-entropy (see Appendix D.1).

Second, we can connect the simple Fano bound from Corollary 3.1 to the size loss (7) from [61]. In Appendix F.1, we show that via Jensen's inequality and $\log(|Y| - |C(X)|) \leq \log |Y|$ the bound in (6) can be further upper bounded as

$$\begin{aligned} &\lambda_\alpha := h_b(\alpha) + \alpha \log |\mathcal{Y}| - (1 - \alpha_n) \log(1 - \alpha), \\ &H(Y|X) \leq \lambda_\alpha + (1 - \alpha_n) \log \mathbb{E}\left[|C(X)|\right], \end{aligned} \tag{8}$$

Since λ_α and $(1 - \alpha_n)$ are constants, they do not affect optimization, and minimizing the right hand side in (8) is equivalent to minimizing the size loss in (7). Therefore, we ground ConfTr as minimizing an upper bound to the true conditional entropy that is looser than the simple Fano bound and likely also looser than the model-based Fano bound for an appropriate choice for Q.

Algorithm 1: Conformal training algorithm.

input : A batch of labeled samples $\mathcal{B}$, a model f, a loss function $\mathcal{L}$ that can be any of our upper bounds (4), (5) or (6) or a version of (7)—see variants (25) and (26) in Appendix F.

for each training batch $\mathcal{B}$ **do**
- split $\mathcal{B}$ into $\mathcal{B}_{cal}$ and $\mathcal{B}_{test}$
- **for** each $(x_i, y_i) \in \mathcal{B}_{cal}$ **do**
  - compute score $s_i = s_f(x_i, y_i)$
- sort scores (in a differentiable manner) obtaining $s_{(1)} < s_{(2)} < \ldots < s_{(|\mathcal{B}_{cal}|)}$
- set $\hat{q} = s_{(\lceil (|\mathcal{B}_{cal}|+1)(1-\alpha) \rceil)}$ `// get` $1-\alpha$ `quantile estimate using` $\mathcal{B}_{cal}$
- **for** each $(x_j, y_j) \in \mathcal{B}_{test}$ **do**
  - **for** each $y \in \mathcal{Y}$ **do**
    - $\hat{\mathcal{C}}_f(x_j, y) \leftarrow \sigma(\hat{q} - s_f(x_j, y))$ `// Construct soft prediction set`
- `// Bounds (4) and (5) also require class probabilities under` Q`, which in this case are given by the model` $f(x_j)$
- compute loss according to $\mathcal{L}$ on $\mathcal{B}_{test}$ using $y_j, \hat{\mathcal{C}}_f(x_j)$ and if needed $f(x_j)$
- update f via gradient descent to minimize loss

5 Side Information

With the information theoretical interpretation of conformal prediction, we can translate various intuitions from information theory, for example, about different types of channels or network information theory, to conformal prediction. In this section, we consider the notion of side information.

Let X be the input covariates, Y be the target variable, Z be some side information about the task, and let $Q_{Y|X}$ be the model which we use to perform conformal prediction. As we can relate CP with $Q_{Y|X}$ to an upper bound on the conditional entropy $H(Y|X)$, we would like to do the same for the case of the conditional entropy when side information is available, i.e., $H(Y|X, Z)$. Since we know that the conditional entropy directly affects the expected set size, i.e., the inefficiency of CP, and given that $H(Y|X) \geq H(Y|X, Z)$ we can expect that with the additional side information the inefficiency of CP will decrease. We can take side information into account by defining conformity scores as a function of $Q_{Y|X,Z}$ instead of $Q_{Y|X}$. A simple way to do that would be via the Bayes rule

$$Q_{Y|X,Z} = \frac{Q_{Y|X}Q_{Z|X,Y}}{\sum_Y Q_{Y|X}Q_{Z|X,Y}}, \tag{9}$$

where $Q_{Z|X,Y}$ is an auxiliary model that predicts the side information given the input and target variables. Such a model could be learned separately from the main model $Q_{Y|X}$ given access to a dataset $\mathcal{D}_{side} = \{(x_i, y_i, z_i)\}$. Therefore, we can now calibrate with CP by taking into account the side information and then, at test time, given access to the input and side information, we can use the appropriate probabilities $Q_{Y|X,Z}$ to construct the prediction sets. Intuitively, the prediction sets from such a procedure should be smaller than the prediction sets obtained from using $Q_{Y|X}$ directly. It should be noted that, in the case of side information not being available, we can marginalize $Q_{Y,Z|X}$ over Z, which, by construction, falls back to the original model $Q_{Y|X}$. If the availability pattern of Z is consistent between the calibration and test sets, the conformal prediction guarantees still hold by defining the model as

$$f = \begin{cases} Q_{Y|X,Z} & \text{if } Z \text{ is observed} \\ Q_{Y|X} & \text{otherwise.} \end{cases} \tag{10}$$

With this addition to the split conformal prediction tool set, if new (side) information is made available at test time, we can properly incorporate it into the CP pipeline without having to retrain the main classifier. One needs only train an auxiliary classifier $Q_{Z|X,Y}$, which might be much simpler than $Q_{Y|X}$ (in our experiments, $Q_{Z|X,Y}$ is given by a single linear layer) and require much less data. One notable example of side information arises naturally in the distributed setting, which we discuss next.

5.1 The Distributed Learning Setting

Consider the case where we have a dataset that is distributed among a set of m devices and want to run conformal training to get a global model $Q_{Y|X}$ trained on all the data. Further, assume it is hard to gather the data at a central location (e.g., due to privacy reasons), and thus we have to work with the data staying locally on each device. An example of this would be federated learning (or FL) [42]. In this case, if $Z \in \{1, \ldots, m\}$ identifies the device, the entropy $H(Y|X)$ can be expressed as

$$H(Y|X) = H(Y|X, Z) + I(Y; Z|X) = \mathbb{E}_{P_Z}\left[H(Y|X, Z = z)\right] + I(Y; Z|X),$$

which decomposes into a weighted average of local entropy functions $H(Y|X, Z = z)$. We can now use any of our proposed bounds for each of the conditional entropies $H(Y|X, Z = z)$ by calibrating with CP independently on each device, ending up with

$$H(Y|X) \leq \mathbb{E}_{P_Z}\left[H_{ub}(Y|X, Z = z)\right] + I(Y; Z|X),$$

where $H_{ub}(Y|X, Z = z)$ corresponds to an upper bound of the conditional entropy $H(Y|X, Z = z)$. Furthermore, for the mutual information term we have that

$$I(Y; Z|X) = \mathbb{E}_{P_{Z,X,Y}}\left[\log \frac{P_{Z|X,Y}}{P_{Z|X}}\right] \leq \mathbb{E}_{P_{Z,X}}\left[-\log P_{Z|X}\right] \leq \mathbb{E}_{P_{Z,X}}\left[-\log Q_{Z|X}\right]$$

where the first inequality is due to Z being discrete and having non-negative entropy and the second is due to Gibbs inequality with $Q_{Z|X}$ being an auxiliary model trained to predict the user id $Z = z$

given input $X = x$. A similar upper bound has been considered before in a federated setting [36]. With this upper bound we get

$$H(Y|X) \leq \mathbb{E}_{P_Z}\Big[H_{ub}(Y|X, Z = z) - \mathbb{E}_{P_{X|Z=z}}\left[\log Q_{Z=z|X}\right]\Big]. \tag{11}$$

This gives us an upper bound on the entropy of the entire population that decomposes into a sum of local functions, one for each client, only requiring local information. Thus, we can easily carry out conformal training for $Q_{Y|X}$ by minimizing this upper bound in the federated setting with, e.g., federated averaging [42]. At test time, we can take the device ID z as side information. To this end, we can either train a model $Q_{Z|X,Y}$ in parallel and use (9) with the global model $Q_{Y|X}$ at test time to get $Q_{Y|X,Z}$, or we can obtain $Q_{Y|X,Z}$ by fine-tuning the global model $Q_{Y|X}$ with local data.

6 Related Work

Conformal prediction, a powerful framework for uncertainty quantification developed by Vovk and collaborators [60, 64], has recently witnessed a wide adoption in many fields, e.g., healthcare [3, 38, 39, 48] and finance [6, 67]. The marriage of conformal prediction and machine learning has been especially fruitful. Since the seminal work by Vovk et al. [64], many extensions and applications have been proposed, covering topics such as survival analysis [10], treatment effect evaluation [32], classification [4, 21] and regression [54] settings, risk control [5, 7], and covariate shift [62].

To our knowledge, our work represents the first attempt to bridge conformal prediction and information theory. Among other things, this allows us to build on the conformal training ideas of Bellotti [8] and Stutz et al. [61], deriving principled learning objectives that generalize their approaches, dispense with some of their hyperparameters and result in more efficient prediction sets. Further, we empirically show that our conformal training objectives provide a strong enough learning signal to train complex architectures from scratch, with strong results on ResNet-34 and ResNet-50 [23] fitted on CIFAR10 and CIFAR100, respectively. In contrast, the previous state-of-the-art approach, ConfTr, struggles in those settings (see experiments in Section 7) and required pretrained models for consistent results [61]. Further, our information-theoretic interpretation of CP provides a new simple and effective mechanism to leverage side information in split conformal prediction. We are unaware of any other approaches to treat side information within the conformal prediction framework in the literature.

On the information theory side, the notion of f-divergence and related inequalities have appeared in many different works. The use of f-divergence goes back to works of Ali and Silvey, Csiszár, and Morimoto in the 60s, as in, for instance, [2, 15, 44]. A key f-divergence inequality is the data processing inequality—see [57, 58] for an extensive survey. It provides a unified way of obtaining many classical and new results, including Fano's inequality [19]. The tightness of the data processing inequalities is discussed in terms of Bregman's divergence in [12, 34] and in terms of χ^2-divergence in [57]. List decoding, which is closely connected to CP, was introduced in the context of communication design by Elias [18] and Wozencraft [68]. A generalization of Fano's inequality to list decoding was given in [16] in the context of multi-user information theory, see also the general Fano inequality for list decoding presented in [53]. Variable-size list decoding was discussed in [57] using ideas first introduced in [52] and [35]. A selection of relevant inequalities for list decoding can be found in [57].

7 Experiments

In this section, we empirically study two applications of our theoretical results, namely conformal prediction with side information and conformal training with our upper bounds on the conditional entropy as optimization objectives. We focus our experiments on classification tasks since this is the most common setting in previous works in conformal training [8, 13, 61].

7.1 Conformal Training

We test the effectiveness of our upper bounds as objectives for conformal training in five data sets: MNIST [29], Fashion-MNIST [69], EMNIST [11], CIFAR10 and CIFAR100 [25]. In addition to our three upper bounds, we also evaluate the cross-entropy loss (CE, also another upper bound to the entropy), and the two main variants proposed in [61], namely ConfTr, which minimizes (7) and $\text{ConfTr}_{\text{class}}$ that optimizes an additional classification loss term (see Appendix F). We follow a similar

Table 1: **Inefficiency results for conformal training in the centralized setting.** We report the mean prediction set size ($\pm$ standard deviation) across 10 different calib./test splits for $\alpha = 0.01$, showing in bold all values within one std. of the best result. Results for THR and APS correspond to different models trained with different hyperparameters (see Appendix G). Lower is better.

Method	MNIST		F-MNIST		EMNIST		CIFAR 10		CIFAR 100	
	THR	APS	THR	APS	THR	APS	THR	APS	THR	APS
CE	$\mathbf{2.29}_{\pm 0.18}$	$2.50_{\pm 0.08}$	$2.39_{\pm 0.13}$	$2.41_{\pm 0.17}$	$\mathbf{2.06}_{\pm 0.11}$	$3.40_{\pm 0.18}$	$\mathbf{1.69}_{\pm 0.11}$	$2.34_{\pm 0.22}$	$19.70_{\pm 2.05}$	$26.02_{\pm 1.31}$
ConfTr	$6.28_{\pm 0.71}$	$\mathbf{2.10}_{\pm 0.07}$	$\mathbf{1.73}_{\pm 0.06}$	$\mathbf{1.89}_{\pm 0.09}$	$\mathbf{1.99}_{\pm 0.10}$	$\mathbf{2.36}_{\pm 0.11}$	$9.90_{\pm 0.02}$	$9.98_{\pm 0.00}$	$32.80_{\pm 2.75}$	$40.58_{\pm 1.23}$
$\text{ConfTr}_{\text{class}}$	$\mathbf{2.09}_{\pm 0.11}$	$\mathbf{2.13}_{\pm 0.13}$	$5.11_{\pm 0.49}$	$\mathbf{1.79}_{\pm 0.07}$	$\mathbf{2.01}_{\pm 0.09}$	$\mathbf{2.38}_{\pm 0.11}$	$2.16_{\pm 0.09}$	$2.18_{\pm 0.06}$	$66.48_{\pm 3.67}$	$32.91_{\pm 1.53}$
Fano	$\mathbf{2.09}_{\pm 0.12}$	$\mathbf{2.12}_{\pm 0.08}$	$\mathbf{1.70}_{\pm 0.05}$	$\mathbf{1.87}_{\pm 0.05}$	$\mathbf{2.10}_{\pm 0.11}$	$2.75_{\pm 0.14}$	$2.05_{\pm 0.05}$	$2.35_{\pm 0.10}$	$40.30_{\pm 1.10}$	$33.80_{\pm 0.93}$
MB Fano	$\mathbf{2.24}_{\pm 0.12}$	$2.49_{\pm 0.19}$	$\mathbf{1.80}_{\pm 0.08}$	$2.25_{\pm 0.14}$	$\mathbf{2.01}_{\pm 0.11}$	$3.67_{\pm 0.13}$	$\mathbf{1.66}_{\pm 0.09}$	$\mathbf{1.89}_{\pm 0.06}$	$\mathbf{14.61}_{\pm 0.84}$	$21.68_{\pm 1.44}$
DPI	$\mathbf{2.24}_{\pm 0.17}$	$2.64_{\pm 0.07}$	$\mathbf{1.73}_{\pm 0.07}$	$2.08_{\pm 0.06}$	$\mathbf{1.98}_{\pm 0.09}$	$4.07_{\pm 0.23}$	$\mathbf{1.64}_{\pm 0.07}$	$\mathbf{1.97}_{\pm 0.08}$	$17.55_{\pm 1.33}$	$\mathbf{17.41}_{\pm 0.62}$

optimization procedure and experimental setup to that of [61], but with the key differences that we learn the classifiers from scratch in all cases (without the need of pretrained CIFAR models), and that we use the larger "by class" split of EMNIST. For each data set, we use the default train and test splits but transfer 10% of the training data to the test data set. We train the classifiers only on the remaining 90% of the training data and, at test time, run SCP with 10 different calibration/test splits by randomly splitting the enlarged test data set. See Appendix G for a complete description of the experimental setup, with extra results and details on model architectures and hyperparameter search.

In Table 1, we report the empirical inefficiency on test data considering two SCP methods, threshold CP with probabilities (or THR) [56] and APS [55]—see Appendix G.1.1 for results with RAPS [4]. In all cases, our upper bounds proved effective loss functions to train efficient classifiers end-to-end and from scratch. For the simpler data sets (MNIST, Fashion-MNIST and EMNIST), all conformal training methods achieved similar results, but both ConfTr methods proved less consistent. This is noticeable in the oftentimes sharp difference in performance between THR and APS, since even after fine-tuning the hyperparameters (see Appendix G) some of the models failed to converge properly. For the remaining and more challenging data sets, both ConfTr variants lagged behind, probably because they do not provide a strong enough signal to train ResNets from scratch (on CIFAR data sets, Stutz et al. [61] only used ConfTr to fine tune pretrained models). A similar observation applies to the simple Fano bound (6), where the relaxed prediction set size is the only learning signal.

In all experiments, we run conformal training with a target coverage rate of 99%, i.e., $\alpha = 0.01$. It is then important to assess whether the performance of the resulting models deteriorates at different coverage rates, "overfitting" to the value of α used for training. In Table 2, we see how inefficiency varies with α at test time for models trained via conformal training with $\alpha = 0.01$. In particular, we can contrast their performance against that of models trained via the CE loss, which is agnostic to the desired coverage rate. In all cases, our model-based Fano and DPI bound performs best with the THR and APS methods, respectively, proving conformal training is worthwhile even if the desired coverage rate might vary at test time. Still, as noticed in [61], there is a drop in performance in comparison to the CE loss for higher values of α at test time. This could be due to some degree of overfitting, but it could also be attributed to the conformal prediction problem becoming easier for lower coverage rates, thus reducing the gap between our bounds and the CE loss.

Table 2: **Inefficiency results with varying α at test time**. Average prediction set size on CIFAR100 for different α targets at test time, averaged across 10 random calib./test splits. All methods were only optimized for α=0.01. The models used for THR and APS might not be the same according to the best hyperparameters found in Table 11. Lower is better.

Method	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	THR	APS	THR	APS	THR	APS
CE	$19.70_{\pm 2.05}$	$26.02_{\pm 1.31}$	$6.11_{\pm 0.34}$	$9.19_{\pm 0.34}$	$3.02_{\pm 0.10}$	$4.52_{\pm 0.12}$
ConfTr	$32.80_{\pm 2.75}$	$40.58_{\pm 1.23}$	$12.25_{\pm 0.47}$	$21.60_{\pm 0.78}$	$7.13_{\pm 0.23}$	$14.58_{\pm 0.47}$
$\text{ConfTr}_{\text{class}}$	$66.48_{\pm 3.67}$	$32.91_{\pm 1.53}$	$14.18_{\pm 0.60}$	$16.80_{\pm 0.60}$	$8.90_{\pm 0.40}$	$11.29_{\pm 0.42}$
Fano	$40.30_{\pm 1.10}$	$33.80_{\pm 0.93}$	$19.43_{\pm 0.80}$	$16.17_{\pm 0.49}$	$11.46_{\pm 0.58}$	$9.72_{\pm 0.25}$
MB Fano	$\mathbf{14.61}_{\pm 0.84}$	$21.68_{\pm 1.44}$	$\mathbf{5.24}_{\pm 0.13}$	$9.25_{\pm 0.30}$	$\mathbf{2.88}_{\pm 0.05}$	$5.51_{\pm 0.14}$
DPI	$17.55_{\pm 1.33}$	$\mathbf{17.41}_{\pm 0.62}$	$6.26_{\pm 0.20}$	$\mathbf{6.98}_{\pm 0.32}$	$3.33_{\pm 0.11}$	$\mathbf{4.08}_{\pm 0.14}$

7.2 Side Information

As a first experiment, we consider datasets for which a natural grouping of the labels exists and use the group assignment as side information Z. In CIFAR100, there is a disjoint partition of the 100 classes into 20 superclasses, so we define z as the superclass to which example (x, y) belongs. In EMNIST, z indicates whether the example is a digit, uppercase or lowercase letter. We train an auxiliary model $R_{Z|X,Y}$ and, at test time, assume access to side information z to recompute class probabilities $Q_{Y|X,Z=z}$ from the original classifier $Q_{Y|X}$ as in equation (9).

We report results for two different scenarios in Table 3. The first is the standard SCP setting, where we assess the inefficiency of THR and APS methods, with side information Z observed for 10, 30 and 100% of the instances. We redefine the classifier f as in (10) to account for when Z is missing, but otherwise, the SCP process remains unchanged. The second scenario is Mondrian or group-balanced CP [64], where one splits $\mathcal{D}_{cal}$ into groups and runs CP for each of them individually. In this setting, we group the calibration data points according to Z and base the score function on $Q_{Y|X,Z}$. In all cases, taking the side information into account reduced the inefficiency considerably.

Table 3: **Inefficiency results with side information.** We report the mean prediction set size ($\pm$ std.) across 10 different calib./test splits for $\alpha = 0.01$. The side information is the superclass assignment for CIFAR100 and whether the class is a digit / uppercase letter / lowercase letter for EMNIST.

Method	CIFAR 100			EMNIST		
	THR	APS	Acc.(%)	THR	APS	Acc.(%)
CP	$19.70_{\pm 2.05}$	$26.02_{\pm 1.31}$	72.22	$2.06_{\pm 0.11}$	$3.37_{\pm 0.15}$	85.74
CP w/ 10% SI	$18.13_{\pm 2.63}$	$23.59_{\pm 2.08}$	72.84	$1.91_{\pm 0.07}$	$2.18_{\pm 0.09}$	86.93
CP w/ 30% SI	$15.74_{\pm 1.11}$	$21.63_{\pm 1.45}$	74.83	$1.69_{\pm 0.05}$	$1.88_{\pm 0.07}$	89.43
CP w/ 100% SI	$10.28_{\pm 0.86}$	$15.65_{\pm 1.17}$	78.72	$1.06_{\pm 0.02}$	$1.07_{\pm 0.02}$	97.65
Group CP	$17.59_{\pm 1.89}$	$21.92_{\pm 1.80}$	72.22	$2.32_{\pm 0.14}$	$2.68_{\pm 0.11}$	85.74
Group CP w/ 100% SI	$9.07_{\pm 0.60}$	$13.16_{\pm 0.68}$	78.72	$1.14_{\pm 0.03}$	$1.16_{\pm 0.04}$	97.65

7.3 Federated Learning (FL)

A practically relevant application of side information arises in FL, where we can take the device ID as side information Z. In the federated setting, we train two extra heads on top of the main classifier, one computing $Q_{Z|X}$ so that we can optimize the proper upper bound in (11), and another computing $Q_{Z|X,Y}$ (while detaching gradients to the main classifier so as to not affect the upper bound optimization) that we use to integrate side information into CP using (9). Besides being a practically relevant application of side information to CP, FL also serves as a more challenging test bed for our conformal training methods, which has not been explored in previous work. We ran federated averaging [42] with CE, ConfTr, $\text{ConfTr}_{\text{class}}$, and our upper bounds as local loss functions. In this setting, we consider CIFAR10, CIFAR100, and EMNIST with 100, 500, and 1K devices, resp. We assign data points to devices imposing a *distribution-based label imbalance* [33], i.e., we sample a marginal label distribution for each device from a Dirichlet distribution $\text{Dir}(1.0)$. See Appendix G for results with $\text{Dir}(0.5)$ and $\text{Dir}(0.1)$. As hyperparameter search in FL is notably challenging and costly [66], we keep the same hyperparameters found for the centralized case in Section 7.1.

After convergence, we ran SCP with the final global model assuming calibration and test data sets at the server, or equivalently that the clients share their scores with the server. This reflects the best inefficiency results we can hope for with the global model, as in practice we might need to resort to privacy-preserving methods that are likely to hurt performance. See Appendix G for a discussion and extra results on other possible settings. We report inefficiency results for the global model with THR in Table 4 (see Table 5 in the appendix for APS results), where we observe similar trends to the centralized experiments in Table 1. ConfTr methods still perform well on EMNIST but struggle on CIFAR data (with the notable exception on CIFAR100, where ConfTr excelled) while our methods delivered consistent results across all data sets. This probably reflects the sensitivity of both ConfTr objectives to hyperparameters, which makes them hard to use in practice, especially in FL where hyperparameter optimization is difficult. Conversely, our bounds seem more robust to such variations, as the hyperparameters found in the centralized setting seem to translate well to the federated case.

Table 4: **Inefficiency results for conformal training in the federated setting with THR.** We report the mean prediction set size ($\pm$ standard deviation) of the global federated model across 10 different calib./test splits for $\alpha = 0.01$ and using THR. We use $_{+\text{SI}}$ to indicate the inclusion of side information. We show in bold all values within one standard deviation of the best result. Lower is better.

Method	EMNIST		CIFAR 10		CIFAR 100	
	THR	THR$_{+\text{SI}}$	THR	THR$_{+\text{SI}}$	THR	THR$_{+\text{SI}}$
CE	$2.91_{\pm 0.02}$	$2.46_{\pm 0.02}$	$2.73_{\pm 0.04}$	$2.30_{\pm 0.06}$	$55.41_{\pm 1.09}$	$52.31_{\pm 1.03}$
ConfTr	$4.60_{\pm 0.05}$	$3.30_{\pm 0.02}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$\mathbf{45.60_{\pm 1.30}}$	$\mathbf{41.18_{\pm 1.16}}$
ConfTr$_{\text{class}}$	$2.88_{\pm 0.02}$	$\mathbf{1.98_{\pm 0.02}}$	$3.53_{\pm 0.09}$	$3.39_{\pm 0.08}$	$58.53_{\pm 1.40}$	$56.03_{\pm 1.29}$
Fano	$\mathbf{2.63_{\pm 0.02}}$	$2.37_{\pm 0.02}$	$\mathbf{2.39_{\pm 0.07}}$	$\mathbf{2.07_{\pm 0.07}}$	$\mathbf{47.91_{\pm 1.20}}$	$\mathbf{41.19_{\pm 1.02}}$
MB Fano	$2.84_{\pm 0.04}$	$2.25_{\pm 0.03}$	$\mathbf{2.52_{\pm 0.08}}$	$\mathbf{2.04_{\pm 0.07}}$	$52.94_{\pm 1.40}$	$46.97_{\pm 1.30}$
DPI	$\mathbf{2.60_{\pm 0.02}}$	$2.23_{\pm 0.01}$	$2.76_{\pm 0.07}$	$2.28_{\pm 0.03}$	$52.36_{\pm 0.95}$	$48.64_{\pm 0.70}$

Table 5: **Inefficiency results for conformal training in the federated setting with APS.** We report the mean prediction set size ($\pm$ standard deviation) of the global federated model across 10 different calib./test splits for $\alpha = 0.01$ and using APS. We use $_{+\text{SI}}$ to indicate the inclusion of side information. We show in bold all values within one standard deviation of the best result. Lower is better.

Method	EMNIST		CIFAR 10		CIFAR 100	
	APS	APS$_{+\text{SI}}$	APS	APS$_{+\text{SI}}$	APS	APS$_{+\text{SI}}$
CE	$3.69_{\pm 0.03}$	$3.14_{\pm 0.04}$	$\mathbf{2.83_{\pm 0.07}}$	$2.43_{\pm 0.06}$	$64.73_{\pm 0.34}$	$62.67_{\pm 3.68}$
ConfTr	$6.14_{\pm 0.04}$	$5.25_{\pm 0.04}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$55.18_{\pm 2.10}$	$47.58_{\pm 1.48}$
ConfTr$_{\text{class}}$	$\mathbf{2.65_{\pm 0.02}}$	$\mathbf{2.42_{\pm 0.02}}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$99.92_{\pm 0.02}$	$99.91_{\pm 0.01}$
Fano	$3.12_{\pm 0.04}$	$2.72_{\pm 0.03}$	$\mathbf{2.73_{\pm 0.07}}$	$\mathbf{2.39_{\pm 0.06}}$	$\mathbf{46.95_{\pm 0.67}}$	$\mathbf{42.75_{\pm 0.91}}$
MB Fano	$4.75_{\pm 0.03}$	$\mathbf{2.43_{\pm 0.01}}$	$\mathbf{2.79_{\pm 0.13}}$	$\mathbf{2.33_{\pm 0.05}}$	$50.72_{\pm 1.77}$	$45.72_{\pm 1.38}$
DPI	$2.98_{\pm 0.03}$	$2.58_{\pm 0.02}$	$\mathbf{2.68_{\pm 0.15}}$	$\mathbf{2.22_{\pm 0.09}}$	$51.29_{\pm 1.07}$	$47.18_{\pm 1.27}$

One marked difference between Tables 1 and 4 is that the simple Fano bound (6), which lagged behind the DPI bound and its model-based counterpart in the centralized setting, achieved the best results on the federated setting. We hypothesize this could be due to overfitting of the local optimization procedures to the individual data distribution of each device, which hurts the convergence of the global model. This effect is exacerbated on CIFAR100, where we have 500 devices, each of which with very few data points. The simple Fano bound is less vulnerable to such overfitting since it relies on the main classifier to a much lesser degree. Finally, in almost all cases, our bounds outperformed the CE loss, reassuring the potential of conformal training. Moreover, the inclusion of side information reduced inefficiency in all settings, and markedly so in a few instances. This confirms the effectiveness of our side information approach in a complex and practically relevant scenario, like federated learning.

8 Conclusion

In this work, we established a link between notions of uncertainty coming from conformal prediction and information theory (or more precisely variable-size list decoding). We proved that one can use split conformal prediction methods to upper bound the conditional entropy of the target variable given the inputs, and that these upper bounds form principled objectives for conformal training. We empirically validated our approach to conformal training, with strong results in both centralized and federated settings. Furthermore, the information-theoretic perspective also offers a simple yet rigorous approach to incorporate side information into conformal prediction, which we experimentally show leads to better predictive efficiency. To the best of our knowledge, this is the first attempt at connecting information theory and conformal prediction. Given the limited communication between these two research communities thus far, we expect our work to incite a fruitful exchange of not only ideas but also theory and algorithms between these two research domains. In this paper, we concentrated our exposition and experiments in classification tasks, but we see an extension of our methods to the regression setting, as a particularly promising avenue for future work.

References

[1] Ahsan, M. M., Luna, S. A., and Siddique, Z. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, 10(3), 2022. ISSN 2227-9032. doi: 10.3390/healthcare10030541. URL https://www.mdpi.com/2227-9032/10/3/541. 1

[2] Ali, S. M. and Silvey, S. D. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1): 131–142, 1966. Publisher: [Royal Statistical Society, Wiley]. 7, 21

[3] Alnemer, L. M., Rajab, L., and Aljarah, I. Conformal prediction technique to predict breast cancer survivability. *Int J Adv Sci Technol*, 96:1–10, 2016. 7

[4] Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020. 7, 8, 32

[5] Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. 7

[6] Bastos, J. A. Conformal prediction of option prices. *Expert Systems with Applications*, 245: 123087, 2024. 7

[7] Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021. Publisher: ACM New York, NY. 7

[8] Bellotti, A. Optimized conformal classification using gradient descent approximation. *arXiv preprint arXiv:2105.11255*, 2021. 5, 7, 15, 28

[9] Blondel, M., Teboul, O., Berthet, Q., and Djolonga, J. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pp. 950–959. PMLR, 2020. 5, 28, 29

[10] Candès, E. J., Lei, L., and Ren, Z. Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*, 2021. 7

[11] Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017. 7, 30

[12] Collet, J.-F. An Exact Expression for the Gap in the Data Processing Inequality for f - Divergences. *IEEE Transactions on Information Theory*, 65(7):4387–4391, July 2019. ISSN 1557-9654. 7, 21

[13] Colombo, N. and Vovk, V. Training conformal predictors. In *Conformal and Probabilistic Prediction and Applications*, pp. 55–64. PMLR, 2020. 5, 7, 15, 28

[14] Cover, T. M. and Thomas, J. A. *Elements of information theory*. Wiley-Interscience, Hoboken, N.J, 2nd ed edition, 2006. 4, 18

[15] Csiszár, I. On topological properties of f-divergence. *Studia Sci. Math. Hungar.*, 2:330–339, 1967. 7, 21

[16] Csiszár, I. and Körner, J. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, Cambridge ; New York, 2nd ed edition, 2011. 7, 21

[17] Cuturi, M., Teboul, O., and Vert, J.-P. Differentiable ranking and sorting using optimal transport. *Advances in neural information processing systems*, 32, 2019. 5, 28

[18] Elias, P. List Decoding for Noisy Channels. In *IRE WESCON Convention Record, 1957*, volume 2, pp. 94–104, 1957. 2, 3, 7, 19, 21

[19] Fano, R. M. *The transmission of information*, volume 65. Massachusetts Institute of Technology, 1949. 1, 4, 7, 21, 24

[20] Gallager, R. G. *Information theory and reliable communication*. Wiley, New York, NY, 1986. ISBN 978-0-471-29048-3. 21

[21] Gupta, C., Podkopaev, A., and Ramdas, A. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020. 7

[22] Guruswami, V. *List decoding of error-correcting codes: winning thesis of the 2002 ACM doctoral dissertation competition*, volume 3282. Springer Science & Business Media, 2004. 2, 21

[23] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 7

[24] Humbert, P., Le Bars, B., Bellet, A., and Arlot, S. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pp. 14153–14177. PMLR, 2023. 33

[25] Krizhevsky, A. et al. Learning multiple layers of features from tiny images, 2009. 7, 30

[26] Kuchibhotla, A. K. Exchangeability, conformal prediction, and rank tests. *arXiv preprint arXiv:2005.06095*, 2020. 15, 16

[27] Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 3

[28] Kuutti, S., Bowden, R., Jin, Y., Barber, P., and Fallah, S. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 22(2): 712–733, 2020. 1

[29] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 7, 30

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

[31] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. 2, 17

[32] Lei, L. and Candès, E. J. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021. Publisher: Wiley Online Library. 7

[33] Li, Q., Diao, Y., Chen, Q., and He, B. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 965–978. IEEE, 2022. 9, 33

[34] Liese, F. and Vajda, I. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, October 2006. 7, 21

[35] Liu, J., Cuff, P., and Verdú, S. e_γ-Resolvability. *IEEE Transactions on Information Theory*, 63 (5):2629–2658, May 2017. Conference Name: IEEE Transactions on Information Theory. 7, 21

[36] Louizos, C., Reisser, M., and Korzhenkov, D. A mutual information perspective on federated contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JrmPG9ufKg. 7

[37] Lu, C. and Kalpathy-Cramer, J. Distribution-free federated learning with conformal predictions. *arXiv preprint arXiv:2110.07661*, 2021. 33

[38] Lu, C., Angelopoulos, A. N., and Pomerantz, S. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 545–554. Springer, 2022. 7

[39] Lu, C., Lemay, A., Chang, K., Höbel, K., and Kalpathy-Cramer, J. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12008–12016, 2022. 7

[40] maintainers, T. and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016. 30, 40

[41] Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009. 4, 22

[42] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017. 6, 7, 9, 32

[43] Merhav, N. List Decoding—Random Coding Exponents and Expurgated Exponents. *IEEE Transactions on Information Theory*, 60(11):6749–6759, November 2014. ISSN 1557-9654. Conference Name: IEEE Transactions on Information Theory. 20

[44] Morimoto, T. Markov Processes and the H-Theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, March 1963. Publisher: The Physical Society of Japan. 7, 21

[45] Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M. I., et al. Ray: A distributed framework for emerging {AI} applications. In *13th USENIX symposium on operating systems design and implementation (OSDI 18)*, pp. 561–577, 2018. 30

[46] Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6): 1191–1253, 2003. 35

[47] Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pp. 345–356. Springer, 2002. 1, 2, 17, 28

[48] Papadopoulos, H., Gammerman, A., and Vovk, V. Reliable diagnosis of acute abdominal pain with conformal prediction. *Engineering Intelligent Systems*, 17(2):127, 2009. 7

[49] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 30

[50] Petersen, F., Borgelt, C., Kuehne, H., and Deussen, O. Monotonic differentiable sorting networks. In *International Conference on Learning Representations (ICLR)*, 2022. 5, 15, 28, 29, 30

[51] Polyanskiy, Y. and Wu, Y. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014. 18, 19

[52] Polyanskiy, Y., Poor, H. V., and Verdu, S. Channel Coding Rate in the Finite Blocklength Regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, May 2010. 7, 21

[53] Raginsky, M. and Sason, I. *Concentration of measure inequalities in information theory, communications, and coding*. Number vol. 10, no. 1-2 in Foundations and Trends in Communications and Information Theory. Now Publ, Boston, Mass., 2013. ISBN 978-1-60198-724-2. 7, 20, 21

[54] Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019. 7

[55] Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020. 8, 20, 30

[56] Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. 8, 30

[57] Sason, I. On data-processing and majorization inequalities for f-divergences with applications. *Entropy*, 21(10):1022, 2019. Publisher: MDPI. 7, 19, 21

[58] Sason, I. and Verdú, S. f-Divergence Inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, November 2016. 3, 7, 18, 21

[59] Sason, I. and Verdú, S. Arimoto–Rényi conditional entropy and Bayesian M -ary hypothesis testing. *IEEE Transactions on Information theory*, 64(1):4–25, 2017. Publisher: IEEE. 21

[60] Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008. 7

[61] Stutz, D., Dvijotham, K. D., Cemgil, A. T., and Doucet, A. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. 5, 7, 8, 15, 28, 29, 30, 32

[62] Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019. 7

[63] Vovk, V. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012. 17

[64] Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005. 1, 2, 7, 9, 17

[65] Vovk, V., Fedorova, V., Nouretdinov, I., and Gammerman, A. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings 5*, pp. 23–39. Springer, 2016. 2, 5, 20

[66] Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021. 9

[67] Wisniewski, W., Lindsay, D., and Lindsay, S. Application of conformal prediction interval estimations to market makers' net positions. In *Conformal and Probabilistic Prediction and Applications*, pp. 285–301. PMLR, 2020. 7

[68] Wozencraft, J. M. List decoding. *Quarterly Progress Report*, 48:90–95, 1958. 2, 3, 7, 19, 21

[69] Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. This work is licensed under the MIT License. To view a copy of this license, visit https://github.com/zalandoresearch/fashion-mnist/blob/master/LICENSE. 7, 30

A Broader Impact, Limitations and Other Remarks

Broader Impact. This work explores the connection between conformal prediction and information theory, with the end goal of advancing the state of the art of uncertainty quantification in machine learning. With that in mind, we believe its potential societal consequences are chiefly positive, since our work might contribute to a larger adoption of uncertainty estimates, in a number of safety-critical applications. Notwithstanding, conformal prediction, like any other uncertainty estimation method, should be applied with care and a proper understanding of the provided guarantees so as not to create an illusion of safety when there is none. That is why our work aims to develop new methods and algorithms from first principles, with the hope of providing new techniques that we can study and understand at a deeper level so as to mitigate, or at least foresee, some of their failure modes.

Limitations. Our work is not without limitations. In practice, estimating the conditional entropy $H(Y|X)$ is a difficult problem. That is part of the reason upper bounds might be useful, but that also means it is hard to evaluate how tight our upper bounds are. This also limits potential applications of our upper bounds like, for instance, estimating the expected prediction set size. Our experiments in Appendix G.3 rely on quantization to get a reasonable lower bound on the conditional entropy, and thus lower bound the expected prediction set size. Finally, in terms of our experimental results, we unfortunately have not been able to single out which of our new upper bounds performs best. The general trend we observe is that the DPI and model-based Fano bounds perform better on complex classification tasks, while simple Fano seems to excel in relatively simple tasks or where there is a high risk of overfitting, like in the federated setting.

Regression Setting. In this paper, we focus our experiments on the classification setting, similarly to previous works on conformal training [8, 13, 61]. However, in principle, our bounds pose no assumptions on the underlying prediction problem and we see the application of our results to the regression setting as a promising avenue for future work. In practice, only the simple Fano bound in (6) would not be directly applicable to the regression setting, but that is mainly because it assumes a uniform distribution over the output space and we get the, potentially infinite, $|\mathcal{Y}|$ term. The other two bounds we propose, DPI in (4) and model-based Fano in (5), do not include the $|\mathcal{Y}|$ and $|C(X)|$ terms and can be applied to regression problems as is.

Computational Cost. The computational cost of using our bounds for conformal training is the same as that of previously proposed conformal training [61]. In comparison to regular training, i.e. minimizing the cross-entropy loss, the additional cost is given by the (differentiable) sorting operation of the scores, which in our case was performed with diffsort [50] with bitonic networks, which has complexity $\mathcal{O}(b^2 \log b^2)$ for b the batch size. Since the batch size is typically small, the additional computation cost is only marginal in practice.

B Background on Conformal Prediction

In this section, we outline a brief introduction to conformal prediction, providing the reader with the necessary background to follow our main results. Readers already familiar with conformal prediction can safely skip this section. We start by reviewing the definitions of quantiles and exchangeability, which are central to the main results in conformal prediction.

B.1 Exchangeability, Ranks and Quantiles

The main assumption in conformal prediction is that the data points used for calibration and testing are exchangeable. Next, we define the concept of exchangeability and discuss how it leads to the main results in conformal prediction via properties of ranks of exchangeable random variables. Our exposition is markedly brief, and we refer the reader to [26] for a more thorough discussion on exchangeability and its importance in conformal prediction.

Formally, the concept of exchangeability can be defined as follows.

Definition B.1 (Exchangeable Random Variables)**.** Random variables $X_1, \ldots, X_n$ are said to be exchangeable if for any permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$, the sequences $(X_1, \ldots, X_n)$ and $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ have the same joint probability distribution.

Note that exchangeability is a weaker assumption than the i.i.d. (independent and identically distributed) assumption commonly relied upon in machine learning. More precisely, exchangeable random variables must be identically distributed but not necessarily independent [26]. Naturally, i.i.d. random variables are also exchangeable.

One relevant consequence of exchangeability that is central to conformal prediction is that the ranks of exchangeable random variables are uniformly distributed. We define ranks and this property more formally in Definition B.2 and Lemma B.4, respectively.

Definition B.2 (Rank)**.** For a set of n elements $\mathcal{X} = \{x_1, \ldots, x_n\}$, the rank of any one element x_i in $\mathcal{X}$ is defined as

$$\text{rank}(x_i; \mathcal{X}) = |\{j \in \{1, \ldots, n\} : x_j + \xi U_j \leq x_i + \xi U_i\}|,$$

for $\xi \geq 0$ and $\mathcal{U} = \{U_1, \ldots, U_n\}$ a set of i.i.d. random variables uniformly distributed in $[-1, -1]$.

Remark B.3. The addition of i.i.d. uniform noise serves as a tie-breaking mechanism. Since $\{U_1, \ldots, U_n\}$ are almost surely distinct, $\{x_i + \xi U_i\}_{i=1}^n$ are also distinct with probability one. This is necessary to render the rank independent of the distribution of X_i, which is key to ensure the distribution-free quality of conformal prediction.

Lemma B.4. *If $(X_1, \ldots, X_n)$ are exchangeable random variables, then*

$$(\text{rank}(X_i; \{X_1, \ldots, X_n\}))_{i=1}^n \sim \text{Unif}\left(\{\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}\}\right).$$

In words, Lemma B.4 tell us that the ranking of exchangeable random variables is uniformly distributed among all possible permutations $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$. That means the probability of observing any one ranking is equal to $1/n!$ and, importantly, independent of the distribution of X. The corollary below follows directly from Lemma B.4.

Corollary B.1. If $(X_1, \ldots, X_n)$ are exchangeable random variables, then

$$\mathbb{P}\left(\text{rank}(X_i; \{X_1, \ldots, X_n\}) \leq t\right) = \frac{\lfloor t \rfloor}{n},$$

for $t \in [0, n]$ and $\lfloor t \rfloor$ the smallest integer smaller or equal to t. Moreover, we can define a valid p-value as $P := \text{rank}(X_i; \{X_1, \ldots, X_n\})/n$ since

$$\mathbb{P}\left(P \leq \alpha\right) \leq \alpha \quad \text{for all } \alpha \in [0, 1].$$

Proof.

$$\begin{aligned}
\mathbb{P}(\text{rank}(X_i; \{X_1, \ldots, X_n\}) \leq t) &= \mathbb{P}(\text{rank}(X_i; \{X_1, \ldots, X_n\}) \leq \lfloor t \rfloor) \\
&= \sum_{i=1}^{\lfloor t \rfloor} \mathbb{P}(\text{rank}(X_i; \{X_1, \ldots, X_n\}) = i) \\
&= \sum_{i=1}^{\lfloor t \rfloor} \frac{(n-1)!}{n!} = \frac{\lfloor t \rfloor}{n},
\end{aligned}$$

where the first equality follows because rank(.) returns an integer, and the third equality follows directly from Lemma B.4: each permutation of $(\text{rank}(X_i; \{X_1, \ldots, X_n\}))_{i=1}^n$ has the same probability $1/n!$, and there are $(n-1)!$ configurations where $\text{rank}(X_i; \{X_1, \ldots, X_n\}) = i$ since X_i is fixed at rank i, and we can permute the other $(n-1)$ variables. □

As we shall see, Corollary B.1 is central to the main result in conformal prediction, but before proving that result, we should first define the concept of quantiles.

Definition B.5 (Quantile)**.** For Z a random variable with probability distribution F, the level β quantile of distribution F is defined as

$$\text{Quantile}(\beta; F) = \inf\{z : \mathbb{P}\{Z \leq z\} \geq \beta\}.$$

Similarly, for a sample $\{z_i\}_{i=1}^n$ and δ_{z_i} a point mass concentrated at z_i, the quantile of the empirical distribution is given by

$$\text{Quantile}\left(\beta; \{z_i\}_{i \in [n]}\right) = \text{Quantile}\left(\beta; \frac{1}{n}\sum_{i=1}^n \delta_{z_i}\right).$$

B.2 Conformal Prediction

Armed with the definitions of quantiles and exchangeability, we are prepared to study conformal prediction, a distribution-free uncertainty quantification framework with the following goal: given a set of data points $\{(X_i, Y_i)\}_{i=1}^n$ sampled from some distribution P on $\mathcal{X} \times \mathcal{Y}$, to construct a set predictor $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ such that for a new data point (X_{test}, Y_{test}) and target error rate $\alpha \in (0, 1)$, we have the guarantee

$$\mathbb{P}(Y_{test} \in \mathcal{C}(X_{test})) \geq 1 - \alpha,$$

where the probability is over the randomness of $\{(X_i, Y_i)\}_{i=1}^n \cup \{(X_{test}, Y_{test})\}$. Since the probability is taken over both X_{test} and Y_{test}, this means that *on average* across all possible values in $\mathcal{X}$, the correct label Y_{test} is included in the constructed set $\mathcal{C}(X_{test})$ with probability at least $1 - \alpha$. This property is known as *marginal coverage*. This is in contrast to the stronger *conditional coverage*, which requires

$$\mathbb{P}(Y_{test} \in \mathcal{C}(X_{test})|X_{test}) \geq 1 - \alpha.$$

That is, the guarantee holds for each X_{test} individually, with X_{test} fixed and the probability taken over the randomness of Y_{test} only. However, distribution-free conditional coverage is known to admit no non-trivial solution (i.e., besides $\mathcal{C}(X_{test}) = \mathcal{Y}$) [30, 63]. Therefore, in this paper, whenever we refer to the lower and upper bounds provided by conformal prediction, we mean those provided by Theorem 2.1, which guarantees only marginal coverage.

There are different ways to achieve marginal coverage, but for simplicity we will focus on split conformal prediction or SCP [47], since it is easier to grasp than other variants of conformal prediction and is also the main object of study in this paper. We start by assuming a calibration dataset $\mathcal{D}_{cal}$ which consists of n i.i.d. samples (X_i, Y_i) drawn from an unknown distribution over $\mathcal{X} \times \mathcal{Y}$. We also assume access to a model $f : \mathcal{X} \to \hat{\mathcal{Y}}$, where the output space $\hat{\mathcal{Y}}$ can be different from $\mathcal{Y}$. Prediction sets satisfying the guarantee above can be constructed via the following three steps.

1. Define a *non-conformity* score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which assigns high scores to unusual pairs (x, y). The score function is typically a function of the model f itself.
2. Compute $S_i = s(X_i, Y_i)$ for all $(X_i, Y_i) \in \mathcal{D}_{cal}$ and compute Quantile$(1 - \alpha; \{S_i\}_{i=1}^n \cup \{\infty\})$, the empirical $1 - \alpha$ quantile of the scores in $\mathcal{D}_{cal}$.
3. Construct prediction sets $\mathcal{C}(X_{test}) = \left\{y : s(X_{test}, y) \leq \text{Quantile}(1-\alpha; \{S_i\}_{i=1}^n \cup \{\infty\})\right\}$

Theorem 2.1. *(same result as in the main text (Lei et al. [31], Vovk et al. [64])*
Let $\{(X_i, Y_i)\}_i^n$ be i.i.d. (or only exchangeable) random variables, and $\{S_i\}_{i=1}^n$ be the corresponding set of scores $S_i = s(X_i, Y_i)$ given to each pair (X_i, Y_i) by some score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Then for a new i.i.d. draw (X_{test}, Y_{test}) and any target error rate $\alpha \in (0, 1)$, the prediction set constructed as

$$\mathcal{C}(X_{test}) = \left\{y : s(X_{test}, y) \leq \text{Quantile}(1 - \alpha; \{S_i\}_{i=1}^n \cup \{\infty\})\right\}$$

satisfies the marginal coverage property

$$\mathbb{P}(Y_{test} \in \mathcal{C}(X_{test})) \geq 1 - \alpha,$$

Moreover, if $\{S_i\}_{i=1}^n$ are almost surely distinct, this probability is upper bounded by $1 - \alpha + 1/n+1$.

Proof. For simplicity, we assume that the set of scores $\{S_i\}_{i=1}^n$ are distinct (or have been made distinct by a suitable random tie-breaking rule). Denote $S_{test} = s(X_{test}, Y_{test})$ and observe that since $s(\cdot)$ is applied element-wise to each pair (X_i, Y_i) it preserves exchangeability, and thus random variables $\{S_i\}_{i=1}^n \cup \{S_{test}\}$ are also exchangeable. Next, we show that the following events are all equivalent

$$\begin{aligned} Y_{test} \in \mathcal{C}(X_{test}) &\overset{(i)}{\iff} S_{test} \leq \text{Quantile}(1 - \alpha; \{S_i\}_{i=1}^n \cup \{\infty\}) \\ &\overset{(ii)}{\iff} S_{test} \leq \text{Quantile}(1 - \alpha; \{S_i\}_{i=1}^n \cup \{S_{test}\}) \\ &\overset{(iii)}{\iff} \text{rank}(S_{test}; \{S_i\}_{i=1}^n \cup \{S_{test}\}) \leq \lceil (n+1)(1-\alpha) \rceil. \end{aligned}$$

(i) follows from the construction of the prediction set itself $\mathcal{C}(X_{test})$ itself.

(ii) can be easily verified as follows. If $S_{test} \leq \text{Quantile}(1-\alpha; \{S_i\}_{i=1}^n \cup \{\infty\})$, then shifting all values $S_i \geq S_{test}$ to arbitrary values larger than S_{test} will not change the validity of the inequality, since the $1-\alpha$ quantile remains unchanged. In particular, the inequality holds when replacing $\{S_{test}\}$ with $\{\infty\}$ and vice-versa.

(iii) follows from the fact that if $S_{test} \leq \text{Quantile}(1-\alpha; \{S_i\}_{i=1}^n \cup \{S_{test}\})$, then S_{test} is among the $\lceil (n+1)(1-\alpha) \rceil$ smallest values of the set $\{S_i\}_{i=1}^n \cup \{S_{test}\}$.

Finally, this is where the crucial exchangeability assumption comes into play, allowing us to apply Corollary B.1 to get

$$\mathbb{P}(Y_{test} \in \mathcal{C}(X_{test})) = \mathbb{P}(\text{rank}(S_{test}; \{S_i\}_{i=1}^n \cup S_{test}) \leq \lceil (n+1)(1-\alpha) \rceil) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}$$

From there, it is easy to verify that the right hand side is at least $1-\alpha$ and at most $1-\alpha+1/n+1$. □

C Background on List Decoding and Basic Information Theoretic Results

In this section, we enunciate some classical results from information theory that are instrumental in deriving our main theoretical results, as we show in the detailed proofs in Appendix D. We also provide a brief introduction to list decoding and demonstrate conformal prediction can be framed as a list decoding problem.

C.1 Data processing inequalities for f-divergence

We start by presenting the data processing inequality (DPI) for f-divergence, which we define below.

Definition C.1 (f-Divergence)**.** Consider two probability measures P and Q and assume that the measure P is absolutely continuous with respect to Q, i.e., $P \ll Q$. For a convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, the f-divergence is defined as:

$$D_f(P||Q) := \mathbb{E}_Q \left[f\left(\frac{dP}{dQ} \right) \right],$$

where $\frac{dP}{dQ}$ is a Radon-Nikodym derivative. In particular, using $f(x) = x \log x$ we recover the familiar notion of KL-divergence

$$D_{KL}(P||Q) := \mathbb{E}_P \left[\log \left(\frac{dP}{dQ} \right) \right].$$

We can similarly define conditional f-divergence.

Definition C.2 (Conditional f-Divergence)**.** Consider two probability measures P and Q such that $P := P_X P_{Y|X}$ and $Q := P_X Q_{Y|X}$ and that the measure P is absolutely continuous with respect to Q. For a convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, the conditional f-divergence is defined as:

$$D_f(P_{Y|X}||Q_{Y|X}|P_X) := \mathbb{E}_{P_X} \left[D_f(P_{Y|X=x}||Q_{Y|X=x}) \right].$$

Theorem C.3 is the classical data processing inequality (DPI), which we restate below for the sake of completeness. The classical versions of DPI, stated in terms of mutual information, can be found in classical information theoretic text books like [14], while the generalization of DPI to f-divergences can be found in other works with a comprehensive survey in [51, 58].

Theorem C.3 (Data Processing Inequality for f-divergence)**.** *Consider a conditional distribution $W_{Y|X}$. Suppose that P_Y and Q_Y are two distributions obtained by marginalization of $P_X W_{Y|X}$ and $Q_X W_{Y|X}$ over X. For any f-divergence, we have*

$$D_f(P_X||Q_X) \geq D_f(P_Y||Q_Y).$$

The proof can be found in standard textbooks in information theory see, for example, Chapter 7, Section 7.2 in [51] for a derivation of the DPI enunciated in the same form as above. Next, we consider the application of the DPI to conformal prediction.

Theorem C.4 (Data Processing Inequality for Conformal Prediction)**.** *For any set function $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$, any f-divergence, and all distribution pairs P, Q on (X, Y), we have:*

$$D_f(P||Q) \geq d_f\left(P(Y \in \mathcal{C}(x))||Q(Y \in \mathcal{C}(x))\right),$$

where $d_f(p||q)$ is the binary f-divergence, namely $d_f(p||q) = qf(p/q) + (1-q)f(1-p/1-q)$.

Proof. Consider the random variable $E = \mathbf{1}(Y \in \mathcal{C}(x))$ denoting the event of valid coverage, and the conditional distribution $P_{E|X,Y} = \mathbb{E}_{\mathcal{D}_{cal}}[\mathbf{1}(Y \in \mathcal{C}(x))|X, Y]$. Note that $P_{E|X,Y}$ maps distributions $P_{X,Y}$ and $Q_{X,Y}$ to P_E and Q_E, resp. From here, the result follows directly from Theorem C.3. □

The construction in Theorem C.4 can be further improved as follows.

Theorem C.5 (Conditional Data Processing Inequality for Conformal Prediction)**.** *For any set function $C : \mathcal{X} \to 2^{\mathcal{Y}}$, any f-divergence, and all conditional distribution pairs $P_{Y|X}, Q_{Y|X}$, and P_X, we have:*

$$\mathbb{E}_{P_X} D_f(P_{Y|X=x}||Q_{Y|X=x}) \geq \mathbb{E}_{P_X} d_f\left(P_{Y|X}(Y \in \mathcal{C}(x)|X = x)||Q_{Y|X}(Y \in \mathcal{C}(x)|X = x)\right),$$

where $d_f(p||q)$ is the binary f-divergence, namely $d_f(p||q) = qf(p/q) + (1-q)f(1-p/1-q)$.

Proof. Consider the conditional distribution $P_{E|X=x,Y} = \mathbb{E}_{\mathcal{D}_{cal}}[\mathbf{1}(Y \in \mathcal{C}(x))|X = x, Y]$. We have that

$$\begin{aligned} D_f(P_{Y|X=x}||Q_{Y|X=x}) &= D_f(P_{Y|X=x}P_{E|X=x,Y}||Q_{Y|X=x}P_{E|X=x,Y}) \\ &= D_f(P_{Y,E|X=x}||Q_{Y,E|X=x}) \end{aligned}$$

and from the monotonicity property of f-divergences [51] we have that

$$\begin{aligned} D_f(P_{Y|X=x}||Q_{Y|X=x}) &= D_f(P_{Y,E|X=x}||Q_{Y,E|X=x}) \\ &\geq D_f(P_{E|X=x}||Q_{E|X=x}) \\ &= d_f(P_{Y|X}(Y \in C(x)|X = x)||Q_{Y|X}(Y \in C(x)|X = x)). \end{aligned}$$

By taking the expectation with respect to P_X, we conclude the proof. □

C.2 List Decoding

List decoding [18, 68] is a notion coming from coding theory, a large branch of engineering and mathematics that arises from the application of information theory to the design of reliable communication systems and robust information processing and storage. In particular, we are interested in channel coding, a field concentrated on the design of so-called error-correcting codes to enable reliable communication over inaccurate or noisy communication channels.

The general setup studied in channel coding, including list decoding, can be summarized as follows. A message $y \in \mathcal{Y}$ is encoded and transmitted over a noisy channel, and a message $x \in \mathcal{X}$ is received. The noisy channel is governed by probability density $p(x|y)$ that describes the probability of observing output $x \in \mathcal{X}$ given input $y \in \mathcal{Y}$. The receiver attempts to *decode* x, that is, to guess the originally transmitted message y from the received one, x. At this point, parallels to machine learning should already have become clear. If the receiver provides a single guess of the transmitted message y, we are in a unique-decoding scenario, which is akin to a point prediction in machine learning. Conversely, if the receiver is allowed to guess a set (or a list in the list decoding formalism) of the most likely messages, we have list decoding, which closely resembles conformal prediction. Note that in many settings, a list decoding algorithm is constrained to output a list of fixed size. While simple conformal prediction methods for regression settings show the same limitation, the parallel between the two domains is more pertinent when we consider *variable-size* list decoding [57].

More formally, a list-decoding algorithm can be defined by a set predictor $L : \mathcal{X} \to 2^{\mathcal{Y}}$, with maximum output set size $|L(X)| \leq M$. Naturally, the goal is to design the function L so as to maximize the probability of $Y \in L(X)$. Similarly, for a given input-label pair (X, Y) the goal of conformal prediction is to provide a set that contains Y with a certain pre-determined probability. It is this connection that motivates our bounds on the conditional entropy $H(Y|X)$. However, the nonconformity score in conformal prediction is typically a function of the output of a given model

$f : \mathcal{X} \to \mathcal{Y}$. For instance, $\hat{Y} = f(X)$ could be the logits output by the model, i.e., a vector with the dimension equal to number of classes. If we consider $\hat{Y}$ the noisy observation of the ground-truth Y, then the problem of determining a set containing Y is again the list decoding problem. Therefore, we can also consider the communication channel $p(\hat{y}|y;x)$ directly, which justifies applying the same upper bounds to $H(Y|\hat{Y})$. Note that in this last case, the input data X can also be taken as side information, although it is rarely used directly in building the conformal prediction set.

This reinterpretation of conformal prediction as list decoding allows us to apply some of the standard results from the list decoding literature to conformal prediction, as we show in the next section.

C.3 List decoding: Information Theoretic Inequalities

Fano's inequality for variable size list decoding. The following generalization of Fano's inequality is given in [[53], Appendix 3.E].

Theorem C.6. *Consider a scenario where a decoder upon observing $\hat{Y}$ provides a nonempty list $L(\hat{Y})$ that contains another random variable $Y \in \mathcal{Y}$ with $|\mathcal{Y}| = M$. Define $P_e := \mathbb{P}(Y \notin L(\hat{Y}))$. We have:*

$$H(Y|\hat{Y}) \leq h_b(P_e) + P_e \log(M) + \mathbb{E}(\log |L(\hat{Y})|).$$

Optimal list decoding and conformal prediction. It can be shown that the optimal list decoder consists of selecting $L(\hat{y})$ elements of $\mathcal{Y}$ with highest conditional probability $p(y|\hat{y})$. That is, consider the sorted posteriors under the true distribution $p(y_1|\hat{y}) \geq p(y_2|\hat{y}) \geq \cdots \geq p(y_M|\hat{y})$ and choose the first $\{y_1, \ldots, y_{|L|}\}$ for some list size $|L|$. However, this rule is for fixed-size list decoding and does not determine how to select the coverage set size to guarantee a given coverage. We can modify this rule to obtain a variable-size list decoding with the required coverage. Assuming again the sorted posteriors $p(y_1|\hat{y}) \geq p(y_2|\hat{y}) \geq \cdots \geq p(y_M|\hat{y})$, we can select the set as follows:

$$L(\hat{y}) = \{y_1, \ldots, y_{\ell_y}\} \quad \text{where } \ell_y := \inf \left\{ j : \sum_{i=1}^{j} p(y_i|\hat{y}) \geq 1 - \alpha \right\}.$$

It is easy to see that the above set is the smallest set given each y and the confidence level $1 - \alpha$ (see, for example, [43]). The same result holds in conformal prediction [55, 65].

C.4 Fano and Data Processing Inequalities for Conformal Prediction

First, from Fano's inequality for list decoding, Theorem C.6, we get the next result "out-of-the-box".

Proposition C.7. Suppose that $|\mathcal{Y}| = M$. Any conformal prediction method with the prediction set $\mathcal{C}(x)$ and confidence level $1 - \alpha$, $\alpha \in (0, 0.5)$, satisfies the following inequality:

$$H(Y|X) \leq h_b(\alpha) + \alpha \log(M) + \mathbb{E}([\log |\mathcal{C}(x)|]^+),$$

where $h_b(\cdot)$ is the binary entropy function, $[x]^+ := \max\{x, 0\}$ and $H(Y|X)$ is computed using the true distribution P_{XY}. When the conformal prediction is merely based on the model output $\hat{Y} = f(X)$, the inequality can be modified to:

$$H(Y|\hat{Y}) \leq h_b(\alpha) + \alpha \log(M) + \mathbb{E}([\log |\mathcal{C}(x)|]^+).$$

The proposition follows easily from Theorem C.6 by using the condition $\mathbb{P}(Y \in \mathcal{C}(x)) \geq 1 - \alpha$. Note that in Theorem C.6, we assume non-empty lists, whereas in Proposition C.7 we allow empty prediction sets but apply the maximum operator $[x]^+ := \max\{x, 0\}$. This is justified because the last term of Fano's inequality relates to the probability of correct assignments $Y \in \mathcal{C}(X)$, which never happens for empty sets. See Proposition E.1 for the proof, where the same result appears.

The bounds that we present in the main paper, model-based and simple Fano bounds, are actually derived through a slightly different root by leveraging the lower and upper bounds in the finite-sample guarantee of conformal prediction (Theorem 2.1). We derive these other two bounds in Appendix D.2.

C.5 Related Work on Information Theoretic Inequalities

The use of f-divergences goes back to works of Ali and Silvey, Csiszár, and Morimoto in the 60s—see for instance [2, 15, 44]. A key f-divergence inequality is the data processing inequality, which was used in information theory to establish various upper bounds on the achievability of coding schemes for different tasks—see [57, 58] for an extensive survey. Moreover, the data processing inequality for f-divergences provides a unified way of obtaining many classical and new results, for example Fano's inequality [19]. The tightness of data processing inequalities is discussed in terms of Bregman's divergence in [12, 34] and in terms of χ^2-divergence in [57].

When it comes to list decoding, there are a number of relevant inequalities in the literature. List decoding was introduced in the context of communication design by Elias [18] and Wozencraft [68]. See also [22] for a more recent overview of list decoding. For fixed list size, the information theoretic bounds on list decoding were obtained in [20] using error exponent analysis. A generalization of Fano's inequality to list decoding was given in [16] in the context of multi-user information theory, see also the general Fano inequality for list decoding presented in [53]. For fixed list size, stronger inequalities, some based on Arimoto-Rényi conditional entropy, were presented in [59]. Variable size list decoding was discussed in [57] using the notion of E_γ resolvability first introduced in [52] related to the dependence testing bound. It was used again in the context of channel resolvability in [35], where some relevant inequalities have been obtained and discussed. A selection of the most relevant inequalities for list decoding can be found in [57].

D Proofs of Main Theoretical Results

In this section, we provide the proofs of our main results, namely the *DPI bound* in Proposition 3.1, the *model-based Fano bound* in Proposition 3.2, and the *simple Fano bound* in Corollary 3.1. For notational convenience, we use the shorthand $\alpha_n = \alpha - 1/n+1$ in most of the steps of the derivations.

D.1 DPI Bound

We start with the DPI bound which we restate and proof below using the data processing inequalities discussed in Appendix C. Note that, when clear from the context, we remove explicit dependence on the calibration set $\mathcal{D}_{cal}$ from the derivations. It is implicitly assumed that the probability of the event $Y \in \mathcal{C}(x)$ is computed by marginalizing over $\mathcal{D}_{cal}$.

Proposition 3.1. *Consider any conformal prediction method with the prediction set $\mathcal{C}(x)$ with the following finite sample guarantee:*

$$1 - \alpha \leq \mathbb{P}(Y \in \mathcal{C}(x)) \leq 1 - \alpha + \frac{1}{n+1}$$

for any $\alpha \in (0, 0.5)$. For any arbitrary conditional distribution $Q_{Y|X}$, the true conditional distribution $P_{Y|X}$ and the input measure P_X, define the following two measures $Q := P_X Q_{Y|X}$ and $P := P_X P_{Y|X}$. We have for any $\alpha \in (0, 0.5)$

$$H(Y|X) \leq h_b(\alpha) + \left(1 - \alpha + \frac{1}{n+1}\right) \log Q(Y \in \mathcal{C}(x)) + \alpha \log Q(Y \notin \mathcal{C}(x)) - \mathbb{E}_{P_{XY}}\left[\log Q_{Y|X}\right].$$

Proof. Consider an arbitrary distribution $Q_{Y|X}$. Then we can use P_{XY}, and $P_X \times Q_{Y|X}$ in the data processing inequality for KL-divergence (Theorem C.4) to get:

$$D_{KL}(P_X P_{Y|X} || P_X Q_{Y|X}) \geq d_{KL}(P(Y \in \mathcal{C}(x)) || Q(Y \in \mathcal{C}(x))) \tag{12}$$

Now note that we can decompose $D_{KL}(P_X P_{Y|X} || P_X Q_{Y|X})$ in terms of the conditional entropy $H(Y|X)$ and the cross-entropy $-\mathbb{E}_{P_{XY}}[\log Q_{Y|X}]$:

$$\begin{aligned} D_{KL}(P_X P_{Y|X} || P_X Q_{Y|X}) &= \mathbb{E}_{P_{XY}}\left[\log \frac{P_X P_{Y|X}}{P_X Q_{Y|X}}\right] = \mathbb{E}_{P_{XY}}\left[\log \frac{P_{Y|X}}{Q_{Y|X}}\right] \\ &= \mathbb{E}_{P_{XY}}[\log P_{Y|X}] - \mathbb{E}_{P_{XY}}[\log Q_{Y|X}] \\ &= -H(Y|X) - \mathbb{E}_{P_{XY}}[\log Q_{Y|X}]. \end{aligned}$$

With the decomposition above, we can rearrange the terms in (12) to get the following upper bound on $H(Y|X)$

$$
\begin{aligned}
-H(Y|X) - \mathbb{E}_{P_{XY}}\left[\log Q_{Y|X}\right] &\geq d_{KL}(P(Y \in \mathcal{C}(x))||Q(Y \in \mathcal{C}(x))) \\
H(Y|X) &\leq -d_{KL}(P(Y \in \mathcal{C}(x))||Q(Y \in \mathcal{C}(x))) - \mathbb{E}_{P_{XY}}\left[\log Q_{Y|X}\right]. \quad (13)
\end{aligned}
$$

We can then apply the upper and lower bounds from conformal prediction, i.e. $P(Y \in \mathcal{C}(x)) \geq 1-\alpha$ and $P(Y \notin \mathcal{C}(x)) \geq \alpha_n$, to upper bound $d_{KL}(P(Y \in \mathcal{C}(x))||Q(Y \in \mathcal{C}(x)))$ as follows.

$$
\begin{aligned}
&-d_{KL}(P(Y \in \mathcal{C}(x))||Q(Y \in \mathcal{C}(x))) = \\
&\quad h_b\left(P(Y \in \mathcal{C}(x))\right) + P(Y \in \mathcal{C}(x)) \log Q(Y \in \mathcal{C}(x)) + P(Y \notin \mathcal{C}(x)) \log Q(Y \notin \mathcal{C}(x)) \\
&\qquad\qquad \leq h_b\left(\alpha\right) + (1-\alpha) \log Q(Y \in \mathcal{C}(x)) + \alpha_n \log Q(Y \notin \mathcal{C}(x)),
\end{aligned}
$$

where $h_b(\cdot)$ is the binary entropy function, that is, $h_b(\alpha) = -\alpha\log(\alpha) - (1-\alpha)\log(1-\alpha)$. The equality in the second line follows from the definition of the binary KL divergence, and we get the inequality simply by upper bounding each of the terms individually. In particular, note that $\log Q(Y \in \mathcal{C}(x))$ and $\log Q(Y \notin \mathcal{C}(x))$ are both negative, and $h_b(\cdot)$ is decreasing in $[0.5, 1.0]$ and symmetric about 0.5, such that for typical values of $\alpha < 0.5$

$$
P(Y \in \mathcal{C}(X)) \geq 1-\alpha \implies h_b\left(P(Y \in \mathcal{C}(x))\right) \leq h_b(1-\alpha) = h_b(\alpha). \quad (14)
$$

Finally, we can replace the upper bound above into (13) to conclude the proof.

$$
H(Y|X) \leq h_b(\alpha_n) + (1-\alpha)\log Q(Y \in \mathcal{C}(x)) + \alpha_n \log Q(Y \notin \mathcal{C}(x)) - \mathbb{E}_{P_{XY}}\left[\log Q_{Y|X}\right].
$$

□

Remark D.1. The DPI bound always provides a tighter upper bound on the conditional entropy $H(Y|X)$ than the cross-entropy, which is easily verified in (13) since the KL term is always non-negative. This serves as an important motivation to optimize the DPI bound instead of the cross-entropy in conformal training.

Remark D.2. The derivation of the DPI bound places no assumptions on the conditional distribution $Q_{Y|X}$. However, in practice, the underlying model f often already provides such a distribution, and since it is typically trained to approximate $P_{Y|X}$ well, it makes sense to take $Q_{Y|X}$ as the distribution defined by the model itself. That is how we evaluate the DPI in all of our experiments. Finally, we can again use $H(Y|\hat{Y})$ instead of $H(Y|X)$ if the conformal method uses merely $\hat{Y}$.

Remark D.3. Typically, we have $\alpha \in (0.0, 0.5)$, and we use this fact in the proof to bound the binary entropy $h_b\left(P(Y \in \mathcal{C}(x))\right)$. The same could have been done for $\alpha \in (0.5, 1.0)$, but since $1-\alpha$ lands in the increasing part of the binary entropy function between 0 and 0.5, we have to resort to the lower bound from conformal prediction to get

$$
P(Y \in \mathcal{C}(X)) \leq 1-\alpha_n \implies h_b\left(P(Y \in \mathcal{C}(x))\right) \leq h_b(1-\alpha_n) = h_b(\alpha_n).
$$

Remark D.4. One of the appeals of the DPI bound is that the terms can be computed in a data-driven way using samples. While the cross-entropy can be estimated in an unbiased way with samples from the true distribution P_{XY}, we must be careful when estimating $Q(Y \in \mathcal{C}(x))$. The main challenge is that $Q(Y \in \mathcal{C}(x))$ appears inside a log, and thus an empirical estimate $\hat{Q}(Y \in \mathcal{C}(x))$ would yield a lower bound of the negative KL divergence and would be biased. We can get an upper confidence bound on this estimate via the empirical Bernstein inequality [41], which we restate below.

Theorem D.5 (Empirical Bernstein Inequality [41]). *Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables with values in [0, 1] and let $\delta > 0$. Then with probability at least $1-\delta$ in the i.i.d. vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ we have that*

$$
\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i \leq \sqrt{\frac{2V_n(\mathbf{Z})\log(2/\delta)}{n}} + \frac{7\log(2/\delta)}{3(n-1)},
$$

where $V_n(\mathbf{Z})$ is the empirical variance over the $(Z_1, \ldots, Z_n)$ samples.

By using this bound, we have with probability $1-\delta$:

$$Q(Y \in \mathcal{C}(x)) \leq \hat{Q}(Y \in \mathcal{C}(x)) + \sqrt{\frac{2V_n(\mathbf{Z})\log(2/\delta)}{n}} + \frac{7\log(2/\delta)}{3(n-1)} := \tilde{Q}(Y \in \mathcal{C}(x)),$$
$$Q(Y \notin \mathcal{C}(x)) \leq \hat{Q}(Y \notin \mathcal{C}(x)) + \sqrt{\frac{2V_n(\mathbf{Z})\log(2/\delta)}{n}} + \frac{7\log(2/\delta)}{3(n-1)} := \tilde{Q}(Y \notin \mathcal{C}(x)),$$

where $Z_i = Q(y_i \in C(x_i))$ for x_i sampled from P_X and a prediction set $C(x_i)$ obtained with a calibration dataset sampled from $P_{\mathcal{D}_{cal}}$. This yields the following inequality with probability $1-\delta$:

$$H(Y|X) \leq h_b(\alpha) + (1-\alpha)\log\tilde{Q}(Y \in \mathcal{C}(x)) + \alpha_n \log\tilde{Q}(Y \notin \mathcal{C}(x)) - \mathbb{E}_{P_{XY}}\left[\log Q_{Y|X}\right].$$

We can use this inequality to evaluate the bounds.

D.2 Model-Based Fano Bound

In this section, prove the model-based Fano bound, which we restate in Proposition 3.2 below.

Proposition 3.2. *Consider any conformal prediction method with the prediction set $\mathcal{C}(x)$, and any distribution Q, with the following finite sample guarantee:*

$$1-\alpha \leq \mathbb{P}(Y \in \mathcal{C}(x)) \leq 1-\alpha+\frac{1}{n+1},$$

for $\alpha \in (0, 0.5)$. Then, for the true distribution P, and for any probability distribution Q, we have

$$H(Y|X) \leq h_b(\alpha) + \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\notin\mathcal{C}(X)}}\left[-\log Q_{Y|X,\mathcal{C}(x),Y\notin\mathcal{C}(X)}\right] + \left(1-\alpha+\frac{1}{n+1}\right)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\in\mathcal{C}(X)}}\left[-\log Q_{Y|X,\mathcal{C}(x),Y\in\mathcal{C}(X)}\right].$$

Proof. For notational convenience, define $E = \mathbf{1}(Y \in \mathcal{C}(x))$, which by our assumption on the conformal prediction method means that $1-\alpha \leq \mathbb{P}(E=1) \leq 1-\alpha_n$. The starting point is similar to the well-known proof of Fano's inequality:

$$H(E,Y|X,\mathcal{D}_{cal}) = H(Y|X,\mathcal{D}_{cal}) + H(E|Y,X,\mathcal{D}_{cal}) = H(Y|X,\mathcal{D}_{cal}),$$

where the last step follows because knowing X, Y and $\mathcal{D}_{cal}$, we know if $Y \in \mathcal{C}(x)$, and therefore $H(E|Y,X,\mathcal{D}_{cal}) = 0$. Furthermore, given the structure of the graphical model of the conformal prediction process (c.f. Figure 1), we have that $H(Y|X) = H(Y|X,\mathcal{D}_{cal})$, We now find an upper bound on $H(E,Y|X,\mathcal{D}_{cal})$:

$$\begin{aligned}
H(E,Y|X,\mathcal{D}_{cal}) &= H(E|X,\mathcal{D}_{cal}) + H(Y|X,E,\mathcal{D}_{cal}) \\
&= H(E|X,\mathcal{D}_{cal}) + P(E=0)H(Y|X,E=0,\mathcal{D}_{cal}) + P(E=1)H(Y|X,E=1,\mathcal{D}_{cal}) \\
&\leq H(E) + P(E=0)H(Y|X,E=0,\mathcal{D}_{cal}) + P(E=1)H(Y|X,E=1,\mathcal{D}_{cal}) \\
&\leq h_b(\alpha) + P(E=0)H(Y|X,E=0,\mathcal{D}_{cal}) + P(E=1)H(Y|X,E=1,\mathcal{D}_{cal}), \qquad (15)
\end{aligned}$$

where the first inequality follows from the fact that $H(E|X,\mathcal{D}_{cal}) \leq H(E)$, and the last one comes from the same argument for $a < 0.5$ in (14). We can continue as follows

$$\begin{aligned}
H(Y|X) &\leq h_b(\alpha) + P(E=0)H(Y|X,E=0,\mathcal{D}_{cal}) + P(E=1)H(Y|X,E=1,\mathcal{D}_{cal}) \\
&= h_b(\alpha) + P(E=0)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log P_{Y|X,\mathcal{D}_{cal},E=0}\right] \\
&\quad + P(E=1)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log P_{Y|X,\mathcal{D}_{cal},E=1}\right] \\
&= h_b(\alpha) + P(E=0)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log Q_{Y|X,\mathcal{C}(x),E=0}\right] \\
&\quad - \mathbb{E}[D_{KL}(P_{Y|X,\mathcal{D}_{cal},E=0}||Q_{Y|X,\mathcal{C}(x),E=0})] \\
&\quad + P(E=1)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log Q_{Y|X,\mathcal{C}(x),E=1}\right] \\
&\quad - \mathbb{E}[D_{KL}(P_{Y|X,\mathcal{D}_{cal},E=1}||Q_{Y|X,\mathcal{C}(x),E=1})] \\
&\leq h_b(\alpha) + P(E=0)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log Q_{Y|X,\mathcal{C}(x),E=0}\right] \\
&\quad + P(E=1)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log Q_{Y|X,\mathcal{C}(x),E=1}\right] \qquad (16)
\end{aligned}$$

The first equality comes from the definition of the conditional entropy, whereas in the second equality we replace the true distribution $P_{Y|X}$ with an arbitrary conditional distribution $Q_{Y|X}$ plus the KL divergence between the two distributions. The last inequality then follows simply from the fact that the KL divergence is non-negative. Finally, we can leverage the finite-sample guarantees from conformal prediction, namely $P(E=0) \leq \alpha$ and $P(E=1) \leq 1-\alpha_n$, to upper bound each term, which yields the proof.

$$\begin{aligned} H(Y|X) &\leq h_b(\alpha) + P(E=0)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log Q_{Y|X,\mathcal{C}(x),E=0}\right] \\ &\quad + P(E=1)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log Q_{Y|X,\mathcal{C}(x),E=1}\right] \\ &\leq h_b(\alpha) + \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log Q_{Y|X,\mathcal{C}(x),E=0}\right] \\ &\quad + (1-\alpha_n)\,\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log Q_{Y|X,\mathcal{C}(x),E=1}\right] \end{aligned} \tag{17}$$

□

Remark D.6. When the conformal prediction is merely based on the model output $\hat{Y} = f(X)$, the model-based Fano bound can be modified to

$$\begin{aligned} H(Y|\hat{Y}) \leq h_b(\alpha) &+ \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\notin\mathcal{C}(X)}}\left[-\log Q_{Y|X,\mathcal{C}(x),Y\notin\mathcal{C}(X)}\right] \\ &+ \left(1-\alpha+\frac{1}{n+1}\right)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\in\mathcal{C}(X)}}\left[-\log Q_{Y|X,\mathcal{C}(x),Y\in\mathcal{C}(X)}\right]. \end{aligned}$$

D.3 Simple Fano

In the derivation of the model-based Fano bound above, we placed no assumptions on the distribution $Q_{Y|X}$. One simple choice that we consider in this section is the uniform distribution $Q_{Y|X} = 1/|\mathcal{Y}|$, akin to the classical Fano's inequality [19] and the list decoding result in Proposition C.7.

Corollary 3.1. *Consider any conformal prediction method with the prediction set $\mathcal{C}(x)$, and any distribution Q, with the following finite sample guarantee:*

$$1-\alpha \leq \mathbb{P}(Y \in \mathcal{C}(x)) \leq 1-\alpha+\frac{1}{n+1}.$$

For $\alpha \in (0, 0.5)$ we have the following inequality:

$$\begin{aligned} H(Y|X) \leq h_b(\alpha) &+ \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\notin\mathcal{C}(X)}}\left[\log(|\mathcal{Y}| - |C(X)|)\right] \\ &+ (1-\alpha_n)\,\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\in\mathcal{C}(X)}}\left[\log|C(X)|\right]. \end{aligned}$$

Proof. Note that if we make an error ($E=0$) the correct class will not be inside $\mathcal{C}(X)$ and since $Q_{Y|X}$ is uniform the probability will be spread equally among the remaining $|\mathcal{Y}| - |\mathcal{C}(X)|$ labels, and we have $Q_{Y|X,\mathcal{C}(x),E=0} = \frac{1}{|\mathcal{Y}|-|\mathcal{C}(X)|}$. Through the same logic, we get that $Q_{Y|X,\mathcal{C}(x),E=1} = \frac{1}{|\mathcal{C}(X)|}$, and plugging both into (17) we get the simple Fano bound □

Remark D.7. Once again, when the conformal prediction prediction is merely based on the model output $\hat{Y} = f(X)$, the inequality can be modified to

$$\begin{aligned} H(Y|\hat{Y}) \leq h_b(\alpha) &+ \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\notin\mathcal{C}(X)}}\left[\log(|\mathcal{Y}| - |C(X)|)\right] \\ &+ (1-\alpha_n)\,\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\in\mathcal{C}(X)}}\left[\log|C(X)|\right]. \end{aligned}$$

E Further Theoretical Results and Proofs on the Prediction Set Size

E.1 Fano's inequality for maximal prediction set size

If we leverage the upper bound on the prediction set size, we can find a lower bound on the maximum coverage set size.

Proposition E.1. Suppose that $|\mathcal{Y}| = M$. Consider any conformal prediction method that constructs the prediction set $\mathcal{C}(X)$ with the following finite-sample guarantee:

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

for $\alpha \in (0, 0.5)$. Then, we have the following inequality:

$$H(Y|X) \leq h_b(\alpha) + \alpha \log M + (1 - \alpha + \frac{1}{n+1}) \sup_{\mathcal{D}_{cal}} \sup_{x \in \text{supp}(P_X)} \log |\mathcal{C}(x)|.$$

We can similarly replace $H(Y|X)$ with $H(Y|\hat{Y})$.

Proof. We use the conditional data processing inequality with $f(x) = x \log x$, $\mathrm{P} = P_{\mathcal{D}_{cal}XY}$, and $\mathrm{Q} = P_{\mathcal{D}_{cal}X} \times U_M$ where U_M is the uniform distribution over $\mathcal{Y}$. We fix the input to $X = x$, and the calibration set $\mathcal{D}_{cal}$. The conditional f-divergence, conditioned on $\mathcal{D}_{cal}$ and X, is given by:

$$\begin{aligned} D_f(\mathrm{Q}_{Y|X=x,\mathcal{D}_{cal}} || \mathrm{P}_{Y|X=x,\mathcal{D}_{cal}}) &= D_{KL}(P_{Y|\mathcal{D}_{cal}X=x} || U_M) \\ &= \log M - H(Y|X=x), \end{aligned}$$

where the last step follows from the independence of the calibration set $\mathcal{D}_{cal}$ and (X, Y). This means that

$$\mathbb{E}_{\mathcal{D}_{cal},X} D_f(P_{Y|X=x,\mathcal{D}_{cal}} || Q_{Y|X=x,\mathcal{D}_{cal}}) = \log M - H(Y|X). \tag{18}$$

On the other hand, we have

$$\begin{aligned} &d_f\left(Q(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X=x) || P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X=x)\right) \\ &\quad = d_{KL}\left(P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X=x) || Q(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X=x)\right) \\ &\quad = -h_b\left(P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X=x)\right) - P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X=x) \log \frac{|\mathcal{C}(x)|}{M} \\ &\qquad - P(Y \notin \mathcal{C}(x)|\mathcal{D}_{cal}, X=x) \log \frac{M - |\mathcal{C}(x)|}{M}. \end{aligned} \tag{19}$$

Now, we can plug both (18) and (19) into the data processing inequality of Theorem C.5. Rearranging the terms and getting the expectation from both sides w.r.t. $\mathcal{D}_{cal}$ and X would yield:

$$\begin{aligned} H(Y|X) &\leq \mathbb{E}[h_b\left(P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X)\right)] \\ &\quad + \mathbb{E}[P(Y \notin \mathcal{C}(x)|\mathcal{D}_{cal}, X) \log(M - |\mathcal{C}(x)|)] + \mathbb{E}[P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X) \log |\mathcal{C}(x)|] \\ &\quad - \mathbb{E}[P(Y \notin \mathcal{C}(x)|\mathcal{D}_{cal}, X) \log M] - \mathbb{E}[P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X) \log M] + \log M \\ &\leq h_b\left(P(Y \in \mathcal{C}(x))\right) + P(Y \notin \mathcal{C}(x)) \log M + \mathbb{E}[P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X) \log |\mathcal{C}(x)|] \\ &\leq h_b(\alpha) + \alpha \log M + \mathbb{E}[P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X) \log |\mathcal{C}(x)|]. \end{aligned}$$

Note that the $\log M$ terms in the third line cancel each other out, and that the second inequality comes from resolving the expectations and $\log(M - |\mathcal{C}(X)|) \leq \log M$. Further, in the last inequality, we used the concavity of binary entropy as in (14). Finally, using respectively the inequalities $P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X) \leq 1$ and $\log |\mathcal{C}(x)| \leq \sup_{\mathcal{D}_{cal}} \sup_{x \in \mathcal{X}} \log |\mathcal{C}(x)|$, we get two inequalities:

$$\begin{aligned} H(Y|X) &\leq h_b(\alpha) + \alpha \log M + \mathbb{E}\left([\log |\mathcal{C}(x)|]^+\right) \\ H(Y|X) &\leq h_b(\alpha) + \alpha \log M + (1 - \alpha + \frac{1}{n+1}) \sup_{\mathcal{D}_{cal}} \sup_{x \in \text{supp}(P_X)} \log |\mathcal{C}(x)|. \end{aligned}$$

The first one is the Fano's inequality for list decoding from Proposition C.7 while the second one yields the theorem. Note that if $\mathcal{C}(x)$ is an empty set, the probability $P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X)$ is zero, and the term $P(Y \in \mathcal{C}(x)|\mathcal{D}_{cal}, X) \log |\mathcal{C}(x)|$ disappears. Therefore, if we use the inequality $P(Y \in \mathcal{C}(x)| \sup_{\mathcal{D}_{cal}}, X) \leq 1$, we need to introduce the term $[\log |\mathcal{C}(x)|]^+$ to keep the expectation well defined. □

E.2 Fano's inequality for lower bound on prediction set size

In a similar manner, we can also obtain lower bounds for the set size. More specifically, we have that

$$Q_{Y|X,\mathcal{C}(x),E=1} = \frac{q(y|x)\mathbb{I}[y \in \mathcal{C}(x)]}{\sum_{y\in\mathcal{C}(x)} q(y|x)} = \frac{1}{|\mathcal{C}(x)|\mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]} q(y|x)\mathbb{I}[y \in \mathcal{C}(x)]$$
$$:= \frac{1}{|\mathcal{C}(x)|\mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]} \hat{Q}^1_{Y|X} \tag{20}$$
$$Q_{Y|X,\mathcal{C}(x),E=0} = \frac{q(y|x)\mathbb{I}[y \notin \mathcal{C}(x)]}{\sum_{y\notin\mathcal{C}(x)} q(y|x)} = \frac{1}{(M-|\mathcal{C}(x)|)\mathbb{E}_{u(y_{\bar{\mathcal{C}}(x)})}[q(y|x)]} q(y|x)\mathbb{I}[y \notin \mathcal{C}(x)]$$
$$:= \frac{1}{(M-|\mathcal{C}(x)|)\mathbb{E}_{u(y_{\bar{\mathcal{C}}(x)})}[q(y|x)]} \hat{Q}^0_{Y|X} \tag{21}$$

where $u(y_{\mathcal{C}(x)})$ and $u(y_{\bar{\mathcal{C}}(x)})$ denote uniform distributions over the labels in the confidence set. By considering the standard conformal prediction bounds on the error probabilities, we have that

$$\begin{aligned}
H(Y|X) &\leq h_b(\alpha) + \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log \hat{Q}^0_{Y|X} + \log(M-\mathcal{C}(x)) + \log \mathbb{E}_{u(y_{\bar{\mathcal{C}}(x)})}[q(y|x)]\right] \\
&+ \left(1-\alpha+\frac{1}{n+1}\right)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log \hat{Q}^1_{Y|X} + \log|\mathcal{C}(x)| + \log \mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]\right] \\
&\leq h_b(\alpha) + \alpha\log M + \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log \hat{Q}^0_{Y|X} + \log \mathbb{E}_{u(y_{\bar{\mathcal{C}}(x)})}[q(y|x)]\right] \\
&+ \left(1-\alpha+\frac{1}{n+1}\right)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log \hat{Q}^1_{Y|X} + \log|\mathcal{C}(x)| + \log \mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]\right]
\end{aligned}$$

which leads to

$$\begin{aligned}
&\mathbb{E}_{E=1}[\log|\mathcal{C}(x)|] \geq \\
&\frac{H(Y|X) - h_b(\alpha) - \alpha\log M - \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log \hat{Q}^0_{Y|X} + \log \mathbb{E}_{u(y_{\bar{\mathcal{C}}(x)})}[q(y|x)]\right]}{1-\alpha+\frac{1}{n+1}} \\
&- \mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log \hat{Q}^1_{Y|X} + \log \mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]\right].
\end{aligned}$$

Note that when $E=1$, we know that $Y \in \mathcal{C}(x)$, and therefore $|\mathcal{C}(x)| > 0$ and $[\log|\mathcal{C}(x)|]^+ = \log|\mathcal{C}(x)|$. Using this, we can find an upper bound on $\mathbb{E}_{E=1}(\log(|\mathcal{C}(x)|))$ as follows:

$$\begin{aligned}
\mathbb{E}_{E=1}(\log(|\mathcal{C}(x)|)) &= E_{E=1}([\log(|\mathcal{C}(x)|)]^+) \\
&= \left(\frac{\mathbb{E}([\log|\mathcal{C}(x)|]^+)}{P(E=1)} - \frac{P(E=0)}{P(E=1)}\mathbb{E}_{E=0}([\log|\mathcal{C}(x)|]^+)\right) \\
&\leq \left(\frac{\mathbb{E}([\log|\mathcal{C}(x)|]^+)}{P(E=1)}\right) \leq \left(\frac{\mathbb{E}([\log|\mathcal{C}(x)|]^+)}{1-\alpha}\right).
\end{aligned}$$

This leads to:

$$\begin{aligned}
&\mathbb{E}([\log|\mathcal{C}(x)|]^+) \geq \\
&(1-\alpha)\frac{H(Y|X) - h_b(\alpha) - \alpha\log M - \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log \hat{Q}^0_{Y|X} + \log \mathbb{E}_{u(y_{\bar{\mathcal{C}}(x)})}[q(y|x)]\right]}{1-\alpha+\frac{1}{n+1}} \\
&- (1-\alpha)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log \hat{Q}^1_{Y|X} + \log \mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]\right]
\end{aligned}$$

All the above terms can be approximated from samples. We summarize this in the following proposition.

Proposition E.2. For any conformal prediction scheme with the coverage guarantee of $1-\alpha$, and any distribution $q(\cdot)$, we have:

$$\mathbb{E}([\log|\mathcal{C}(x)|]^+) \geq$$
$$(1-\alpha)\frac{H(Y|X) - h_b(a) - a\log M - \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log\hat{Q}^0_{Y|X} + \log\mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]\right]}{1-\alpha+\frac{1}{n+1}}$$
$$-(1-\alpha)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log\hat{Q}^1_{Y|X} + \log\mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]\right] \tag{22}$$

where $\hat{Q}^0_{Y|X} = q(y|x)\mathbb{I}[y \notin \mathcal{C}(x)]$ and $\hat{Q}^1_{Y|X} = q(y|x)\mathbb{I}[y \in \mathcal{C}(x)]$. When the conformal prediction is merely based on the model output $\hat{Y} = f(X)$, the inequality can be modified to

$$\mathbb{E}([\log|\mathcal{C}(x)|]^+) \geq$$
$$(1-\alpha)\frac{H(Y|\hat{Y}) - h_b(a) - a\log M - \alpha\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=0}}\left[-\log\hat{Q}^0_{Y|X} + \log\mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]\right]}{1-\alpha+\frac{1}{n+1}}$$
$$-(1-\alpha)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|E=1}}\left[-\log\hat{Q}^1_{Y|X} + \log\mathbb{E}_{u(y_{\mathcal{C}(x)})}[q(y|x)]\right] \tag{23}$$

Remark E.3. If we use the uniform distribution in the above bound, we get a bound similar to what is obtained from Fano's inequality given in Proposition C.7, but with an additional factor of $\frac{1-\alpha}{1-\alpha+\frac{1}{n+1}}$. Since the factor is smaller than one, the current bound with the choice of uniform distribution is looser than Fano's bound, although the gap vanishes for large n.

F Conformal Training

Split conformal prediction (SCP) [47] has quickly become a popular framework for uncertainty quantification, largely thanks to its computational efficiency. One only needs access to a separate calibration data set to derive prediction sets with valid marginal coverage from any pretrained model. Given that training new machine learning models is becoming ever more time-consuming and expensive with new, larger architectures, this ability to apply conformal prediction to existing models is invaluable in a number of applications. Yet, it is reasonable to expect that the performance of the final set predictor could be improved if the conformal prediction process were to be accounted for during training of the model as well, steering the model towards better predictive efficiency, as it were. That is the motivation behind the line of work that we broadly refer to as *conformal training* [8, 13, 61]. In a nutshell, conformal training introduces a differentiable, and hence approximate, conformal prediction step during training so that one can directly optimize for desired properties of the set predictor, most notably its predictive efficiency. In what follows we give an overview of conformal training, focusing on classification tasks.

The idea of conformal training has been proposed concomitantly in [8, 13]. Here we follow the approach in [8], where the key idea is to relax the prediction set defined by the model f and define "soft" prediction sets $\hat{\mathcal{C}}_f(x)$, which contain each of the labels $y \in \mathcal{Y}$ with a certain probability. That is, if $\mathcal{C}_f(x, y) \in \{0, 1\}$ is the hard assignment of label y to $\mathcal{C}(x)$, a corresponding soft version of this assignment can be defined as

$$\hat{\mathcal{C}}_f(x, y) := \sigma\left(\frac{\hat{q} - s_f(x, y)}{T}\right) \tag{24}$$

where $s_f(x, y)$ is the non-conformity score function defined by model f evaluated at (x, y), $\hat{q}$ is a thresholding value, σ is the logistic sigmoid function, and T is a temperature parameter controlling the smoothness of the soft assignment. Then we can define $\hat{\mathcal{C}}_f(x)$ as the vector collection of all the soft assignments $\hat{\mathcal{C}}_f(x, y)$ for all labels $y \in \mathcal{Y}$. Similarly the size of $\hat{\mathcal{C}}_f(x)$ can be naturally defined as

$$|\hat{\mathcal{C}}_f(x)| := \sum_{y \in \mathcal{Y}} \hat{\mathcal{C}}_f(x, y).$$

Bellotti [8] then proposes the following loss functions which are computed for each training batch $\mathcal{B}$

$$\mathcal{L}_{size}(f) = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} g\left(|\hat{\mathcal{C}}_f(x))|\right) \quad \mathcal{L}_{coverage}(f) = \left(\left[\frac{1}{|\mathcal{B}|} \sum_{(x,y) \in \mathcal{B}} \hat{\mathcal{C}}_f(x, y)\right] - (1 - \alpha)\right)^2,$$

where α is the desired coverage rate and g is a user-defined function of the prediction set size, e.g. the log function. Intuitively, $\mathcal{L}_{size}$ encourages small (efficient) prediction sets, whereas $\mathcal{L}_{coverage}$ penalizes deviations from the target coverage of $1 - \alpha$. Naturally, there is a trade-off between these two objectives, inefficiency and coverage, so both loss terms are optimized together with a hyperparameter λ governing the influence of each term:

$$\mathcal{L}(f) = \mathcal{L}_{size}(f) + \lambda \mathcal{L}_{coverage}(f). \tag{25}$$

Importantly, Bellotti [8] argues the choice of the threshold $\hat{q}$ in (24) is immaterial since the model can learn to shift its outputs (in logit space) accordingly to match the constraints in $\mathcal{L}_{coverage}(f)$. We can then directly optimize (25) via stochastic gradient descent methods during training since it is fully differentiable with respect to the model parameters.

Stutz et al. [61] build on the work of Bellotti [8] by noticing that the calibration step is an important component in conformal prediction that should also be accounted for during training. To that end, they propose to split each training batch $\mathcal{B}$ in two: the $\mathcal{B}_{cal}$ half used for calibration, and the $\mathcal{B}_{test}$ used for testing. Now, instead of using an arbitrary threshold $\hat{q}$, we compute it using the quantile of $\mathcal{B}_{cal}$, or concretely

$$\hat{q} = \text{Quantile}(1 - \alpha; \{s_f(x, y) : (x, y) \in \mathcal{B}_{cal}\})$$

With this modification, we no longer need to enforce valid coverage via $\mathcal{L}_{coverage}(f)$ and can optimize for low inefficiency directly by minimizing $\mathcal{L}_{size}(f)$ on $\mathcal{B}_{test}$. In that case, however, we only get a learning signal from $\mathcal{B}_{test}$, since the quantile operation applied to $\mathcal{B}_{cal}$ is non-differentiable. Stutz et al. [61] bypass that limitation via differentiable sorting operators [9, 17, 50], in particular via

a version of differentiable sorting networks. In our experiments, we considered both fast sort [9] and the monotonic differentiable sorting networks of [50] but finally chose the latter since they proved more stable and provided richer gradient signals.

This version of their approach, which we refer to as ConfTr, only optimizes the size loss, but Stutz et al. [61] also proposed another variant which includes a classification loss term as follows

$$\mathcal{L}_{class}(f) = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \sum_{\hat{y}\in\mathcal{Y}} L_{y,\hat{y}} \left[\left(1 - \hat{\mathcal{C}}_f(x,y)\right) \delta[\hat{y} = y] + \hat{\mathcal{C}}_f(x,y)\delta[\hat{y} \neq y] \right], \tag{26}$$

where δ is the indicator function, and L is a user-defined square matrix of size $|\mathcal{Y}|^2$ with $L_{y,\hat{y}}$ capturing some similarity notion between y and $\hat{y}$. In our experiments, as well as most experiments in the original paper [61], no prior information about the classification problem is assumed, in which case L is taken to be the identity matrix of size $|\mathcal{Y}|$. Therefore, we have two variants of conformal training as proposed in [61]: **ConfTr** that optimizes only $\mathcal{L}_{size}$, and **ConfTr$_{\text{class}}$** that jointly optimizes $\mathcal{L}_{size}$ and $\mathcal{L}_{class}$, both of which are included in our experiments.

Our own approach to conformal training follows the same recipe from [61], i.e., we also simulate a split conformal prediction step during training by splitting each training batch into two and using differentiable sorting operators (see Algorithm 1). The key difference is in how we define the training objectives, which we derive from first principles and standard information theory inequalities. Not only do our upper bounds, DPI (4), MB Fano (5) and Fano (6), outperform the ConfTr objectives in many cases, but they also do away with a few hyperparameters. Namely, the function g in $\mathcal{L}_{size}$, and hyperparameters controlling the relative importance of $\mathcal{L}_{size}$ and $\mathcal{L}_{class}$.

F.1 Deriving Conformal Training from Fano's bound

Through Proposition 3.2, we can connect Fano's bound for list decoding to the size loss from [61], proposed for conformal training. Assuming a uniform distribution for Q we can show that

$$\begin{aligned}
H(Y|X) &\leq h_b(\alpha) + \alpha \mathbb{E}_{E=0}\left[\log(|Y| - |\mathcal{C}(x)|)\right] + (1-\alpha_n)\,\mathbb{E}_{E=1}\left[\log|\mathcal{C}(x)|\right] \\
&\leq h_b(\alpha) + \alpha \log|Y| + (1-\alpha_n)\,\mathbb{E}_{E=1}\left[\log|\mathcal{C}(x)|\right] \\
&\leq h_b(\alpha) + \alpha \log|Y| + (1-\alpha_n)\log\mathbb{E}_{E=1}\left[|\mathcal{C}(x)|\right] \\
&\leq h_b(\alpha) + \alpha \log|Y| - (1-\alpha_n)\log(1-\alpha) + (1-\alpha_n)\log\mathbb{E}\left[|\mathcal{C}(x)|\right].
\end{aligned}$$

Note that in the first line we have the simple Fano bound, whereas in the last one we have the ConfTr objective, namely $\log\mathbb{E}\left[|\mathcal{C}(x)|\right]$, multiplied by $1-\alpha_n$ plus a constant that depends only on α. Therefore, we ground ConfTr as minimizing a looser upper bound to the true conditional entropy of the data than the simple Fano bound we provide in Corollary 3.1. Moreover, the simple Fano bound can be further improved with an appropriate choice of Q, for instance as given by the model f, in the model-based Fano bound of Proposition 3.2.

G Experiments

In this section, we present further experimental results for conformal training in the centralized and federated setting. We start by defining the splits and architectures used for each data set, which are listed in Table 6. In most aspects, we follow the experimental design of [61]. All experiments were conducted on commercially available NVIDIA GPUs using our own implementation in Python 3 and Pytorch [49], which can be found at `github.com/Qualcomm-AI-research/info_cp`. All data sets were retrieved directly from `torchvision` [40].

Table 6: Experimental settings for each data set, with $|\mathcal{D}_{train}|$, $|\mathcal{D}_{cal}|$ and $|\mathcal{D}_{test}|$ the sizes of train, calibration and test splits, respectively.

Data set	$\lvert\mathcal{D}_{train}\rvert$	$\lvert\mathcal{D}_{cal}\rvert$	$\lvert\mathcal{D}_{test}\rvert$	Epochs	Architecture
MNIST [29]	55K	5K	10K	50	1-layer MLP
Fashion-MNIST [69]	55K	5K	10K	150	2-layer MLP
EMNIST [11]	628K	70K	116K	75	2-layer MLP
CIFAR10 [25]	45K	5K	10K	150	ResNet-34
CIFAR100 [25]	45K	5K	10K	150	ResNet-50

Regarding the architectures, we also closely follow the experimental setup in [61]. For MNIST we have a simple linear model, whereas for Fashion-MNIST and EMNIST we use 2-layer MLPs with 64 and 128 hidden units for first and second layers, respectively. For the CIFAR data sets, we use the default ResNet implementations from *torchvision* [40], but changing the first convolution to have a kernel size of 3 and unitary stride and padding. We use Pytorch's default weight initialization strategy for all architectures. For all datasets, we use a regular SGD optimizer with momentum 0.9 and Nesterov gradients, accompanied by a step scheduler multiplying the initial learning rate by 0.1 after $2/5$, $3/5$ and $4/5$ of the total number of epochs. We only use data augmentations on the CIFAR datasets, and differently from [61], we only apply random flipping and cropping for both CIFAR10 and CIFAR100.

G.1 Centralized Setting

We followed the experimental procedure of [61], and for each dataset and each method, we ran a grid search over the following hyperparameters using *ray tune* [45]:

- **Batch size** with possible values in $\{100, 500, 1000\}$.
- **Learning rate** with possible values in $\{0.05, 0.01, 0.005\}$.
- **Temperature** used in relaxing the construction of prediction sets at training time. We considered temperature values in $\{0.01, 0.1, 0.5, 1.0\}$.
- **Steepness** of the differentiable sorting algorithm (monotonic sorting networks with Cauchy distribution [50]), which regulates the smoothness of the sorting operator; the higher the steepness value, the closer the differentiable sorting operator is to standard sorting. We considered steepness values in $\{1, 10, 100\}$.

In Tables 7, 8, 9, 10 and 11, we report the best hyperparameters found for each dataset and method as well as the average prediction set size for threshold CP [56] computed in the probability domain (THR) and APS [55], as well as the test accuracy. Importantly, similarly to [61], in all cases we only train the models to optimize threshold CP with log-probabilities. We confirm the observation in [61] that other methods, and notably APS, are unstable during training, probably because it forces us to operate in the probability domain, as opposed to the more optimization-friendly logits or log-probabilities. Nevertheless, we still select hyperparameters according to the best performance with respect to each CP method, and that is why we have different optimal hyperparameters for THR and APS for each data set and each conformal training objective. We note ConfTr and ConfTr$_{\text{class}}$ require extra hyperparameters like the target size and weights attributed to each loss term (see Appendix F). For those hyperparameters, we use the best values for each data set as reported in [61].

As described in the main paper, we use the default train and test splits of each data set but transfer 10% of the training data to the test data set. We train the classifiers only on the remaining 90% of the

training data and, at test time, run SCP with 10 different calibration/test splits by randomly splitting the enlarged test data set. All results reported in the paper are given by the average ($\pm$ one standard deviation) computed across these 10 random splits. Crucially, to avoid overfitting to the test data, the grid search was done solely on the 90% of the training data not used for testing.

Table 7: Hyperparameter Search for MNIST.

Bound	Optimized for	batch size	lr	temperature	steepness	THR	APS	Test Acc.
CE	THR	100	0.01	-	-	$2.29_{\pm 0.18}$	$2.50_{\pm 0.08}$	0.93
	APS	500	0.05	-	-	$2.28_{\pm 0.19}$	$2.50_{\pm 0.08}$	0.93
ConfTr	THR	500	0.05	0.1	100	$6.28_{\pm 0.71}$	$9.81_{\pm 0.06}$	0.90
	APS	500	0.01	1.0	100	$2.08_{\pm 0.10}$	$2.10_{\pm 0.07}$	0.90
ConfTr-class	THR	500	0.005	1.0	10	$2.09_{\pm 0.11}$	$2.15_{\pm 0.12}$	0.91
	APS	500	0.01	0.1	100	$2.12_{\pm 0.11}$	$2.13_{\pm 0.13}$	0.90
Fano	THR	100	0.005	0.1	100	$2.09_{\pm 0.12}$	$2.12_{\pm 0.09}$	0.91
	APS	100	0.01	0.5	100	$2.11_{\pm 0.08}$	$2.12_{\pm 0.08}$	0.91
MB Fano	THR	1000	0.005	0.1	100	$2.24_{\pm 0.12}$	$2.96_{\pm 0.11}$	0.91
	APS	100	0.005	0.1	10	$2.43_{\pm 0.22}$	$2.49_{\pm 0.19}$	0.92
DPI	THR	1000	0.005	0.01	100	$2.24_{\pm 0.17}$	$2.87_{\pm 0.12}$	0.92
	APS	500	0.05	0.1	100	$2.29_{\pm 0.17}$	$2.64_{\pm 0.07}$	0.92

Table 8: Hyperparameter Search for Fashion-MNIST.

Bound	Optimized for	batch size	lr	temperature	steepness	THR	APS	Test Acc.
CE	THR	1000	0.005	-	-	$2.39_{\pm 0.13}$	$2.74_{\pm 0.20}$	0.87
	APS	100	0.05	-	-	$2.18_{\pm 0.17}$	$2.41_{\pm 0.17}$	0.89
ConfTr	THR	100	0.01	1.0	1.0	$1.73_{\pm 0.06}$	$1.89_{\pm 0.09}$	0.89
	APS	100	0.01	1.0	1.0	$1.73_{\pm 0.06}$	$1.89_{\pm 0.09}$	0.89
ConfTr-class	THR	100	0.01	1.0	10.0	$5.11_{\pm 0.49}$	$8.28_{\pm 3.17}$	0.88
	APS	500	0.005	1.0	100.0	$1.79_{\pm 0.06}$	$1.79_{\pm 0.07}$	0.88
Fano	THR	100	0.05	0.5	100.0	$1.70_{\pm 0.05}$	$2.06_{\pm 0.08}$	0.88
	APS	100	0.01	1.0	100.0	$1.76_{\pm 0.04}$	$1.87_{\pm 0.05}$	0.87
MB Fano	THR	100	0.05	0.5	10.0	$1.80_{\pm 0.08}$	$2.71_{\pm 0.14}$	0.89
	APS	100	0.05	1.0	100.0	$1.84_{\pm 0.10}$	$2.25_{\pm 0.14}$	0.89
DPI	THR	100	0.05	0.01	10.0	$1.73_{\pm 0.07}$	$2.12_{\pm 0.08}$	0.89
	APS	100	0.05	0.01	100.0	$1.75_{\pm 0.05}$	$2.08_{\pm 0.06}$	0.89

Table 9: Hyperparameter Search for EMNIST.

Bound	Optimized for	batch size	lr	temperature	steepness	THR	APS	Test Acc.
CE	THR	100	0.01	-	-	$2.06_{\pm 0.11}$	$3.37_{\pm 0.15}$	0.86
	APS	100	0.005	-	-	$2.06_{\pm 0.10}$	$3.40_{\pm 0.18}$	0.86
ConfTr	THR	100	0.01	0.1	100	$1.99_{\pm 0.10}$	$4.94_{\pm 0.26}$	0.85
	APS	100	0.005	1.0	100	$1.99_{\pm 0.08}$	$2.36_{\pm 0.11}$	0.83
ConfTr-class	THR	100	0.005	0.1	100	$2.01_{\pm 0.09}$	$5.07_{\pm 0.24}$	0.85
	APS	100	0.05	1.0	100	$1.99_{\pm 0.10}$	$2.38_{\pm 0.11}$	0.84
Fano	THR	100	0.01	0.1	100	$2.10_{\pm 0.11}$	$8.27_{\pm 0.53}$	0.84
	APS	100	0.01	1.0	100	$2.05_{\pm 0.11}$	$2.75_{\pm 0.14}$	0.82
MB Fano	THR	100	0.01	1.0	100	$2.01_{\pm 0.11}$	$5.21_{\pm 0.26}$	0.86
	APS	100	0.01	0.1	1	$2.94_{\pm 0.17}$	$3.67_{\pm 0.13}$	0.86
DPI	THR	100	0.05	0.5	100	$1.98_{\pm 0.09}$	$3.86_{\pm 0.20}$	0.86
	APS	100	0.05	0.1	100	$2.04_{\pm 0.12}$	$4.07_{\pm 0.23}$	0.86

Table 10: Hyperparameter Search for CIFAR10.

Bound	Optimized for	batch size	lr	temperature	steepness	THR	APS	Test Acc.
CE	THR	100	0.05	-	-	$1.69_{\pm 0.11}$	$2.12_{\pm 0.21}$	0.93
	APS	100	0.05	-	-	$1.74_{\pm 0.07}$	$2.34_{\pm 0.22}$	0.93
ConfTr	THR	100	0.05	0.5	10	$9.90_{\pm 0.02}$	$10.00_{\pm 0.00}$	0.10
	APS	1000	0.005	0.1	1	$9.90_{\pm 0.01}$	$9.98_{\pm 0.00}$	0.10
ConfTr-class	THR	100	0.01	0.5	10	$2.16_{\pm 0.09}$	$2.19_{\pm 0.10}$	0.86
	APS	100	0.01	0.5	10	$2.13_{\pm 0.08}$	$2.18_{\pm 0.06}$	0.86
Fano	THR	100	0.01	1.0	1	$2.05_{\pm 0.05}$	$2.34_{\pm 0.09}$	0.89
	APS	100	0.01	1.0	1	$2.06_{\pm 0.10}$	$2.35_{\pm 0.10}$	0.89
MB Fano	THR	100	0.05	0.5	100	$1.66_{\pm 0.09}$	$2.40_{\pm 0.08}$	0.92
	APS	100	0.01	1.0	10	$1.69_{\pm 0.09}$	$1.89_{\pm 0.06}$	0.91
DPI	THR	100	0.05	0.01	100	$1.64_{\pm 0.07}$	$1.88_{\pm 0.05}$	0.92
	APS	100	0.005	0.01	10	$1.79_{\pm 0.12}$	$1.97_{\pm 0.08}$	0.91

Table 11: Hyperparameter Search for CIFAR100.

Bound	Optimized for	batch size	lr	temperature	steepness	THR	APS	Test Acc.
CE	THR	100	0.05	-	-	$19.70_{\pm 2.05}$	$26.02_{\pm 1.31}$	0.72
	APS	100	0.05	-	-	$19.70_{\pm 2.05}$	$26.02_{\pm 1.31}$	0.72
ConfTr	THR	100	0.005	1.0	1	$32.80_{\pm 2.75}$	$34.09_{\pm 2.54}$	0.52
	APS	100	0.01	0.5	1	$30.04_{\pm 1.36}$	$40.58_{\pm 1.23}$	0.53
ConfTr-class	THR	100	0.01	1.0	1	$66.48_{\pm 3.67}$	$54.30_{\pm 17.12}$	0.43
	APS	100	0.01	1	10	$33.32_{\pm 1.89}$	$32.91_{\pm 1.53}$	0.37
Fano	THR	100	0.05	0.5	1	$40.30_{\pm 1.10}$	$48.21_{\pm 1.26}$	0.42
	APS	500	0.05	1	1	$30.43_{\pm 1.61}$	$33.80_{\pm 0.93}$	0.57
MB Fano	THR	100	0.05	0.5	100	$14.61_{\pm 0.84}$	$21.69_{\pm 0.71}$	0.70
	APS	100	0.05	1	10	$16.36_{\pm 0.93}$	$21.68_{\pm 1.44}$	0.68
DPI	THR	100	0.05	1.0	1	$17.55_{\pm 1.33}$	$20.13_{\pm 0.78}$	0.69
	APS	100	0.05	1	10	$14.90_{\pm 0.80}$	$17.41_{\pm 0.62}$	0.70

G.1.1 Results with Regularized Adaptive Prediction Sets (RAPS)

We report additional results with a conformal method known as RAPS (regularized adaptive prediction sets) [4]. In a nutshell, RAPS works exactly like APS but adds a penalization term λ_{reg} to the score of each label that would make the prediction set larger than k_{reg}. In Table 12, we have results with RAPS hyperparameters set as k_{reg}=1 and λ_{reg}=0.01, where we can see a pattern of performance across conformal training objectives that is similar to what we previously observed for THR and APS. As expected, the regularization introduced by RAPS results in prediction sets that are smaller than those obtained via APS in almost all cases. Again, similarly to [61], in all of our experiments we perform conformal training by thresholding on log-probabilities, which we observed to work best. The results we report for thresholding on probabilities (THR), APS and now RAPS show that this translates well to other score functions as well. Nonetheless, the improvement we get via RAPS still does not offset the gains from conformal training. In most cases, RAPS applied to a model trained via CE produces less efficient prediction sets than APS applied to a model trained to optimize one of our bounds.

We also study the impact of the two hyperparameter in RAPS, namely λ_{reg} and k_{reg} in the inefficiency of each of the conformal training methods. We concentrate our analysis on CIFAR100 and report results with varying k_{reg} in Table 13 and varying λ_{reg} in Table 14. Interestingly, RAPS was much more sensitive to variations in k_{reg} than in λ_{reg}. Still, we observe models trained via our model-based Fano and DPI bounds produce more efficient prediction sets than the baselines in all cases.

G.2 Federated Setting

In the federated setting, we run conformal training exactly in the same fashion, but including the additional $Q_{Z|X}$ term in (11) to get the proper distributed bound that can be optimized locally and independently in each device. We optimize the conformal training objective with SGD for one epoch in each device, and then communicate the resulting "personalized" model to the server, which aggregates the model parameters of each device via federated averaging [42]. After aggregation, the

Table 12: **Inefficiency restuls with RAPS**. Average prediction set size with both APS and RAPS ($k_{\text{reg}} = 1$ and $\lambda_{\text{reg}} = 0.01$) methods for each conformal training objective and data set for a target $\alpha = 0.01$. We report mean and standard deviation across 10 different calib./test splits and show in bold all values within one standard deviation of the best result. Both APS and RAPS results were computed with the same classifier. Lower is better.

Method	MNIST		F-MNIST		EMNIST		CIFAR 10		CIFAR 100	
	APS	RAPS	APS	RAPS	APS	RAPS	APS	RAPS	APS	RAPS
CE	$2.50_{\pm 0.08}$	$2.35_{\pm 0.16}$	$2.41_{\pm 0.17}$	$2.40_{\pm 0.17}$	$3.40_{\pm 0.18}$	$3.07_{\pm 0.16}$	$2.34_{\pm 0.22}$	$2.31_{\pm 0.22}$	$26.02_{\pm 1.31}$	$26.02_{\pm 1.31}$
ConfTr	$\mathbf{2.10_{\pm 0.07}}$	$\mathbf{2.10_{\pm 0.07}}$	$\mathbf{1.89_{\pm 0.09}}$	$1.84_{\pm 0.09}$	$\mathbf{2.36_{\pm 0.11}}$	$\mathbf{2.19_{\pm 0.06}}$	$9.99_{\pm 0.00}$	$9.99_{\pm 0.00}$	$40.59_{\pm 1.23}$	$39.87_{\pm 1.02}$
$\text{ConfTr}_{\text{class}}$	$\mathbf{2.13_{\pm 0.13}}$	$\mathbf{2.13_{\pm 0.13}}$	$\mathbf{1.79_{\pm 0.09}}$	$\mathbf{1.79_{\pm 0.09}}$	$\mathbf{2.38_{\pm 0.11}}$	$\mathbf{2.21_{\pm 0.07}}$	$2.18_{\pm 0.06}$	$2.18_{\pm 0.06}$	$32.91_{\pm 1.53}$	$32.94_{\pm 1.54}$
Fano	$\mathbf{2.12_{\pm 0.08}}$	$\mathbf{2.12_{\pm 0.08}}$	$\mathbf{1.87_{\pm 0.05}}$	$1.80_{\pm 0.04}$	$\mathbf{2.75_{\pm 0.14}}$	$2.50_{\pm 0.07}$	$2.35_{\pm 0.10}$	$2.20_{\pm 0.04}$	$33.80_{\pm 0.93}$	$33.75_{\pm 1.02}$
MB Fano	$\mathbf{2.49_{\pm 0.19}}$	$\mathbf{2.48_{\pm 0.19}}$	$\mathbf{2.25_{\pm 0.14}}$	$2.07_{\pm 0.11}$	$3.67_{\pm 0.13}$	$3.67_{\pm 0.13}$	$\mathbf{1.89_{\pm 0.06}}$	$\mathbf{1.78_{\pm 0.07}}$	$\mathbf{21.68_{\pm 1.44}}$	$20.97_{\pm 1.30}$
DPI	$\mathbf{2.64_{\pm 0.07}}$	$2.44_{\pm 0.12}$	$\mathbf{2.08_{\pm 0.06}}$	$1.95_{\pm 0.05}$	$\mathbf{4.07_{\pm 0.23}}$	$3.74_{\pm 0.17}$	$\mathbf{1.97_{\pm 0.08}}$	$\mathbf{1.87_{\pm 0.11}}$	$\mathbf{17.41_{\pm 0.62}}$	$\mathbf{17.25_{\pm 0.58}}$

Table 13: **Inefficiency results for RAPS with varying k_{reg}**. Average prediction set size for $\lambda_{\text{reg}} = 0.01$ and varying k_{reg} on CIFAR100 for a target $\alpha = 0.01$. We report mean and standard deviation across 10 different calib./test splits and show in bold all values within one standard deviation of the best result. Both APS and RAPS results were computed with the same classifier. Lower is better.

Method	$k_{\text{reg}} = 0$ (APS)	$k_{\text{reg}} = 1$	$k_{\text{reg}} = 3$	$k_{\text{reg}} = 5$	$k_{\text{reg}} = 10$	$k_{\text{reg}} = 15$
CE	$26.02_{\pm 1.31}$	$26.02_{\pm 1.31}$	$22.43_{\pm 1.88}$	$20.43_{\pm 1.70}$	$18.95_{\pm 1.18}$	$20.63_{\pm 0.89}$
ConfTr	$40.59_{\pm 1.23}$	$39.87_{\pm 1.02}$	$38.73_{\pm 0.84}$	$38.31_{\pm 0.70}$	$38.06_{\pm 0.48}$	$38.06_{\pm 0.59}$
$\text{ConfTr}_{\text{class}}$	$32.91_{\pm 1.53}$	$32.94_{\pm 1.54}$	$32.94_{\pm 1.53}$	$32.95_{\pm 1.53}$	$33.16_{\pm 1.47}$	$33.74_{\pm 1.39}$
Fano	$33.80_{\pm 0.93}$	$33.75_{\pm 1.02}$	$32.83_{\pm 1.60}$	$31.71_{\pm 1.55}$	$30.82_{\pm 1.72}$	$30.31_{\pm 1.71}$
MB Fano	$21.68_{\pm 1.44}$	$20.97_{\pm 1.30}$	$19.32_{\pm 1.40}$	$19.04_{\pm 1.36}$	$\mathbf{17.35_{\pm 0.86}}$	$\mathbf{17.68_{\pm 0.94}}$
DPI	$\mathbf{17.41_{\pm 0.62}}$	$\mathbf{17.25_{\pm 0.58}}$	$\mathbf{17.02_{\pm 0.75}}$	$\mathbf{16.40_{\pm 0.93}}$	$\mathbf{17.68_{\pm 0.82}}$	$19.49_{\pm 2.08}$

Table 14: **Inefficiency results for RAPS with varying λ_{reg}**. Average prediction set size for $k_{\text{reg}} = 1$ and varying λ_{reg} on CIFAR100 for a target $\alpha = 0.01$. We report mean and standard deviation across 10 different calib./test splits and show in bold all values within one standard deviation of the best result. Both APS and RAPS results were computed with the same classifier. Lower is better.

Method	$\lambda_{\text{reg}} = 0$ (APS)	$\lambda_{\text{reg}} = 0.01$	$\lambda_{\text{reg}} = 0.03$	$\lambda_{\text{reg}} = 0.1$	$\lambda_{\text{reg}} = 0.3$	$\lambda_{\text{reg}} = 1.0$
CE	$26.02_{\pm 1.31}$	$26.02_{\pm 1.31}$	$26.02_{\pm 1.31}$	$26.02_{\pm 1.31}$	$26.02_{\pm 1.31}$	$26.02_{\pm 1.31}$
ConfTr	$40.59_{\pm 1.23}$	$39.87_{\pm 1.02}$	$39.17_{\pm 1.05}$	$39.09_{\pm 1.12}$	$39.09_{\pm 1.12}$	$39.09_{\pm 1.12}$
$\text{ConfTr}_{\text{class}}$	$32.91_{\pm 1.53}$	$32.94_{\pm 1.54}$	$32.94_{\pm 1.54}$	$32.93_{\pm 1.53}$	$32.94_{\pm 1.53}$	$32.94_{\pm 1.55}$
Fano	$33.80_{\pm 0.93}$	$33.75_{\pm 1.02}$	$33.75_{\pm 1.02}$	$33.75_{\pm 1.02}$	$33.75_{\pm 1.02}$	$33.75_{\pm 1.02}$
MB Fano	$21.68_{\pm 1.44}$	$20.97_{\pm 1.30}$	$20.96_{\pm 1.29}$	$20.96_{\pm 1.29}$	$20.96_{\pm 1.29}$	$20.96_{\pm 1.29}$
DPI	$\mathbf{17.41_{\pm 0.62}}$	$\mathbf{17.25_{\pm 0.58}}$	$\mathbf{17.25_{\pm 0.58}}$	$\mathbf{17.25_{\pm 0.58}}$	$\mathbf{17.25_{\pm 0.58}}$	$\mathbf{17.25_{\pm 0.58}}$

global model thus computed is communicated to the client and the process restarts. We do 5K such communication rounds for EMNIST, and 10K for CIFAR10 and CIFAR100. As described in the main text, we divide the data among 100, 500, and 1K clients for CIFAR10, CIFAR100 and EMNIST, respectively. We assign data points to devices imposing a *distribution-based label imbalance* [33], where we sample a marginal label distribution for each device from a Dirichlet distribution. We use $\text{Dir}(1.0)$ for all experiments, but also study the effect of $\text{Dir}(0.5)$ and $\text{Dir}(0.1)$ on CIFAR10, as shown in Tables 15 and 16. The remainder of the experimental setup is similar to that used for the centralized setting, with the same architectures, data augmentation strategies, and optimal hyperparameters reported in Tables 7, 8, 9, 10. In Tables 15 and 16, as well as 17 and 18, we report inefficiency results for two different settings, global (GLO) and personalized (PER).

GLO We run SCP with the final global model assuming calibration and test data sets at the server, or equivalently that the clients share their scores with the server. This reflects the best inefficiency results we can hope for with the global model, as in practice we might need to resort to privacy-preserving methods that are likely to hurt performance. Notably, we considered the quantile of quantiles approach of Humbert et al. [24] as well as the simpler alternative proposed in [37], but in

both cases we got varying degrees of coverage due to the data heterogeneity among devices introduced in the *distribution-based label imbalance* setup. Addressing these shortcomings is a promising avenue for future research for conformal prediction in the federated setting.

PER After learning the global model, we fine-tune it on the local training data of each device to obtain a personalized model. We then run SCP individually for each device with local calibration and test data sets and report the average prediction sets across all clients. Importantly, since each client has access to only a small number of data points, we do not always achieve valid coverage in the personalized setting. More precisely, all personalized models on CIFAR10 and EMNIST achieved marginal coverage of around 97%, while for CIFAR100 that value dropped to 90%. Nonetheless, all methods get similar coverages so the results remain comparable. Interestingly, even with personalized models, which implicitly already estimate $Q_{Y|X,Z}$, updating the personalized model as in (9) with the global head $Q_{Z|X,Y}$, still results in non-negligible improvements in performance, as shown in Tables 15 and 16, as well as 17 and 18.

Table 15: **Inefficiency results for different degrees of data heterogeneity in FL with THR**. Average prediction set size with THR for CIFAR10 with different levels of data heterogeneity among clients for both global (GLO) and personalized (PER) models in the federated setting for a target $\alpha = 0.01$. We use $_{+\text{SI}}$ to indicate the inclusion of side information. Lower is better.

Method	Dirichlet(1.0)				Dirichlet(0.5)				Dirichlet(0.1)			
	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$
CE	$2.73_{\pm 0.04}$	$2.30_{\pm 0.06}$	$2.07_{\pm 0.51}$	$1.82_{\pm 0.39}$	$2.81_{\pm 0.11}$	$2.41_{\pm 0.08}$	$2.03_{\pm 0.49}$	$1.79_{\pm 0.38}$	$2.73_{\pm 0.13}$	$2.32_{\pm 0.06}$	$2.05_{\pm 0.52}$	$1.82_{\pm 0.38}$
ConfTr	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.84_{\pm 0.05}$	$9.84_{\pm 0.05}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.84_{\pm 0.05}$	$9.84_{\pm 0.05}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.87_{\pm 0.06}$	$9.87_{\pm 0.06}$
ConfTr$_{\text{class}}$	$3.53_{\pm 0.09}$	$3.39_{\pm 0.08}$	$2.86_{\pm 0.38}$	$2.79_{\pm 0.35}$	$3.54_{\pm 0.06}$	$3.38_{\pm 0.07}$	$2.91_{\pm 0.40}$	$2.84_{\pm 0.37}$	$3.53_{\pm 0.05}$	$3.39_{\pm 0.05}$	$2.84_{\pm 0.41}$	$2.78_{\pm 0.38}$
Fano	$2.39_{\pm 0.07}$	$2.07_{\pm 0.07}$	$1.84_{\pm 0.37}$	$1.63_{\pm 0.30}$	$2.47_{\pm 0.08}$	$2.10_{\pm 0.06}$	$1.94_{\pm 0.43}$	$1.71_{\pm 0.35}$	$2.46_{\pm 0.08}$	$2.11_{\pm 0.06}$	$1.91_{\pm 0.41}$	$1.68_{\pm 0.33}$
MB Fano	$2.52_{\pm 0.08}$	$2.04_{\pm 0.07}$	$1.56_{\pm 0.29}$	$1.40_{\pm 0.22}$	$2.66_{\pm 0.12}$	$2.09_{\pm 0.06}$	$1.57_{\pm 0.31}$	$1.40_{\pm 0.23}$	$2.55_{\pm 0.11}$	$2.06_{\pm 0.10}$	$1.98_{\pm 0.45}$	$1.64_{\pm 0.32}$
DPI	$2.76_{\pm 0.07}$	$2.28_{\pm 0.03}$	$1.64_{\pm 0.36}$	$1.49_{\pm 0.28}$	$2.50_{\pm 0.07}$	$2.11_{\pm 0.49}$	$1.91_{\pm 0.42}$	$1.64_{\pm 0.31}$	$2.51_{\pm 0.07}$	$2.07_{\pm 0.05}$	$2.04_{\pm 0.41}$	$1.70_{\pm 0.30}$

Table 16: **Inefficiency results for different degrees of data heterogeneity in FL with APS**. Average prediction set size with APS for CIFAR10 with different levels of data heterogeneity among clients for both global (GLO) and personalized (PER) models in the federated setting for a target $\alpha = 0.01$. We use $_{+\text{SI}}$ to indicate the inclusion of side information. Lower is better.

Method	Dirichlet(1.0)				Dirichlet(0.5)				Dirichlet(0.1)			
	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$
CE	$2.83_{\pm 0.07}$	$2.43_{\pm 0.06}$	$2.22_{\pm 0.06}$	$1.94_{\pm 0.41}$	$2.70_{\pm 0.79}$	$2.33_{\pm 0.42}$	$2.13_{\pm 0.47}$	$1.87_{\pm 0.36}$	$2.81_{\pm 0.14}$	$2.37_{\pm 0.08}$	$2.15_{\pm 0.05}$	$1.89_{\pm 0.39}$
ConfTr	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.87_{\pm 0.14}$	$9.87_{\pm 0.14}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.88_{\pm 0.09}$	$9.88_{\pm 0.09}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.87_{\pm 0.14}$	$9.87_{\pm 0.14}$
ConfTr$_{\text{class}}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.92_{\pm 0.03}$	$9.92_{\pm 0.03}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.94_{\pm 0.07}$	$9.95_{\pm 0.05}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.92_{\pm 0.03}$	$9.92_{\pm 0.03}$
Fano	$2.73_{\pm 0.07}$	$2.39_{\pm 0.06}$	$2.17_{\pm 0.45}$	$1.94_{\pm 0.36}$	$2.61_{\pm 0.07}$	$2.30_{\pm 0.05}$	$2.15_{\pm 0.04}$	$1.92_{\pm 0.34}$	$2.67_{\pm 0.08}$	$2.35_{\pm 0.07}$	$2.13_{\pm 0.42}$	$1.92_{\pm 0.34}$
MB Fano	$2.79_{\pm 0.13}$	$2.33_{\pm 0.05}$	$2.28_{\pm 0.48}$	$1.97_{\pm 0.35}$	$2.71_{\pm 0.09}$	$2.31_{\pm 0.07}$	$2.27_{\pm 0.48}$	$1.96_{\pm 0.35}$	$2.87_{\pm 0.12}$	$2.40_{\pm 0.06}$	$2.32_{\pm 0.51}$	$1.99_{\pm 0.39}$
DPI	$2.68_{\pm 0.15}$	$2.22_{\pm 0.09}$	$2.11_{\pm 0.51}$	$1.84_{\pm 0.36}$	$2.54_{\pm 0.10}$	$2.18_{\pm 0.04}$	$2.10_{\pm 0.45}$	$1.84_{\pm 0.33}$	$2.65_{\pm 0.10}$	$2.25_{\pm 0.08}$	$2.13_{\pm 0.44}$	$1.89_{\pm 0.34}$

Table 17: **Inefficiency results for global (GLO) and personalized (PER) models with THR**. Average prediction set size with THR for federated conformal training with a target $\alpha = 0.01$. We use $_{+\text{SI}}$ to indicate the inclusion of side information. We show in bold all values within one standard deviation of the best result. Lower is better.

Method	EMNIST				CIFAR 10				CIFAR 100			
	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$
CE	$2.91_{\pm 0.02}$	$2.46_{\pm 0.02}$	$\mathbf{2.46_{\pm 0.65}}$	$\mathbf{2.08_{\pm 0.51}}$	$2.73_{\pm 0.04}$	$2.30_{\pm 0.06}$	$\mathbf{2.07_{\pm 0.51}}$	$\mathbf{1.82_{\pm 0.39}}$	$55.41_{\pm 1.09}$	$52.31_{\pm 1.03}$	$\mathbf{20.84_{\pm 8.02}}$	$\mathbf{18.85_{\pm 7.53}}$
ConfTr	$4.60_{\pm 0.05}$	$3.30_{\pm 0.02}$	$\mathbf{3.95_{\pm 1.05}}$	$3.05_{\pm 0.78}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.84_{\pm 0.05}$	$9.84_{\pm 0.05}$	$\mathbf{45.60_{\pm 1.30}}$	$\mathbf{41.18_{\pm 1.16}}$	$\mathbf{16.23_{\pm 6.27}}$	$\mathbf{14.01_{\pm 5.66}}$
ConfTr$_{\text{class}}$	$2.88_{\pm 0.02}$	$\mathbf{1.98_{\pm 0.02}}$	$\mathbf{1.69_{\pm 0.31}}$	$\mathbf{1.55_{\pm 0.24}}$	$3.53_{\pm 0.09}$	$3.39_{\pm 0.08}$	$2.86_{\pm 0.38}$	$2.79_{\pm 0.35}$	$58.53_{\pm 1.40}$	$56.03_{\pm 1.29}$	$\mathbf{25.22_{\pm 7.40}}$	$\mathbf{23.87_{\pm 7.02}}$
Fano	$\mathbf{2.63_{\pm 0.02}}$	$2.37_{\pm 0.02}$	$\mathbf{2.14_{\pm 0.46}}$	$\mathbf{1.98_{\pm 0.40}}$	$\mathbf{2.39_{\pm 0.07}}$	$\mathbf{2.07_{\pm 0.07}}$	$\mathbf{1.84_{\pm 0.37}}$	$\mathbf{1.63_{\pm 0.30}}$	$\mathbf{47.91_{\pm 1.20}}$	$\mathbf{41.19_{\pm 1.02}}$	$\mathbf{18.42_{\pm 6.95}}$	$\mathbf{14.55_{\pm 5.83}}$
MB Fano	$2.84_{\pm 0.04}$	$2.25_{\pm 0.03}$	$\mathbf{2.47_{\pm 0.64}}$	$\mathbf{1.96_{\pm 0.45}}$	$\mathbf{2.52_{\pm 0.08}}$	$\mathbf{2.04_{\pm 0.07}}$	$\mathbf{1.56_{\pm 0.29}}$	$\mathbf{1.40_{\pm 0.22}}$	$52.94_{\pm 1.40}$	$46.97_{\pm 1.30}$	$\mathbf{20.36_{\pm 7.77}}$	$\mathbf{16.41_{\pm 6.79}}$
DPI	$\mathbf{2.60_{\pm 0.02}}$	$2.23_{\pm 0.01}$	$\mathbf{2.31_{\pm 0.58}}$	$\mathbf{1.97_{\pm 0.46}}$	$2.76_{\pm 0.07}$	$2.28_{\pm 0.03}$	$\mathbf{1.64_{\pm 0.36}}$	$\mathbf{1.49_{\pm 0.28}}$	$52.36_{\pm 0.95}$	$48.64_{\pm 0.70}$	$\mathbf{20.13_{\pm 7.84}}$	$\mathbf{17.73_{\pm 7.23}}$

Table 18: **Inefficiency results for global (GLO) and personalized (PER) models with APS**. Average prediction set size with APS for federated conformal training with a target $\alpha = 0.01$. We use $_{+\text{SI}}$ to indicate the inclusion of side information. We show in bold all values within one standard deviation of the best result. Lower is better.

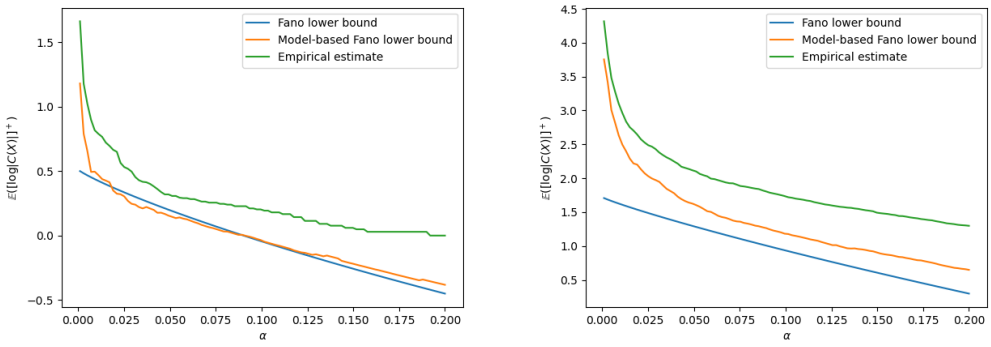
Method	EMNIST				CIFAR 10				CIFAR 100			
	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$	GLO	GLO$_{+\text{SI}}$	PER	PER$_{+\text{SI}}$
CE	$3.69_{\pm 0.03}$	$3.14_{\pm 0.04}$	$1.42_{\pm 0.25}$	$1.40_{\pm 0.24}$	$\mathbf{2.83_{\pm 0.07}}$	$2.43_{\pm 0.06}$	$\mathbf{2.22_{\pm 0.56}}$	$\mathbf{1.94_{\pm 0.41}}$	$64.73_{\pm 0.34}$	$62.67_{\pm 3.68}$	$\mathbf{22.39_{\pm 8.86}}$	$\mathbf{20.22_{\pm 8.15}}$
ConfTr	$6.14_{\pm 0.04}$	$5.25_{\pm 0.04}$	$4.79_{\pm 1.23}$	$3.06_{\pm 0.63}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.87_{\pm 0.14}$	$9.87_{\pm 0.14}$	$55.18_{\pm 2.10}$	$47.58_{\pm 1.48}$	$\mathbf{23.43_{\pm 8.20}}$	$\mathbf{19.12_{\pm 7.14}}$
ConfTr$_{\text{class}}$	$\mathbf{2.65_{\pm 0.02}}$	$\mathbf{2.42_{\pm 0.02}}$	$3.03_{\pm 1.57}$	$1.51_{\pm 0.31}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$9.92_{\pm 0.03}$	$9.92_{\pm 0.03}$	$99.92_{\pm 0.02}$	$99.91_{\pm 0.01}$	$99.97_{\pm 0.22}$	$99.68_{\pm 0.24}$
Fano	$3.12_{\pm 0.04}$	$2.72_{\pm 0.03}$	$\mathbf{1.17_{\pm 0.10}}$	$\mathbf{1.15_{\pm 0.01}}$	$\mathbf{2.73_{\pm 0.07}}$	$\mathbf{2.39_{\pm 0.06}}$	$\mathbf{2.17_{\pm 0.45}}$	$\mathbf{1.94_{\pm 0.36}}$	$\mathbf{46.95_{\pm 0.67}}$	$\mathbf{42.75_{\pm 0.91}}$	$\mathbf{19.19_{\pm 7.24}}$	$\mathbf{16.9_{\pm 6.67}}$
MB Fano	$4.75_{\pm 0.03}$	$\mathbf{2.43_{\pm 0.01}}$	$2.04_{\pm 0.51}$	$2.31_{\pm 0.42}$	$\mathbf{2.79_{\pm 0.13}}$	$\mathbf{2.33_{\pm 0.05}}$	$\mathbf{2.28_{\pm 0.48}}$	$\mathbf{1.97_{\pm 0.35}}$	$50.72_{\pm 1.77}$	$45.72_{\pm 1.38}$	$\mathbf{21.12_{\pm 7.49}}$	$\mathbf{18.12_{\pm 6.84}}$
DPI	$2.98_{\pm 0.03}$	$2.58_{\pm 0.02}$	$2.73_{\pm 0.65}$	$1.29_{\pm 0.16}$	$\mathbf{2.68_{\pm 0.15}}$	$\mathbf{2.22_{\pm 0.09}}$	$\mathbf{2.11_{\pm 0.51}}$	$\mathbf{1.84_{\pm 0.36}}$	$51.29_{\pm 1.07}$	$47.18_{\pm 1.27}$	$\mathbf{20.76_{\pm 7.51}}$	$\mathbf{18.15_{\pm 6.82}}$

G.3 Evaluating the lower bounds on the set size

In this section, we evaluate our two lower bounds on the expected $[\log |C(X)|]^+$. The first one can be obtained by rearranging the simple Fano bound (c.f. Proposition C.7) whereas the second one can be obtained from the model-based Fano entropy upper bounds (c.f. Theorem E.2). The main challenge in evaluating these bounds is in that we require the ground truth entropy $H(Y|X)$ or $H(Y|\hat{Y})$ which in general are not available. To proceed, we adopt the versions of the bounds that depend on $H(Y|\hat{Y})$. When $\hat{Y}$ is discrete, we can get a tractable lower bound to $H(Y|\hat{Y})$ via a maximum likelihood estimate of the entropy [46]. More specifically, we have that

$$H(Y|\hat{Y}) = H(Y, \hat{Y}) - H(\hat{Y}) \geq H_{MLE}(Y, \hat{Y}) - \log |\hat{\mathcal{Y}}|,$$

where $|\hat{\mathcal{Y}}|$ is the cardinality of $\hat{Y}$.



(a) CIFAR 10 (b) CIFAR 100

Figure 2: Expected $[\log |C(X)|]^+$ as a function of α.

Based on this, we evaluate our set size bounds on ResNet models trained with CE on CIFAR 10 and CIFAR 100. As the logits $\hat{Y}$ used for CP are not discrete, we perform K-means clustering on them and construct a vector quantized $\hat{Y}_{vq}$ by assigning to each logit vector its closest cluster centroid. We then use $\hat{Y}_{vq}$ to perform CP with thresholding and also use $\hat{Y}_{vq}$ to obtain a lower bound on $H(Y|\hat{Y}_{vq})$ via a maximum likelihood estimate for $H(Y, \hat{Y}_{vq})$. For this maximum likelihood estimate we also perform the Miller-Madow bias correction [46].

For CIFAR 10 we cluster the logits into 32 clusters whereas for CIFAR 100 we use 256 clusters. The centroids are learned on a calibration set of 5k logits. For the model-based Fano bound, as it needs to compute terms that depend on the model probabilities and on whether the label was correctly covered by CP or not, we further split the calibration set into two equal-sized chunks; the first is used to find the quantile for thresholding and the second one is used to evaluate the terms of the bound. The obtained lower bounds on $[\log |C(X)|]^+$ for various α's can be seen at Figure 2. We also include an "empirical estimate" which is obtained by computing the quantile with the quantized calibration logits and then measuring the average $[\log |C(X)|]^+$ on the test set via thresholding on the quantized test logits. We see that model-based Fano provides relatively tight estimates for small values of alpha.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All of our claims are supported by our experimental results, and we provide detailed proofs for all theoretical results in the supplemental material.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We comment on the limitations of our work in Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We include detailed proofs for all theoretical results, clearly stating the assumptions (mainly exchangeable data and a valid conformal prediction method). We also provide the necessary background on information theory and conformal prediction in Appendix B to help the reader to follow our derivations.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We did our best to provide as many experimental details as possible, which can be found in the main text as well as in Appendix G.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code will be hosted at github.com/Qualcomm-AI-research/info_cp.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We give a detailed description of all of our experiments in Appendix G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report one standard deviation confidence intervals, computed across 10 random calibration/test splits, for all of our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiments are not particularly demanding in the terms of compute and can be reproduced on any modern hardware. We do state in Appendix G that we ran our experiments in commercially available NVIDIA GPUs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our work is in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have a discussion about the broader impact of our work in Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited the appropriate references for each dataset and, when a license is available, we added that information to the citation. We also made it clear that we downloaded all datasets through the torchvision [40] python package.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.