Label Noise Robustness of Conformal Prediction

Bat-Sheva Einbinder*   BAT-SHEVAB@CAMPUS.TECHNION.AC.IL
Department of Electrical and Computer Engineering
Technion - Israel Institute of Technology

Shai Feldman*   SHAI.FELDMAN@CS.TECHNION.AC.IL
Department of Computer Science
Technion - Israel Institute of Technology

Stephen Bates   STEPHENBATES@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Anastasios N. Angelopoulos   ANGELOPOULOS@BERKELEY.EDU
Department of Electrical Engineering and Computer Science
University of California, Berkeley

Asaf Gendler   ASAFGENDLER@CAMPUS.TECHNION.AC.IL
Department of Electrical and Computer Engineering
Technion—Israel Institute of Technology

Yaniv Romano   YROMANO@TECHNION.AC.IL
Departments of Electrical and Computer Engineering and of Computer Science
Technion—Israel Institute of Technology

Editor: Zaid Harchaoui

Abstract

We study the robustness of conformal prediction, a powerful tool for uncertainty quantification, to label noise. Our analysis tackles both regression and classification problems, characterizing when and how it is possible to construct uncertainty sets that correctly cover the unobserved noiseless ground truth labels. We further extend our theory and formulate the requirements for correctly controlling a general loss function, such as the false negative proportion, with noisy labels. Our theory and experiments suggest that conformal prediction and risk-controlling techniques with noisy labels attain conservative risk over the clean ground truth labels whenever the noise is dispersive and increases variability. In other adversarial cases, we can also correct for noise of bounded size in the conformal prediction algorithm in order to ensure achieving the correct risk of the ground truth labels without score or data regularity.

Keywords: conformal prediction, risk control, uncertainty quantification, label noise, distribution shift

---

*. Equal contribution. These authors are listed in alphabetical order.

1. Introduction

In most supervised classification and regression tasks, one would assume the provided labels reflect the ground truth. In reality, this assumption is often violated (Cheng et al., 2022; Xu et al., 2019; Yuan et al., 2018; Lee and Barber, 2022; Cauchois et al., 2022). For example, doctors labeling the same medical image may have different subjective opinions about the diagnosis, leading to variability in the ground truth label itself. To quote Abdalla and Fine (2023): "Noise in labeling schemas and gold label annotations are pervasive in medical imaging classification and affect downstream clinical deployment." In other settings, such variability may arise due to sensor noise, data entry mistakes, the subjectivity of a human annotator, or many other sources. In other words, the labels we use to train machine learning (ML) models may often be noisy in the sense that these are not necessarily the ground truth. Consequently, this can result in the formulation of invalid data-driven conclusions.

The above discussion emphasizes the critical need for making reliable predictions in real-world scenarios, especially when dealing with imperfect training data. An effective way to enhance the reliability of ML models is to quantify their prediction uncertainty. Conformal prediction (Vovk et al., 2005, 1999; Angelopoulos and Bates, 2023) is a generic uncertainty quantification tool that transforms the output of any ML model into prediction sets that are guaranteed to cover the future, unknown test labels with high probability. This guarantee holds for any data distribution and sample size, under the sole assumption that the training and test data are i.i.d. However, obtaining valid uncertainty quantification via conformal prediction in the presence of noisy labels remains unclear since the noise breaks the i.i.d. assumption. In this paper, we aim to address this specific challenge and precisely characterize under what conditions conformal methods, applied to noisy data, would yield prediction sets guaranteed to cover the unseen clean, ground truth label. Additionally, we analyze the effect of label noise on risk-controlling techniques, which extend the conformal prediction approach to construct uncertainty sets with a guaranteed control over a general risk function (Bates et al., 2021; Angelopoulos et al., 2021, 2024). We also show that our theory can be applied to online settings, e.g., to quantify prediction uncertainty for time series data with noisy labels. Overall, we analyze the behavior of conformal prediction and risk-controlling methods for several common loss functions and noise models, highlighting their built-in robustness to dispersive, variability-increasing noise and vulnerability to adversarial noise. Adversarial noise might reduce the uncertainty of the response variable, potentially causing the prediction sets to be too small and achieve a low coverage rate. We note that a summary of this paper and its key contributions can be found in (Feldman et al., 2023a).

2. Conformal Prediction Under Label Noise

2.1 Problem Setup

Consider a *calibration data set* of i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^n$ sampled from an arbitrary unknown distribution P_{XY}. Here, $X_i \in \mathbb{R}^p$ is the feature vector that contains p features for the i-th sample, and Y_i denotes its response, which can be discrete for classification tasks

or continuous for regression tasks. Given the calibration data set, an i.i.d. test data point $(X_\text{test}, Y_\text{test})$, and a pre-trained model $\hat{f}$, conformal prediction constructs a set $\widehat{\mathcal{C}}_\text{clean}(X_\text{test})$ that contains the unknown test response, Y_test, with high probability, e.g., 90%. We refer to $\widehat{\mathcal{C}}_\text{clean}(X_\text{test})$ as 'clean' to underscore that the prediction set is formed by utilizing samples from the clean data distribution. That is, for a user-specified level $\alpha \in (0, 1)$,

$$\mathbb{P}\left(Y_\text{test} \in \widehat{\mathcal{C}}_\text{clean}(X_\text{test})\right) \geq 1 - \alpha. \tag{1}$$

This property is called *marginal coverage*, where the probability is defined over the calibration and test data.

In the setting of label noise, we only observe the corrupted labels $\tilde{Y}_i = g(Y_i)$ for some *corruption function* $g : \mathcal{Y} \times [0, 1] \to \mathcal{Y}$, so the i.i.d. assumption and marginal coverage guarantee are invalidated. The corruption is random; we will always take the second argument of g to be a random seed U uniformly distributed on $[0, 1]$. To ease notation, we leave the second argument implicit henceforth. Nonetheless, using the noisy calibration data, we seek to form a prediction set $\widehat{\mathcal{C}}_\text{noisy}(X_\text{test})$ that covers the clean, uncorrupted test label, Y_test. More precisely, our goal is to delineate when it is possible to provide guarantees of the form

$$\mathbb{P}\left(Y_\text{test} \in \widehat{\mathcal{C}}_\text{noisy}(X_\text{test})\right) \geq 1 - \alpha, \tag{2}$$

where the probability is taken jointly over the calibration data, test data, and corruption function (this will be the case for the remainder of the paper).

More formally, we use the model $\hat{f}$ to construct a score function, $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which is engineered to be large when the model is uncertain and small otherwise. We will introduce different score functions for both classification and regression as needed in the following subsections. Abbreviate the scores on each calibration data point as $s_i = s(X_i, Y_i)$ for each $i = 1, ..., n$. Conformal prediction tells us that we can achieve a marginal coverage guarantee by picking $\hat{q}_\text{clean} = s_{(\lceil (n+1)(1-\alpha) \rceil)}$ as the $\lceil (n+1)(1-\alpha) \rceil$-smallest of the calibration scores and constructing the prediction sets as

$$\widehat{\mathcal{C}}_\text{clean}(X_\text{test}) = \{y \in \mathcal{Y} : s(X_\text{test}, y) \leq \hat{q}_\text{clean}\}.$$

In this paper, we do not allow ourselves access to the calibration labels, only their noisy versions, $\tilde{Y}_1, \ldots, \tilde{Y}_n$, so we cannot calculate $\hat{q}_\text{clean}$. Instead, we can calculate the noisy quantile $\hat{q}_\text{noisy}$ as the $\lceil (n+1)(1-\alpha) \rceil$-smallest of the *noisy* score functions, $\tilde{s}_i = s(X_i, \tilde{Y}_i)$. The main formal question of our work is whether the resulting prediction set, $\widehat{\mathcal{C}}_\text{noisy}(X_\text{test}) = \{y : s(X_\text{test}, y) \leq \hat{q}_\text{noisy}\}$, covers the clean label as in (2). We state this general recipe algorithmically for future reference:

Recipe 1 (Conformal prediction with noisy labels)

1. *Consider i.i.d. data points* $(X_1, Y_1), \ldots, (X_n, Y_n), (X_\text{test}, Y_\text{test})$, *a corruption model* $g : \mathcal{Y} \to \mathcal{Y}$, *and a score function* $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

2. *Compute the conformal quantile with the corrupted labels,*

$$\hat{q}_\text{noisy} = \text{Quantile}\left(\frac{(n+1)(1-\alpha)}{n}, \{s(X_i, \tilde{Y}_i)\}_{i=1}^n\right),$$

where $\tilde{Y}_i = g(Y_i)$.

3. *Construct the prediction set using the noisy conformal quantile,*

$$\widehat{\mathcal{C}}_{\text{noisy}} = \{y : s(X_{\text{test}}, y) \leq \hat{q}_{\text{noisy}}\}.$$

This recipe produces prediction sets that cover the noisy label at the desired coverage rate (Vovk et al., 2005; Angelopoulos and Bates, 2023):

$$\mathbb{P}\left(\tilde{Y}_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \geq 1 - \alpha.$$

Note that achieving (2) is impossible in general—see Proposition 3. However, we present some realistic noise models under which this goal is satisfied. In the following sections, we provide real-data experiments indicating that conformal prediction and risk controlling methods achieve valid risk/coverage even with access only to noisy labels. In Section 2.3, we outline the precise conditions for these guarantees. There is a particular type of anti-dispersive noise that causes failure, and we see this both in theory and practice. If such noise can be avoided, a user should feel safe deploying uncertainty quantification techniques even with noisy labels. Finally, we note that there exist several works that study the robustness of conformal prediction under ambiguous labels, and we discuss them in more detail in Section 5.1.

2.2 Empirical Evidence

Classification

As a real-world example of label noise, we conduct an image classification experiment where we only observe one annotator's label but seek to cover the majority vote of many annotators. For this purpose, we use the CIFAR-10H data set, used by Peterson et al. (2019); Battleday et al. (2020); Singh et al. (2020), which contains 10,000 images labeled by approximately 50 annotators. We calibrate using only a single annotator's label and seek to cover the majority vote of the 50 annotators. The single annotator differs from the ground truth labels in approximately 5% of the images.

Using the noisy calibration set (i.e., a calibration set containing these noisy labels), we apply vanilla conformal prediction as if the data were i.i.d, and study the performance of the resulting prediction sets. Details regarding the training procedure can be found in Appendix B.1. The fraction of majority vote labels covered is demonstrated in Figure 1. This figure shows that when using the clean calibration set, the marginal coverage is 90%, as expected. When using the noisy calibration set, the coverage on the clean, unknown test point, i.e., (2) increases to approximately 93%. Figure 1 also demonstrates that prediction sets that were calibrated using noisy labels are larger than sets calibrated with clean labels. This experiment illustrates the main intuition behind our paper: adding noise will usually increase the variability in the labels, leading to larger prediction sets that retain the coverage property.
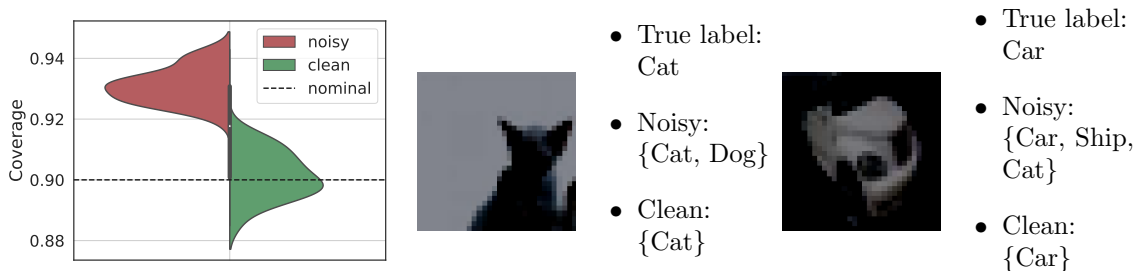
Figure 1: **Effect of label noise on CIFAR-10.** Left: distribution of average coverage on a clean test set over 30 independent experiments with target coverage $1 - \alpha = 90\%$, using noisy and clean labels for calibration. We use a pre-trained resnet 18 model, which has Top-1 accuracy of 93% and 90% on the clean and noisy test set, respectively. The gray bar represents the interquartile range. Center and right: prediction sets achieved using noisy and clean labels for calibration.

REGRESSION

In this section, we present a real-world application with a continuous response, using Aesthetic Visual Analysis (AVA) data set, first presented by Murray et al. (2012). This data set contains pairs of images and their aesthetic scores in the range of 1 to 10, obtained by approximately 200 annotators. Following Kao et al. (2015); Talebi and Milanfar (2018); Murray et al. (2012), the task is to predict the average aesthetic score of a given test image. Therefore, we consider the average aesthetic score taken over all annotators as the clean, ground truth response. The noisy response is the average aesthetic score taken over 10 randomly selected annotators only.

We examine the performance of conformal prediction using two different scores: the CQR score (Romano et al., 2019), defined in (4) and the residual magnitude score (Papadopoulos et al., 2002; Lei et al., 2018) defined in (5). In our experiments, we set the function $\hat{u}(x)$ of the residual magnitude score as $\hat{u}(x) = 1$. We follow Talebi and Milanfar (2018) and take a transfer learning approach to fit the predictive model using a VGG-16 model pretrained on ImageNet data set. Details regarding the training strategy are in Appendix B.2.

Figure 2 portrays the marginal coverage and average interval length achieved using CQR and residual magnitude scores. As a point of reference, this figure also presents the performance of the two conformal methods when calibrated with a clean calibration set; as expected, the two perfectly attain 90% coverage. By constant, when calibrating the same predictive models with a noisy calibration set, the resulting prediction intervals tend to be wider and to over-cover the average aesthetic scores.

Thus far, we have found in empirical experiments that conservative coverage is obtained in the presence of label noise. In the following sections, our objective is to establish conditions that formally guarantee label-noise robustness.
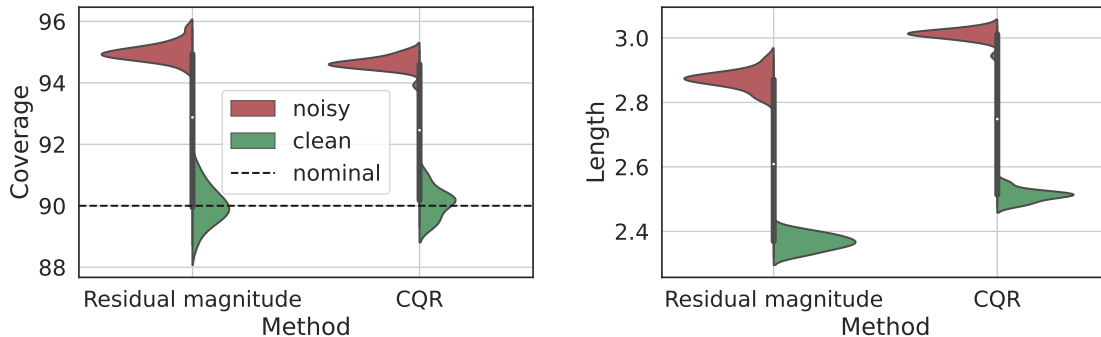
Figure 2: **Results for real-data regression experiment: predicting aesthetic visual rating.** Performance of conformal prediction intervals with 90% marginal coverage based on a VGG-16 model using a noisy training set. We compare the residual magnitude score and CQR methods with both noisy and clean calibration sets. Left: Marginal coverage; Right: Interval length. The results are evaluated over 30 independent experiments and the gray bar represents the interquartile range.

2.3 Theoretical Analysis

2.3.1 GENERAL ANALYSIS

We begin the theoretical analysis with a general statement, showing that Recipe 1 produces valid prediction sets whenever the noisy score distribution stochastically dominates the clean score distribution. The intuition is that the noise distribution 'spreads out' the distribution of the score function such that $\hat{q}_{\text{noisy}}$ is (stochastically) larger than $\hat{q}_{\text{clean}}$.

Theorem 1 *Assume that* $\mathbb{P}(\tilde{s}_{\text{test}} \leq t) \leq \mathbb{P}(s_{\text{test}} \leq t)$ *for all* t. *Then,*

$$\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \geq 1 - \alpha.$$

Furthermore, for any u *satisfying* $\mathbb{P}(\tilde{s}_{\text{test}} \leq t) + u \geq \mathbb{P}(s_{\text{test}} \leq t)$, *for all* t, *then*

$$\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \leq 1 - \alpha + \frac{1}{n+1} + u.$$

Figure 3 illustrates the idea behind Theorem 1, demonstrating that when the noisy score distribution stochastically dominates the clean score distribution, then $\hat{q}_{\text{noisy}} \geq \hat{q}_{\text{clean}}$ and thus uncertainty sets calibrated using noisy labels are more conservative. In practice, however, one does not have access to such a figure, since the scores of the clean labels are unknown. This gap requires an individual analysis for every task and its noise setup, which emphasizes the complexity of this study. In general, for most commonly used score functions the stochastic dominance assumption holds when the noise is dispersive, meaning it "flattens" the density of $Y \mid X$, e.g., when $\text{Var}(\tilde{Y} \mid X) > \text{Var}(Y \mid X)$. In the following subsections, we present example setups in classification and regression tasks under which this stochastic dominance holds, and conformal prediction with noisy labels succeeds in
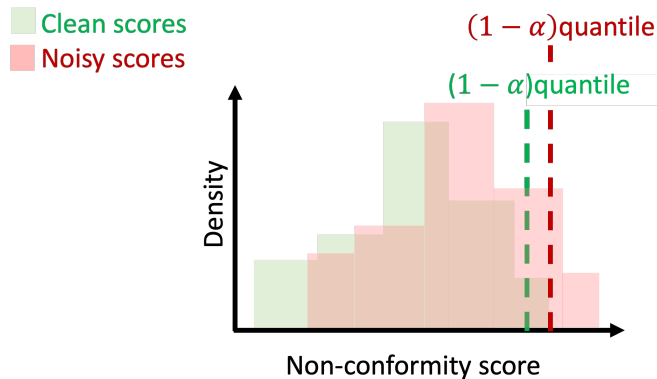
6

Figure 3: Clean (green) and noisy (red) non-conformity scores under dispersive corruption.

covering the true, noiseless label. The purpose of these examples is to illustrate simple and intuitive statistical settings where Theorem 1 holds. Under the hood, all setups given in the following subsections are applications of Theorem 1. Though the noise can be adversarially designed to violate these assumptions and cause under-coverage (as in the impossibility result in Proposition 3), the evidence presented here suggests that in the majority of practical settings, conformal prediction can be applied without modification. The proof is given in Appendix A.1.

2.3.2 REGRESSION

In this section, we analyze a regression task where the labels are continuous-valued and the corruption function is additive:

$$g^{\text{add}}(y) = y + Z \tag{3}$$

for some independent noise sample Z.

We first analyze the setting where the noise Z is symmetric around 0 and the density of $Y \mid X$ is symmetric unimodal. We also assume that the estimated prediction interval contains the true median of $\tilde{Y} \mid X = x$, which is a very weak assumption about the fitted model. The following proposition states that such an interval achieves a conservative coverage rate over the clean labels.

Proposition 1 *Assume a symmetric unimodal distribution of $Y \mid X$ and an independent additive noise that is symmetric around zero. Let $\widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})$ be a prediction interval constructed according to Recipe 1 with the corruption function g^{add}. Further, suppose that $\widehat{\mathcal{C}}_{\text{noisy}}(x)$ contains the true median of $\tilde{Y} \mid X = x$ for all $x \in \mathcal{X}$. Then*

$$\mathbb{P}\Big(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\Big) \geq 1 - \alpha.$$

Importantly, the corruption function may depend on the feature vector, as stated next.

Remark 2 (Corruptions dependent on X) *Proposition 1 holds even if the noise Z depends on X.*

7

Furthermore, Proposition 1 applies for any non-conformity score function, specifically for two popular scores we focus on. The CQR score, developed by Romano et al. (2019), measures the distance of an estimated interval's endpoints from the corresponding label y, formally defined as:

$$s^{\text{CQR}}(x, y) = \max\{\hat{f}_{\text{lower}}(x) - y, y - \hat{f}_{\text{upper}}(x)\}. \tag{4}$$

Above, $\hat{f}_{\text{lower}}$ and $\hat{f}_{\text{upper}}$ are the estimated lower and upper interval's endpoints, e.g., obtained by fitting a quantile regression model to approximate the $\alpha/2$ and $1-\alpha/2$ conditional quantiles of $\tilde{Y} \mid X$ (Koenker and Bassett, 1978). The residual magnitude (RM) score (Papadopoulos et al., 2002; Lei et al., 2018) assesses the normalized prediction error:

$$s^{\text{RM}}(x, y) = \left|\hat{f}(x) - y\right|/\hat{u}(x), \tag{5}$$

where $\hat{f}$ is a regression model, such as a conditional mean estimator of $\tilde{Y} \mid X$, and $\hat{u}$ is some normalization function, e.g., $\hat{u}(x) \equiv 1$. Additionally, the distributional assumption on $Y \mid X$ is not necessary to guarantee label-noise robustness under this noise model; in Appendix A.2.1 we show that this distributional assumption can be relaxed at the expense of additional requirements on the estimated interval. Furthermore, in Section 3.3.3 we extend this analysis and formulate bounds for the coverage rate obtained on the clean labels that do not require any distributional assumptions and impose no restrictions on the constructed intervals. Subsequently, in Section 4.5, we empirically evaluate the proposed bounds and show that they are informative despite their extremely weak assumptions.

2.3.3 CLASSIFICATION

In this section, we formulate the conditions under which conformal prediction is robust to label noise in a K-class classification setting where the labels take one of K values, i.e., $Y \in \{1, 2, ..., K\}$. In a nutshell, robustness is guaranteed when the corruption function transforms the labels' distribution towards a uniform distribution while maintaining the same ranking of labels. Such corruption increases the prediction uncertainty, which drives conformal methods to construct conservative uncertainty sets to achieve the nominal coverage level on the observed noisy labels. We now formalize this intuition. We begin by defining the following noise models:

- **Uniform noise**: A noise model that fulfils the following: for all $x \in \mathcal{X}$:

  1. $\forall i \in \{1, .., K\} : \left|\mathbb{P}(\tilde{Y}_{\text{test}} = i \mid X_{\text{test}} = x) - \frac{1}{K}\right| \leq \left|\mathbb{P}(Y_{\text{test}} = i \mid X_{\text{test}} = x) - \frac{1}{K}\right|$,
  2. $\forall i, j \in \{1, .., K\} : \mathbb{P}(Y_{\text{test}} = i \mid X_{\text{test}} = x) \leq \mathbb{P}(Y_{\text{test}} = j \mid X_{\text{test}} = x) \iff \mathbb{P}(\tilde{Y}_{\text{test}} = i \mid X_{\text{test}} = x) \leq \mathbb{P}(\tilde{Y}_{\text{test}} = j \mid X_{\text{test}} = x)$.

- **Random flip**: A corruption function that follows:

$$g^{\text{flip}}(y) = \begin{cases} y & \text{w.p } 1 - \epsilon \\ Y' & \text{else,} \end{cases} \tag{6}$$

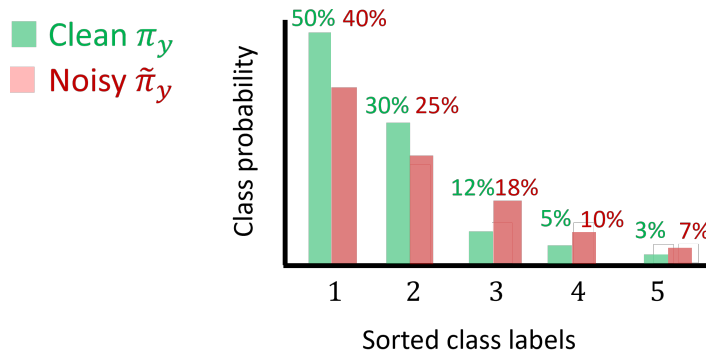 where Y' is uniformly drawn from the set $\{1, ..., K\}$.

Figure 4: Clean (green) and noisy (red) class probabilities under dispersive corruption.

Proposition 2 *Let* $\widehat{\mathcal{C}}_{\text{noisy}}$ *be constructed as in Recipe 1. Then, the coverage rate achieved over the clean labels is upper bounded by:*

$$\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \leq 1 - \alpha + \frac{1}{n+1} + \frac{1}{2}\sum_{i=1}^{K}\left|\mathbb{P}(\tilde{Y}_{\text{test}} = i) - \mathbb{P}(Y_{\text{test}} = i)\right|.$$

Further suppose that $\widehat{\mathcal{C}}_{\text{noisy}}$ *contains the most likely labels, i.e.,* $\forall x \in \mathcal{X}, i \in \widehat{\mathcal{C}}_{\text{noisy}}(x), j \notin \widehat{\mathcal{C}}_{\text{noisy}}(x) : \mathbb{P}(\tilde{Y}_{\text{test}} = i \mid X_{\text{test}} = x) \geq \mathbb{P}(\tilde{Y}_{\text{test}} = j \mid X_{\text{test}} = x)$. *If the noise follows the* `uniform noise` *model, then the coverage rate is guaranteed to be valid:*

$$1 - \alpha \leq \mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right).$$

We emphasize that the key contribution of Proposition 2 is the lower bound, as it guarantees a valid coverage rate. We also note that Barber et al. (2023) provides a sharper upper bound for the coverage rate which also relies on the TV-distance between $Y_{\text{test}}, \tilde{Y}_{\text{test}}$. Nevertheless, the lower bound in Proposition 2 is tighter than the lower bound described in Barber et al. (2023). Figure 4 visualizes the essence of Proposition 2, showing that as the corruption increases the label uncertainty, the prediction sets generated by conformal prediction get larger, having a conservative coverage rate. It should be noted that under the above proposition, achieving valid coverage requires only knowledge of the *noisy* conditional distribution $\mathbb{P}(\tilde{Y} \mid X)$, so a model trained on a large amount of noisy data should approximately have the desired coverage rate. Moreover, one should notice that without the assumptions in Proposition 2, even the oracle model is not guaranteed to achieve valid coverage, as we will show in Section 4.1. For a more general analysis of classification problems where the noise model is an arbitrary confusion matrix, see Appendix A.3.2.

We now turn to examine two conformity scores and show that applying them with noisy data leads to conservative coverage, as a result of Proposition 2. The *adaptive prediction sets* (APS) score, first introduced by Romano et al. (2020), is defined as

$$s^{\text{APS}}(x, y) = \sum_{y' \in \mathcal{Y}} \hat{\pi}_{y'}(x) \mathbb{I}\left\{\hat{\pi}_{y'}(x) > \hat{\pi}_{y}(x)\right\} + \hat{\pi}_{y}(x) \cdot U',$$

where $\mathbb{I}$ is the indicator function, $\hat{\pi}_y(x)$ is the estimated conditional probability $\mathbb{P}(Y = y \mid X = x)$ and $U' \sim \mathrm{Unif}(0,1)$. To make this non-random, a variant of the above where $U' = 1$ is often used. The APS score is one of two popular conformal methods for classification. The other score from Vovk et al. (2005); Lei et al. (2013), is referred to as *homogeneous prediction sets* (HPS) score, $s^{\mathrm{HPS}}(x,y) = 1 - \hat{\pi}_y(x)$, for some classifier $\hat{\pi}_y(x) \in [0,1]$. The next corollary states that with access to an oracle classifier's ranking, conformal prediction covers the noiseless test label.

Corollary 1 *Let $\widehat{\mathcal{C}}_{\mathrm{noisy}}(X_{\mathrm{test}})$ be constructed as in Recipe 1 with either the APS or HPS score functions, with any classifier that ranks the classes in the same order as the oracle classifier $\hat{\pi}_y(x) = \mathbb{P}(\tilde{Y} = y \mid X = x)$. Then,*

$$1 - \alpha \leq \mathbb{P}\Big(Y_{\mathrm{test}} \in \widehat{\mathcal{C}}_{\mathrm{noisy}}(X_{\mathrm{test}})\Big) \leq 1 - \alpha + \frac{1}{n+1} + \frac{1}{2}\sum_{i=1}^{K}\Big|\mathbb{P}(\tilde{Y}_{\mathrm{test}} = i) - \mathbb{P}(Y_{\mathrm{test}} = i)\Big|.$$

We now turn to examine the specific `random flip` noise model in which the noisy label is randomly flipped ϵ fraction of the time. This noise model is well-studied in the literature; see, for example, (Aslam and Decatur, 1996; Angluin and Laird, 1988; Ma et al., 2018; Jenni and Favaro, 2018; Jindal et al., 2016; Yuan et al., 2018).

Corollary 2 *Let $\widehat{\mathcal{C}}_{\mathrm{noisy}}(X_{\mathrm{test}})$ be constructed as in Recipe 1 with the corruption function g^{flip} and either the APS or HPS score functions, with any classifier that ranks the classes in the same order as the oracle classifier $\hat{\pi}_y(x) = \mathbb{P}(\tilde{Y} = y \mid X = x)$. Then,*

$$1 - \alpha \leq \mathbb{P}\Big(Y_{\mathrm{test}} \in \widehat{\mathcal{C}}_{\mathrm{noisy}}(X_{\mathrm{test}})\Big) \leq 1 - \alpha + \frac{1}{n+1} + \epsilon\frac{K-1}{K}.$$

Crucially, the above corollaries apply with any score function that preserves the order of the estimated classifier, which emphasizes the generality of our theory. All proofs are given in Appendix A.3.1. Moreover, although this is not our main focus, in Appendix A.3.3 we investigate the inflation of the prediction set size in the specific case of `random flip` noise with APS scores and the oracle model. Table 1 summarizes all different settings we examine with their corresponding bounds. Finally, we note that in Section 3.3.1 we extend the above analysis to multi-label classification, where there may be multiple labels that correspond to the same sample.

2.4 Disclaimer: Distribution-Free Results

Though the coverage guarantee holds in many realistic cases, conformal prediction may generate uncertainty sets that fail to cover the true outcome. Indeed, in the general case, conformal prediction produces invalid prediction sets, and must be adjusted to account for the size of the noise. The following proposition states that for any nontrivial noise distribution, there exists a score function that breaks naïve conformal.

Table 1: Summary of coverage bounds for different scores and noise models

Task	Score	Noise model	Bounds
Regression	All	Additive symmetric: g^{add} in (3)	$\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \geq 1 - \alpha$ (Proposition 1)
	All	Additive symmetric: g^{add} in (3)	$\mathbb{P}(Y \in C(x) \mid X = x) \geq$ $\mathbb{P}(\tilde{Y} \in C(x) \mid X = x)$ (Theorem A4)
Classification	All	uniform noise	$1 - \alpha \leq \mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \leq$ $1 - \alpha + \frac{1}{n+1}$ $+\frac{1}{2}\sum_{i=1}^{K}\left|\mathbb{P}(\tilde{Y}_{\text{test}} = i) - \mathbb{P}(Y_{\text{test}} = i)\right|$ (Proposition 2)
	HPS, APS	Random flip: g^{flip} in (6)	$1 - \alpha \leq \mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \leq$ $1 - \alpha + \frac{1}{n+1} + \epsilon\frac{K-1}{K}$ (Corollary 2)

Proposition 3 (Coverage is impossible in the general case.) *Take any $\tilde{Y} \overset{d}{\neq} Y$. Then there exists a score function s that yields $\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) < \mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\right)$, for $\widehat{\mathcal{C}}_{\text{noisy}}$ constructed using noisy samples and $\widehat{\mathcal{C}}$ constructed with clean samples.*

The above proposition says that for any noise distribution, there exists an adversarially chosen score function that will disrupt coverage. Furthermore, as we discuss in Appendix A.4, with a noise of a sufficient magnitude, it is possible to get arbitrarily bad violations of coverage. In Appendix A.4 we state an additional impossibility result in which we claim that for any given score function following some conditions, there is an adversarial noise that invalidates the coverage.

Next, we discuss how to adjust the threshold of conformal prediction to account for noise of a known size, as measured by total variation (TV) distance from the clean label.

Corollary 3 (Corollary of Barber et al. (2023)) *Let $\tilde{Y}$ be any random variable satisfying $D_{\text{TV}}(Y, \tilde{Y}) \leq \epsilon$. Take $\alpha' = \alpha + 2\frac{n}{n+1}\epsilon$. Letting $\widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})$ be the output of Recipe 1 with any score function at level α' yields*

$$\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \geq 1 - \alpha.$$

We discuss this strategy more in Appendix A.4—the algorithm implied by Corollary 3 may not be particularly useful, as the TV distance is a badly behaved quantity that is also difficult to estimate, especially since the clean labels are inaccessible.

As a final note, if the noise is bounded in TV norm, then the coverage is also not too conservative.

Corollary 4 (Corollary of Barber et al. (2023) Theorem 3) *Let $\tilde{Y}$ be any random variable satisfying $D_{\mathrm{TV}}(Y, \tilde{Y}) \leq \xi$. Letting $\widehat{\mathcal{C}}_{\mathrm{noisy}}(X_{\mathrm{test}})$ be the output of Recipe 1 with any score function at level α yields*

$$\mathbb{P}\big(Y_{\mathrm{test}} \in \widehat{\mathcal{C}}_{\mathrm{noisy}}(X_{\mathrm{test}})\big) < 1 - \alpha + \frac{1}{n+1} + \frac{n}{n+1}\xi.$$

3. Risk Control Under Label Noise

3.1 Problem Setup

Up to this point, we have focused on the miscoverage loss, and the analysis conducted thus far applies exclusively to this metric. However, in real-world applications, it is often desired to control metrics other than the binary loss $L^{\mathrm{miscoverage}}(y, C) = \mathbb{1}\{y \notin C\}$, where C is a set of predicted labels. Examples of such alternative losses include the F1-score or the false negative rate. The latter is particularly relevant for high-dimensional response Y, as in tasks like multi-label classification or image segmentation. To address this need, researchers have developed extensions of the conformal framework that go beyond the miscoverage loss, providing a rigorous *risk control* guarantee for general loss functions (Bates et al., 2021; Angelopoulos et al., 2021, 2024).

Similarly to the conformal prediction algorithm, in the risk-control setting we post-process the predictions of a model $\hat{f}$ to create a prediction set $\widehat{\mathcal{C}}_\lambda(X_{\mathrm{test}})$ with a parameter λ that determines its level of conservativeness: higher values yield larger and nested sets, in the sense that $\widehat{\mathcal{C}}_{\lambda_2}(\cdot) \subseteq \widehat{\mathcal{C}}_{\lambda_1}(\cdot)$ for $\lambda_2 \leq \lambda_1$. For instance, if $\hat{f}_y(x)$ is an estimator of the conditional probability of $Y \mid X = x$, then the prediction sets can be defined as $\widehat{\mathcal{C}}_\lambda(X_{\mathrm{test}}) = \{y : \hat{f}_y(X_{\mathrm{test}}) > \lambda\}$. To measure the quality of $\widehat{\mathcal{C}}_\lambda(X_{\mathrm{test}})$, we consider a loss function $L(Y_{\mathrm{test}}, \widehat{\mathcal{C}}_\lambda(X_{\mathrm{test}}))$ which we require to be non-increasing as a function of λ. This is analogous to conformal prediction in which the quantile of the scores, $\hat{q}$, encodes the prediction set sizes, and the error measure is simply the miscoverage loss. In the conformal prediction framework, we use a holdout set $\{(X_i, Y_i)\}_{i=1}^n$ in order to calibrate $\hat{q}_{\mathrm{clean}}$ and achieve valid coverage over a new test point. Likewise, in the conformal risk control settings we aim to use the observed losses $\{L(Y_i, \widehat{\mathcal{C}}_\lambda(X_i))\}_{i=1}^n$ derived by the calibration set to find a calibrated threshold $\hat{\lambda}_{\mathrm{clean}}$ that will control the risk of an unseen test point at a pre-specified level α:

$$\mathcal{R}(C) = \mathbb{E}[L(Y_{\mathrm{test}}, \widehat{\mathcal{C}}_{\hat{\lambda}_{\mathrm{clean}}}(X_{\mathrm{test}}))] \leq \alpha.$$

See Appendix C.1 for the conformal risk control procedure. Analogously to conformal prediction, these methods produce valid sets under the i.i.d. assumption, but their guarantees do not hold in the presence of label noise. Provided a noisy calibration set, $\{(X_i, \tilde{Y}_i)\}_{i=1}^n$, the parameter $\hat{\lambda}_{\mathrm{noisy}}$ is constructed using the noisy losses $\{L(\tilde{Y}_i, \widehat{\mathcal{C}}_\lambda(X_i))\}_{i=1}^n$, and therefore the risk of a new clean test point is not guaranteed to be controlled.

Our main goal is to delineate when it is possible to provide a risk control guarantee of the form

$$\mathbb{E}[L(Y_{\mathrm{test}}, \widehat{\mathcal{C}}_{\hat{\lambda}_{\mathrm{noisy}}}(X_{\mathrm{test}}))] \leq \alpha.$$

In the next section, we conduct a real-data experiment, demonstrating that conformal risk control is robust to label noise. In Section 3.3 we explain this phenomenon and specify the conditions under which conservative risk is obtained. To ease notation, we will omit the subscript λ but directly specify whether the prediction set is constructed by clean or noisy calibration data.

3.2 Empirical Evidence

Multi-label Classification

In this section, we analyze the robustness of conformal risk control (Angelopoulos et al., 2024) to label noise in a multi-label classification setting. We use the MS COCO data set (Lin et al., 2014), in which the input image may contain up to $K = 80$ positive labels, i.e., $Y \subseteq \{1, 2, ..., K\}$. In the following experiment, we consider the annotations in this data set as ground-truth labels. We have collected noisy labels from individual annotators who annotated 117 images in total. On average, the annotators missed or mistakenly added approximately 1.75 labels from each image. See Appendix B.5 for additional details about this experimental setup and data collection.

We fit a TResNet (Ridnik et al., 2021) model on 100k clean samples and calibrate it using 105 noisy samples with conformal risk control, as introduced in (Angelopoulos et al., 2024, Section 3.2), to control the false-negative rate (FNR) at different levels. The FNR is the ratio of positive labels of Y that are missed in a prediction set C, formally defined as:

$$L^{\text{FNR}}(Y, C) = 1 - \frac{|Y \cap C|}{|Y|}. \tag{7}$$

We measure the FNR obtained over the clean and noisy test sets which contain 40k and 12 samples, respectively. Figure 5 displays the results, showing that the uncertainty sets attain valid risk even though they were calibrated using corrupted data. In the following section, we aim to explain these results and find the conditions under which label-noise robustness is guaranteed.
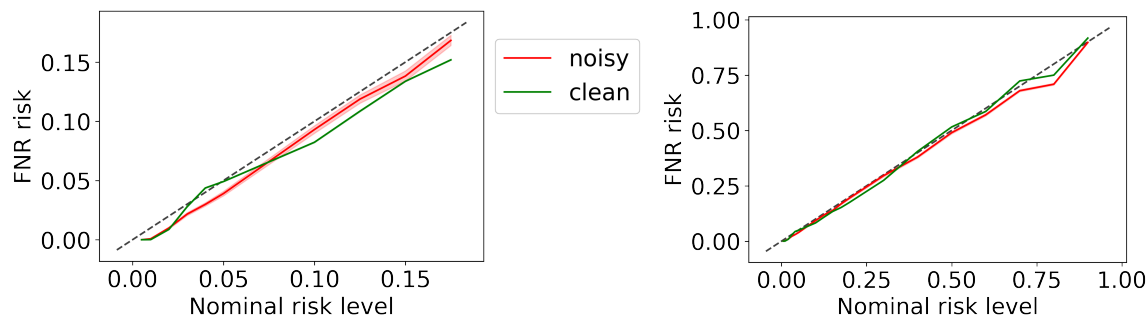

Figure 5: FNR on MS COCO data set, achieved over noisy (red) and clean (green) test sets. The calibration scheme is applied with noisy annotations. Results are averaged over 2000 trials.

3.3 Theoretical Analysis

3.3.1 Multi-label Classification

In this section, we study the conditions under which conformal risk control is robust to label noise in a multi-label classification setting. Recall that each sample contains up to K positive labels, i.e., $Y \subseteq \{1, 2, ..., K\}$. Here, we assume a vector-flip noise model with a binary random variable ϵ_i that flips the i-th label with probability $\mathbb{P}(\epsilon_i = 1)$:

$$g^{\text{vector}-\text{flip}}(y)_i = y_i(1 - \epsilon_i) + (1 - y_i)\epsilon_i. \tag{8}$$

Above, y_i is an indicator that takes the value 1 if the i-th label is present in y, and 0, otherwise. Notice that the random variable ϵ_i takes the value 1 if the i-th label in y is flipped and 0 otherwise. We further assume that the noise is not adversarial, i.e., $\mathbb{P}(\epsilon_i = 1) < 0.5$ for all $i \in \{1, 2, ..., K\}$. We now show that a valid FNR risk is guaranteed in the presence of label noise under the following assumptions.

Proposition 4 *Let $\widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})$ be a prediction set that contains the most likely labels, in the sense of Proposition 2, and controls the FNR risk of the noisy labels at level α. Assume the multi-label noise model $g^{\text{vector}-\text{flip}}$. If*

1. *Y_{test} is a deterministic function of X_{test},*

2. *The number of positive labels, i.e., $|\tilde{Y}_{\text{test}}| \mid X_{\text{test}} = x$ is a constant for all $x \in \mathcal{X}$,*

then

$$\mathbb{E}\left[L^{\text{FNR}}\left(Y_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right)\right] \leq \alpha.$$

Remark 3 (Dependent corruptions) *Proposition 4 holds even if the elements in the noise vector ϵ depend on each other or on X.*

The proof and other additional theoretical results are provided in Appendices A.5.1- A.5.4. Importantly, the determinism assumption on $Y \mid X = x$ is reasonable as it is simply satisfied when the noiseless response is defined as the consensus outcome. Nevertheless, this assumption may not always hold in practice. Thus, in the next section, we propose alternative requirements for the validity of the FNR risk and demonstrate them in a segmentation setting.

3.3.2 Segmentation

In segmentation tasks, the goal is to assign labels to every pixel in an input image such that pixels with similar characteristics share the same label. For example, tumor segmentation can be applied to identify polyps in medical images. Here, the response is a binary matrix $Y \subseteq \{0, 1\}^{W \times H}$ that contains the value 1 in the (i, j) pixel if it includes the object of interest and 0 otherwise. The uncertainty is represented by a prediction set $C \subseteq \{1, ..., W\} \times$

$\{1, ..., H\}$ that includes pixels that are likely to contain the object. Similarly to the multi-label classification problem, here, we assume a vector flip noise model $g^{\text{vector}-\text{flip}}$, that flips the (i, j) pixel in Y, denoted as $Y_{i,j}$, with probability $\mathbb{P}(\epsilon_{i,j} = 1)$. We now show that the prediction sets constructed using a noisy calibration set are guaranteed to have conservative FNR if the clean response matrix Y and the noise variable ϵ are independent given X.

Proposition 5 *Let $\widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})$ be a prediction set that contains the most likely pixels, in the sense of Proposition 2, and controls the FNR risk of the noisy labels at level α. Suppose that:*

1. *The elements of the clean response matrix are independent of each other given X_{test}. That is, $Y_{\text{test}}{}^{i,j} \mid X_{\text{test}} = x \perp\!\!\!\perp Y_{\text{test}}{}^{m,n} \mid X_{\text{test}} = x$ for all $(i, j) \neq (m, n) \subseteq \{1, ..., W\} \times \{1, ..., H\}$ and $x \in \mathcal{X}$.*

2. *For a given input X_{test}, the noise level is the same for all response elements.*

3. *The noise variable ϵ is independent of $Y_{\text{test}} \mid X_{\text{test}} = x$, and the noises of different labels are independent of each other given X_{test}, similarly to condition 1.*

Then,

$$\mathbb{E}\left[L^{\text{FNR}}\left(Y_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right)\right] \leq \alpha.$$

We note that a stronger version of this proposition that allows dependence between the elements in Y_{test} is given in Appendix A.5.1, and the proof is in Appendix A.5.5. The advantage of Proposition 5 over Proposition 4 is that here, the response matrix Y_{test} and the number of positive labels are allowed to be stochastic. For this reason, we believe that Proposition 5 is more suited for segmentation tasks, even though Proposition 4 applies in segmentation settings as well.

3.3.3 REGRESSION WITH A GENERAL LOSS

This section studies the general regression setting in which $Y \in \mathbb{R}$ takes continuous values and the loss function is an arbitrary function $L(y, \widehat{\mathcal{C}}_{\text{noisy}}(x)) \in \mathbb{R}$. Here, our objective is to find tight bounds for the risk over the clean labels using the risk observed over the corrupted labels. The main result of this section accomplishes this goal while making minimal assumptions on the loss function, the noise model, and the data distribution.

Proposition 6 *Let $\widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})$ be a prediction interval that controls the risk at level $\alpha :=$ $\mathbb{E}\left[L(\tilde{Y}_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}}))\right]$. Suppose that the second derivative of the loss $L(y; C)$ is bounded for all y, C: $q \leq \frac{\partial^2}{\partial y^2}L(y; C) \leq Q$ for some $q, Q \in \mathbb{R}$. If the labels are corrupted by the function g^{add} from (3) with a noise Z that satisfies $\mathbb{E}[Z] = 0$, then*

$$\alpha - \frac{1}{2}Q \cdot \text{Var}(Z) \leq \mathbb{E}\left[L(Y_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right] \leq \alpha - \frac{1}{2}q \cdot \text{Var}(Z).$$

If we further assume that L is convex then we obtain valid risk:

$$\mathbb{E}\left[L(Y_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right] \leq \alpha.$$

The proof is detailed in Appendix A.6.1. Remarkably, Proposition 6 applies for any predictive model, calibration scheme, distribution of $Y \mid X$, and loss function that is twice differentiable. The only requirement is that the noise must be additive and with zero mean. We now demonstrate this result on a smooth approximation of the miscoverage loss, formulated as:

$$L^{sm}(y, [a, b]) = \frac{2}{1 + e^{-\left(2\frac{y-a}{b-a}-1\right)^2}} \tag{9}$$

Corollary 5 *Let $\widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})$ be a prediction interval. Denote the smooth miscoverage over the noisy labels as $\alpha := \mathbb{E}\left[L^{sm}(\tilde{Y}_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}}))\right]$. Under the assumptions of Proposition 6, the risk of the clean labels is bounded by*

$$\alpha - \frac{1}{2}Q \cdot \text{Var}(Z) \leq \mathbb{E}\left[L^{sm}(Y_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}}))\right] \leq \alpha - \frac{1}{2}q \cdot \text{Var}(Z),$$

where $q = \mathbb{E}_X\left[\min_y \frac{\partial^2}{\partial y^2}L^{sm}(y, \widehat{\mathcal{C}}_{\text{noisy}}(X))\right]$ and $Q = \mathbb{E}_X\left[\max_y \frac{\partial^2}{\partial y^2}L^{sm}(y, \widehat{\mathcal{C}}_{\text{noisy}}(X))\right]$ are known constants.

We now build upon Corollary 5 and establish a lower bound for the coverage rate achieved by intervals calibrated with corrupted labels.

Proposition 7 *Let $\widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})$ be a prediction interval. Suppose that the labels are corrupted by the function g^{add} from (3) with a noise Z that satisfies $\mathbb{E}[Z] = 0$, then*

$$\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \geq 1 - \frac{\mathbb{E}[L^{\text{sm}}(\tilde{Y}_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}}))] - 0.5P \cdot \text{Var}(Z)}{Q},$$

where P, Q are tunable constants.

In Appendix A.6.2 we give additional details about this result as well as formulate a stronger version of it that provides a tighter coverage bound. Finally, in Appendix A.6.3 we provide an additional miscoverage bound which is more informative and tight for smooth densities of $Y \mid X = x$. Table 2 summarizes all different risk-control settings with their corresponding bounds. Lastly, in Appendix A.6.4 we analyze label-noise robustness in settings where the response Y is a matrix, as in image-to-image regression tasks.

3.4 Online Learning Under Label Noise

In this section, we focus on an online learning setting and show that all theoretical results presented thus far also apply to the online framework. Here, the data is given as a stream $(X_t, \tilde{Y}_t)_{t \in \mathbb{N}}$ in a sequential fashion. Crucially, we have access only to the noisy labels $\tilde{Y}_t$, and the clean labels Y_t are unavailable throughout the entire learning process. At time stamp $t \in \mathbb{N}$, our goal is to construct a prediction set $\widehat{\mathcal{C}}^t_{\text{noisy}}$ given on all previously observed

Table 2: Summary of coverage bounds for different risk control tasks and different noise models

Task	Noise model	Bounds
Multi-label classification	Random multi-label flip: $g^{\text{vector-flip}}$ in (8)	$\mathbb{E}\left[L^{\text{FNR}}\left(Y_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right)\right] \leq \alpha$ (Proposition 4)
Segmentation	Random multi-label flip: $g^{\text{vector-flip}}$ in (8)	$\mathbb{E}\left[L^{\text{FNR}}\left(Y_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right)\right] \leq \alpha$ (Proposition 5)
Regression with a general loss	Additive: g^{add} in (3)	$\mathbb{E}\left[L\left(Y_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right)\right] \leq \alpha$ (Proposition 6)
Regression with miscoverage loss	Additive: g^{add} in (3)	$\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \geq$ $1 - \frac{\mathbb{E}[L^{\text{sm}}(\tilde{Y}_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}}))] - 0.5P \cdot \text{Var}(Z)}{Q}$ (Proposition 7)

samples $(X_{t'}, \tilde{Y}_{t'})_{t'=1}^{t-1}$ along with the test feature vector X_t that achieves a long-range risk controlled at a user-specified level α, i.e.,

$$\mathcal{R}(\widehat{\mathcal{C}}) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} L_t(Y_t, \widehat{\mathcal{C}}_{\text{noisy}}^t(X_t)) = \alpha. \tag{10}$$

Importantly, in this online learning setting, the loss function L_t might be time-dependent and may vary throughout the learning process. There have been developed calibration schemes that generate uncertainty sets with statistical guarantees in online settings in the sense of (10). A popular approach is Adaptive conformal inference (`ACI`), proposed by Gibbs and Candes (2021), which is an innovative online calibration scheme that constructs prediction sets with a pre-specified coverage rate, in the sense of (10) with the choice of the miscoverage loss. In contrast to `ACI`, Rolling risk control (`Rolling RC`) (Feldman et al., 2023b) extends `ACI` by providing a guaranteed control of a general risk that may go beyond the binary loss. The main idea behind both these methods is to tune the calibration parameter that controls the size of the prediction set according to the coverage or risk level achieved in the past. See Appendix C.2 and Appendix C.3 for more details on `ACI` and `RRC`, respectively. Yet, the guarantees of these approaches are invalidated when applied using corrupted data. Nonetheless, we argue that uncertainty sets constructed using corrupted data attain conservative risk in online settings under the requirements for offline label-noise robustness presented thus far.

Proposition 8 *Suppose that for all $t \in \mathbb{N}$ and $x \in \mathcal{X}$ we have a conservative loss:*

$$\mathbb{E}\left[L_t\left(Y_t, \widehat{\mathcal{C}}_{\text{noisy}}^t(X_t)\right)\Big| X_t = x\right] \leq \mathbb{E}\left[L_t\left(\tilde{Y}_t, \widehat{\mathcal{C}}_{\text{noisy}}^t(X_t)\right)\Big| X_t = x\right].$$

Then,

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} L_t(Y_t, \widehat{\mathcal{C}}_{\text{noisy}}^t(X_t)) \leq \alpha,$$

where α is the risk computed over the noisy labels.

The proof is given in Appendix A.7.1. In words, Proposition 8 states that if the expected loss at every timestamp is conservative, then the risk over long-range windows in time is guaranteed to be valid. Practically, this proposition shows that all theoretical results presented thus far apply also in the online learning setting. We now demonstrate this result in two settings: online classification and segmentation and show that valid risk is obtained under the assumptions of Proposition 2 and Proposition 5, respectively.

Corollary 6 (Valid risk in online classification settings) *Suppose that the distributions of $Y_t \mid X_t$ and $\tilde{Y}_t \mid X_t$ satisfy the assumptions in Proposition 2 for all $t \in \mathbb{N}$. If $\widehat{\mathcal{C}}_{\mathrm{noisy}}^t(x)$ contains the most likely labels, in the sense of Proposition 2, for every $t \in \mathbb{N}$ and $x \in \mathcal{X}$, then*

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{Y_t \notin \widehat{\mathcal{C}}_{\mathrm{noisy}}^t(X_t)\} \leq \alpha.$$

Corollary 7 (Valid risk in online segmentation settings) *Suppose that the distributions of $Y_t \mid X_t$ and $\tilde{Y}_t \mid X_t$ satisfy the assumptions in Proposition 5 for all $t \in \mathbb{N}$. If $\widehat{\mathcal{C}}_{\mathrm{noisy}}^t(x)$ contains the most likely labels, in the sense of Proposition 5, for every $t \in \mathbb{N}$ and $x \in \mathcal{X}$ then*

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} L^{FNR}(Y_t, \widehat{\mathcal{C}}_{\mathrm{noisy}}^t(X_t)) \leq \alpha.$$

Finally, in Appendix A.7.2 we analyze the effect of label noise on the *miscoverage counter loss* (Feldman et al., 2023b). This loss assesses conditional validity in online settings by counting occurrences of consecutive miscoverage events. In a nutshell, Proposition A.7 claims that with access to corrupted labels, the miscoverage counter is valid when the miscoverage risk is valid. This is an interesting result, as it connects the validity of the miscoverage counter to the validity of the miscoverage loss, where the latter is guaranteed under the conditions established in Section 2.3.

4. Experiments

Software is available online at `https://github.com/bat-sheva/Conformal-Label-Noise`, with all code needed to reproduce the numerical experiments.

4.1 Synthetic Classification

In this section, we focus on multi-class classification problems, where we study the validity of conformal prediction using different types of label noise distributions, described below.

Class-independent noise. This noise model, which we call `uniform flip`, randomly flips the ground truth label into a *different* one with probability ϵ. Notice that this noise

model slightly differs from the random flip g^{flip} from (6), since in this `uniform flip` setting, a label cannot be flipped to the original label. Nonetheless, Proposition 2 states that the coverage achieved by an oracle classifier is guaranteed to increase in this setting as well.

Class-dependent noise. In contrast to `uniform flip` noise, here, we consider a more challenging setup in which the probability of a label to be flipped depends on the ground truth class label Y. Such a noise label is often called Noisy at Random (NAR) in the literature, where certain classes are more likely to be mislabeled or confused with similar ones. Let T be a row stochastic transition matrix of size $K \times K$ such that $T_{i,j}$ is the probability of a point with label i to be swapped with label j. In what follows, we consider three possible strategies for building the transition matrix T. (1) `Confusion matrix` (Algan and Ulusoy, 2020): we define T as the oracle classifier's confusion matrix, up to a proper normalization to ensure the total flipping probability is ϵ. We provide a theoretical study of this case in Appendix A.3.2. (2) `Rare to most frequent class` (Xu et al., 2019): here, we flip the labels of the least frequent class with those of the most frequent class. This noise model is not uncommon in medical applications: imagine a setting where only a small fraction of the observations are abnormal, and thus likely to be annotated as the normal class. If switching between the rare and most frequent labels does not lead to a total probability of ϵ, we move to the next least common class, and so on.

To set the stage for the experiments, we generate synthetic data with $K = 10$ classes as follows. The features $X \in \mathbb{R}^d$ follow a standard multivariate Gaussian distribution of dimension $d = 100$. The conditional distribution of $Y \mid X$ is multinomial with weights $w_j(x) = \exp((x^\top B)_j)/\sum_{i=1}^{K} \exp((x^\top B)_i)$, where $B \in \mathbb{R}^{d \times K}$ whose entries are sampled independently from the standard normal distribution. In our experiments, we generate a total of $60,000$ data points, where $50,000$ are used to fit a classifier, and the remaining ones are randomly split to form calibration and test sets, each of size $5,000$. The training and calibration data are corrupted using the label noise models we defined earlier, with a fixed flipping probability of $\epsilon = 0.05$. Of course, the test set is not corrupted and contains the ground truth labels. We apply conformal prediction using both the HPS and the APS score functions, with a target coverage level $1 - \alpha$ of 90%. We use two predictive models: a two-layer neural network and an oracle classifier that has access to the conditional distribution of $\tilde{Y} \mid X$. Finally, we report the distribution of the coverage rate as in (2) and the prediction set sizes across 100 random splits of the calibration and test data. As a point of reference, we repeat the same experimental protocol described above on clean data; in this case, we do not violate the i.i.d. assumption required to grant the marginal coverage guarantee in (1).

The results are depicted in Figure 6. As expected, in the clean setting all conformal methods achieve the desired coverage of 90%. Under the `uniform flip` noise model, the coverage of the oracle classifier increases to around 94%, supporting our theoretical results from Section 2.3.3. The neural network model follows a similar trend. Although not supported by a theoretical guarantee, when corrupting the labels using the more challenging `confusion matrix` noise, we can see a conservative behavior similar to the `uniform flip`. By contrast, under the `rare to most frequent class` noise model, we can see a decrease in coverage, which is in line with our disclaimer from Section 2.4. Yet, observe how the
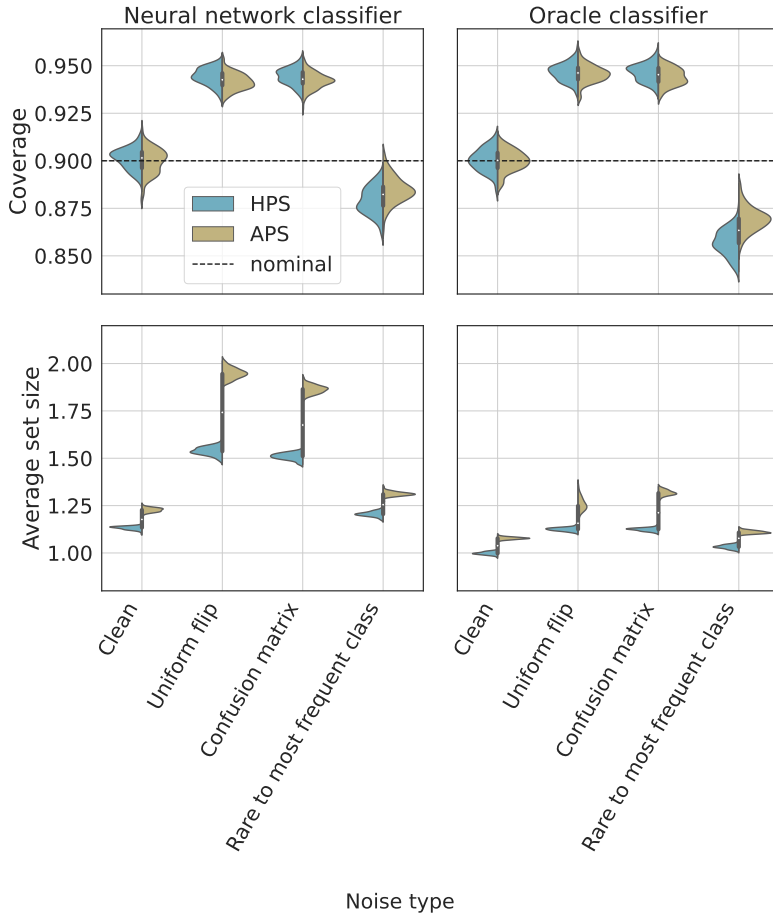
Figure 6: **Effect of label noise on synthetic multi-class classification data**. Performance of conformal prediction sets with target coverage $1 - \alpha = 90\%$, using a noisy training set and a noisy calibration set. Top: Marginal coverage; Bottom: Average size of predicted sets. The results are evaluated over 100 independent experiments and the gray bar represents the interquartile range.

APS score tends to be more robust to label noise than HPS, which emphasizes the role of the score function.

In Appendix B.3 we provide additional experiments with adversarial noise models that more aggressively reduce the coverage rate. Such adversarial cases are more pathological and less likely to occur in real-world settings, unless facing a malicious attacker.

4.2 Regression

Similarly to the classification experiments, we study two types of noise distributions.
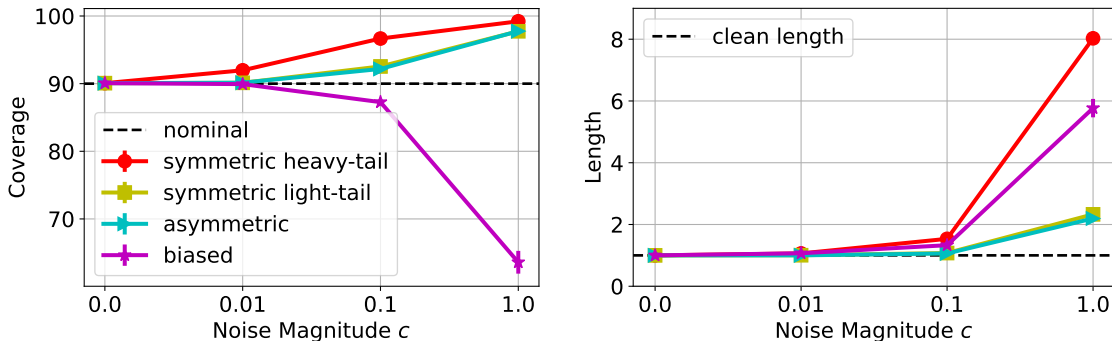
Figure 7: **Response-independent noise.** Performance of conformal prediction intervals with target coverage $1 - \alpha = 90\%$, using a noisy training set and a noisy calibration set. Left: Marginal coverage; Right: Length of predicted intervals (divided by the average clean length) using symmetric, asymmetric and biased noise with a varying magnitude. The results are evaluated over 50 independent experiments, with error bars showing one standard deviation. The standard deviation of Y (without noise) and the square root of $\mathbb{E}\left[\mathrm{Var}\left[Y \mid X\right]\right]$ are both approximately 2.6.

Response-independent noise. We consider an additive noise of the form: $\tilde{Y} = Y + c \cdot Z$, where c is a parameter that allows us to control the noise level. The noise component Z is a random variable sampled from the following distributions. (1) `Symmetric light tailed`: standard normal distribution; (2) `Symmetric heavy tailed`: t-distribution with one degree of freedom; (3) `Asymmetric`: standard Gumbel distribution, normalized to have zero mean and unit variance; and (4) `Biased`: positive noise formulated as the absolute value of the `symmetric heavy tailed noise` above.

Response-dependent noise. Analogously to the class-dependent noise from Section 4.1, we define more challenging noise models as follows. (1) `Contractive`: this corruption pushes the ground truth response variables towards their mean. Formally, $\tilde{Y}_i = Y_i - \left(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i\right) \cdot U$, where U is a random uniform variable defined on the segment [0,0.5], and n is the number of calibration points. (2) `Dispersive`: this noise introduces some form of a dispersion effect on the ground truth response, which takes the opposite form of the `contractive` model, given by $\tilde{Y}_i = Y_i + \left(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i\right) \cdot U$.

Having defined the noise models, we turn to describe the data-generating process. We simulate a 100-dimensional X whose entries are sampled independently from a uniform distribution on the segment $[0, 5]$. Following Romano et al. (2019), the response variable is generated as follows:

$$Y \sim \mathrm{Pois}(\sin^2(\bar{X}) + 0.1) + 0.03 \cdot \bar{X} \cdot \eta_1 + 25 \cdot \eta_2 \cdot \mathbb{1}\left\{U < 0.01\right\}, \qquad (11)$$

where $\bar{X}$ is the mean of the vector X, and $\mathrm{Pois}(\lambda)$ is the Poisson distribution with mean λ. Both η_1 and η_2 are i.i.d. standard Gaussian variables, and U is a uniform random variable on $[0, 1]$. The right-most term in (11) creates a few but large outliers. Figure 17

in the appendix illustrates the effect of the noise models discussed earlier on data sampled from (11).

We apply conformal prediction with the CQR score (Romano et al., 2019) for each noise model as follows. First, we fit a quantile random forest model on $8,000$ noisy training points; we then calibrate the model using $2,000$ fresh noisy samples; and, lastly, test the performance on additional $5,000$ clean, ground truth samples. The results are summarized in Figures 7 and 8. Observe how the prediction intervals tend to be conservative under `symmetric`, both for light- and heavy-tailed noise distributions, `asymmetric`, and `dispersive` corruption models. Intuitively, this is because these noise models increase the variability of Y; in Proposition 1 we prove this formally for any `symmetric` independent noise model, whereas here we show this result holds more generally even for response-dependent noise. By contrast, the prediction intervals constructed under the `biased` and `contractive` corruption models tend to under-cover the response variable. This should not surprise us: following Figure 17(c), the `biased` noise shifts the data 'upwards', and, consequently, the prediction intervals are undesirably pushed towards the positive quadrants. Analogously, the `contractive` corruption model pushes the data towards the mean, leading to intervals that are too narrow. Figure 20 in the appendix illustrates the scores achieved when using the different noise models and the 90%'th empirical quantile of the CQR scores. This figure supports the behavior witnessed in Figures 7, 18 and 19: over-coverage is achieved when $\hat{q}_{\mathrm{noisy}}$ is larger than $\hat{q}_{\mathrm{clean}}$, and under-coverage is obtained when $\hat{q}_{\mathrm{noisy}}$ is smaller.

In Appendix B.4 we study the effect of the predictive model on the coverage property, for all noise models. To this end, we repeat similar experiments to the ones presented above, however, we now fit the predictive model on clean training data; the calibration data remains noisy. We also provide an additional adversarial noise model that reduces the coverage rate, but is unlikely to appear in real-world settings. Figures 18 and 19 in the appendix depict a similar behaviour for most noise models, except the `biased` noise for which the coverage requirement is not violated. This can be explained by the improved estimation of the low and high conditional quantiles, as these are fitted on clean data and thus less biased.

4.3 Multi-label Classification

In this section, we analyze conformal risk control in a multi-label classification task. For this purpose, we use the CIFAR-100N data set (Wei et al., 2022), which contains 50K colored images. Each image belongs to one of a hundred fine classes that are grouped into twenty mutually exclusive coarse super-classes. Furthermore, every image has a noisy and a clean label, where the noise rate of the fine categories is 40% and of the coarse categories is 25%. We turn this single-label classification task into a multi-label classification task by merging four random images into a 2 by 2 grid. Every image is used once in each position of the grid, and therefore this new data set consists of 50K images, where each is composed of four sub-images and thus associated with up to four labels. Figure 9 displays a visualization of this new variant of the CIFAR-100N data set.
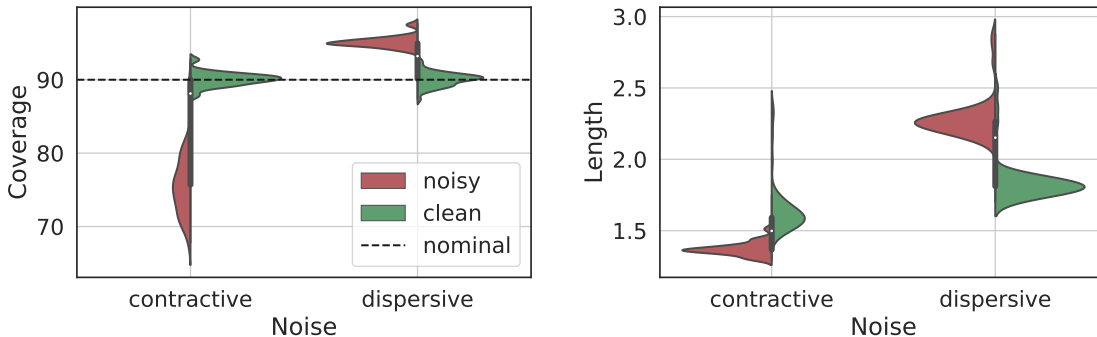
Figure 8: **Dispersive versus contractive noise regression experiment.** Performance of conformal prediction intervals with target coverage $1-\alpha = 90\%$, using a noisy training set and a noisy calibration set. Left: Marginal coverage; Right: Length of predicted intervals. The results are evaluated over 50 independent experiments and the gray bar represents the interquartile range.
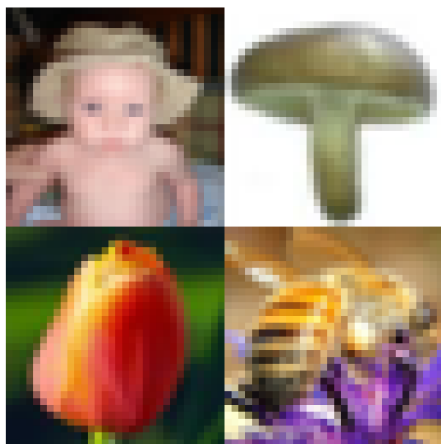
True labels:
baby, mushroom, tulip, bee.

Noisy labels:
baby, mushroom, sweet pepper, bee.

Figure 9: Visualization of the multi-label CIFAR-100N data.

We fit a TResNet (Ridnik et al., 2021) model on 40k noisy samples and calibrate it using 2K noisy samples with conformal risk control, as outlined in (Angelopoulos et al., 2024, Section 3.2). We control the false-negative rate (FNR) defined in (7) at different levels and measure the FNR obtained over clean and noisy versions of the test set, which contains 8k samples. We conducted this experiment twice: once with the fine-classed labels and once with the super-classed labels. Figure 10 presents the results in both settings, showing that the risk obtained over the clean labels is valid for every nominal level. Importantly, this corruption setting violates the assumptions of Proposition 4, as the positive label count may vary across different noise instantiations. This experiment reveals that valid risk can be achieved in the presence of label noise even when the corruption model violates the requirements of our theory.
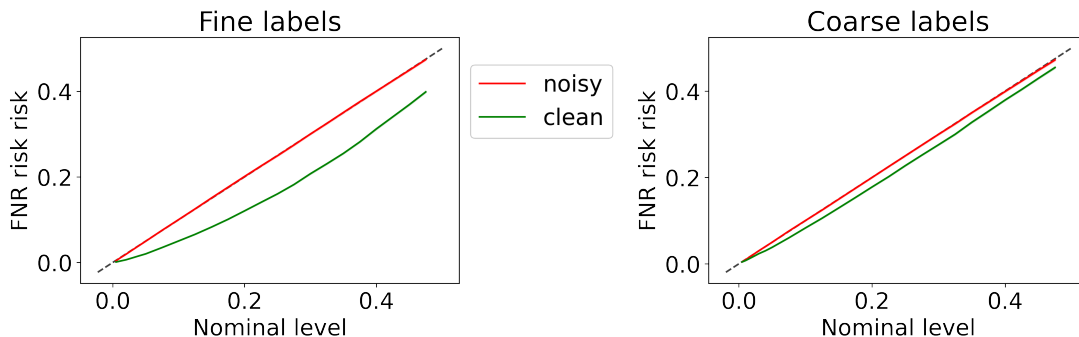
Figure 10: FNR achieved over noisy (red) and clean (green) test sets of the on CIFAR-100N data set. Left: fine labels. Right: coarse labels. The calibration scheme is applied using noisy annotations in all settings. Results are averaged over 50 trials.

4.4 Segmentation

In this section, we follow a common artificial label corruption methodology, following Zhao and Gomes (2021); Kumar et al. (2020), and analyze three corruption setups that are special cases of the vector-flip noise model from (8): independent, dependent, and partial. In the independent setting, each pixel's label is flipped with probability β, independently of the others. In the dependent setting, however, two rectangles in the image are entirely flipped with probability β, and the other pixels are flipped independently with probability β. Finally, in the partial noise setting, only one rectangle in the image is flipped with probability β, and the other pixels are unchanged.
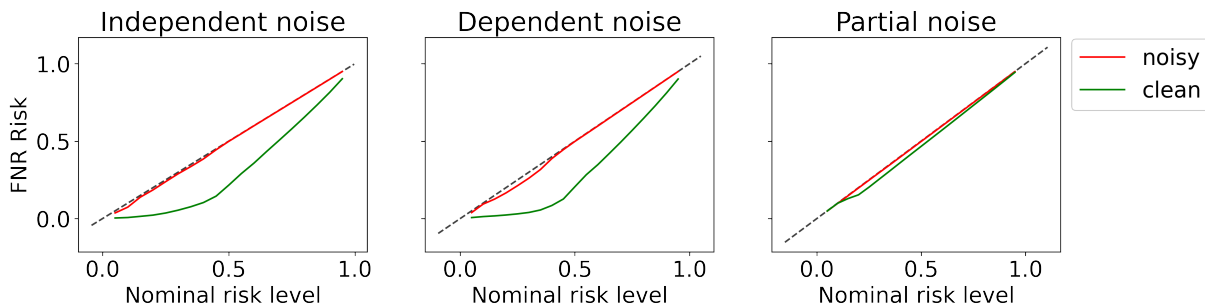

Figure 11: FNR on a polyp segmentation data set, achieved over noisy (red) and clean (green) test sets. Left: independent noise. Middle: dependent noise. Right: partial noise. The predictive model is calibrated using noisy annotations in all settings, where the noise level is set to $\beta = 0.1$. Results are averaged over 1000 trials.

We experiment on a polyp segmentation task, pooling data from several polyp data sets: Kvasir, CVC-ColonDB, CVC-ClinicDB, and ETIS-Larib. We consider the annotations given in the data as ground-truth labels and artificially corrupt them according to the corruption setups described above, to generate noisy labels. We use PraNet (Fan et al., 2020) as a base

model and fit it over 1450 noisy samples. Then, we calibrate it using 500 noisy samples with conformal risk control, as outlined in (Angelopoulos et al., 2024, Section 3.2) to control the false-negative rate (FNR) from (7) at different levels. Finally, we evaluate the FNR over clean and noisy versions of the test set, which contains 298 samples, and report the results in Figure 11. This figure indicates that conformal risk control is robust to label noise, as the constructed prediction sets achieve conservative risk in all experimented noise settings. This is not a surprise, as it is guaranteed by Propositions 5 and A.4.

4.5 Regression Risk Bounds

We now turn to demonstrate the coverage rate bounds derived in Section 3.3.3 on real and synthetic regression data sets. We examine two real benchmarks: `meps_19` and `bio` used in (Romano et al., 2019), and one synthetic data set that was generated from a bimodal density function with a sharp slope, as visualized in Figure 12. The simulated data was deliberately designed to invalidate the assumptions of our label-noise robustness requirements in Proposition 1. Consequentially, prediction intervals that do not cover the two peaks of the density function might undercover the true outcome, even if the noise is dispersive. Therefore, this gap calls for our distribution-free risk bounds from Section 3.3.3, which are applicable in this setup, in contrast to Proposition 1. The former approach can be used to assess the worst risk level that may be obtained in practice. We consider the labels given
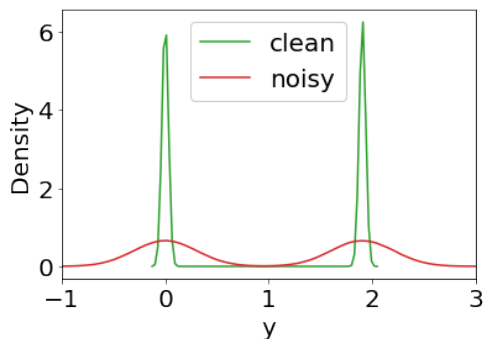

Figure 12: Visualization of the marginal density function of the adversarial synthetic data.

in the real and synthetic data sets as ground truth and artificially corrupt them according to the additive noise model (3). The added noise is independently sampled from a normal distribution with mean zero and variance $0.1\text{Var}(Y)$.

For each data set and nominal risk level α, fit a quantile regression model on 12K samples and learn the $\alpha/2, 1 - \alpha/2$ conditional quantiles of the noisy labels. Then, we calibrate its outputs using another 12K samples of the data with conformal risk control, by Angelopoulos et al. (2024), to control the smooth miscoverage (9) at level α. Finally, we evaluate the performance on the test set which consists of 6K samples. We also compute the smooth miscoverage risk bounds according to Corollary 5 with a noise variance set to 0.1.

Figure 13 presents the risk bound along with the smooth miscoverage obtained over the clean and noisy versions of the test set. This figure indicates that conformal risk control generates invalid uncertainty sets when applied on the simulated noisy data, as we anticipated. Additionally, this figure shows that the proposed risk bounds are valid and tight, meaning that these are informative and effective. Moreover, this figure highlights the main advantage of the proposed risk bounds: their validity is universal across all distributions of the response variable and the noise component. Lastly, we note that in Appendix B.7 we repeat this experiment with the miscoverage loss and display the miscoverage bound derived from Corollary 7.
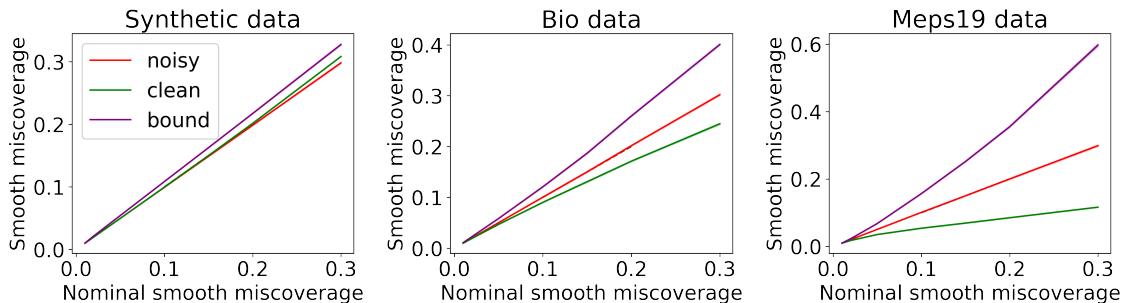

Figure 13: Smooth miscoverage rate achieved over noisy (red) and clean (green) test sets. The calibration scheme is applied using noisy annotations to control the smooth miscoverage level. Results are averaged over 10 random splits of the calibration and test sets.

4.6 Online Learning

This section studies the effect of label noise on uncertainty quantification methods in an online learning setting, as formulated in Section 3.4. We experiment on a depth estimation task (Geiger et al., 2013), where the objective is to predict a depth map given a colored image. In other words, $X \in \mathbb{R}^{W \times H \times 3}$ is an input RGB image of size $W \times H$ and $Y \in \mathbb{R}^{W \times H}$ is its corresponding depth map. We consider the original depth values given in this data as ground truth and artificially corrupt them according to the additive noise model defined in (3) to produce noisy labels. Specifically, we add to each depth pixel an independent random noise drawn from a normal distribution with zero mean and 0.7 variance. Here, the depth uncertainty of the i,j pixel is represented by a prediction interval $C^{i,j}(X) \subseteq \mathbb{R}$. Ideally, the estimated intervals should contain the correct depth values at a pre-specified level $1 - \alpha$. In this high-dimensional setting, this requirement is formalized as controlling the *image miscoverage* loss, defined as:

$$L^{\mathrm{im}}(Y, C(X)) = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} 1\{Y^{i,j} \notin C^{i,j}(X)\}. \tag{12}$$

In words, the image miscoverage loss measures the proportion of depth values that were not covered in a given image.

For this purpose, we employ the calibration scheme `Rolling RC` (Feldman et al., 2023b), which constructs uncertainty sets in an online setting with a valid risk guarantee in the sense of (10). We follow the experimental protocol outlined in (Feldman et al., 2023b, Section 4.2) and apply `Rolling RC` with an exponential stretching to control the image miscoverage loss at different levels on the observed, noisy, labels. We use LeReS (Yin et al., 2021) as a base model, which was pre-trained on a clean training set that corresponds to timestamps 1,...,6000. We continue training it and updating the calibration scheme in an online fashion on the following 2000 timestamps. We consider these samples, indexed by 6001 to 8000, as a validation set and use it to choose the calibration's hyperparameters, as explained in (Feldman et al., 2023b, Section 4.2). Finally, we continue the online procedure on the test samples whose indexes correspond to 8001 to 10000, and measure the performance on the clean and noisy versions of this test set.

Figure 14 displays the risk achieved by this technique over the clean and corrupted labels. This figure indicates that `Rolling RC` attain valid image miscoverage over the unknown noiseless labels. This is not a surprise, as it is supported by Proposition A.6 which guarantees conservative image miscoverage under label noise in an offline setting, and Proposition 8 which states that the former result applies to an online learning setting as well.
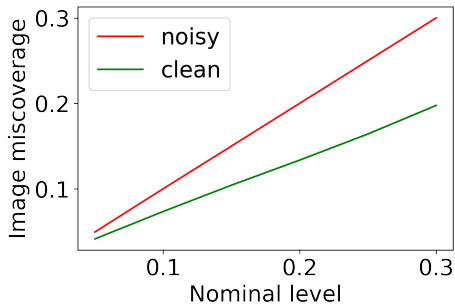

Figure 14: Image miscoverage achieved by `Rolling RC`. Results are averaged over 10 random trials.

27

5. Discussion

5.1 Related Work

Label noise independently of conformal prediction has been well-studied, especially for training more robust predictive models; see, for example (Angluin and Laird, 1988; Frénay and Verleysen, 2013; Tanno et al., 2019; Algan and Ulusoy, 2020; Kumar et al., 2020). Recently, there has been a body of work studying the statistical properties of conformal prediction (Lei et al., 2018; Barber, 2020) and its performance under deviations from exchangeability (Tibshirani et al., 2019; Podkopaev and Ramdas, 2021). This line of work is relevant to us since the label noise setting violates the exchangeability assumption between the training and testing data, where the latter is clean while the former is noisy. However, these works cannot be applied in our setting since they assume covariate or label shift only. The work by Barber et al. (2023) refers to any distribution shift, and we build upon some of its results in our general disclaimer in Section 2.4. Another relevant work is (Farinhas et al., 2024), which studies the effect of a general distribution shift on the obtained risk of risk-controlling techniques, thus extending the work of by Barber et al. (2023) to more general loss functions than the miscoverage loss. Additionally, Angelopoulos et al. (2024) analyzes their proposed risk-controlling method in a covariate shift and a general distribution shift settings. Close works to ours include (Stutz et al., 2023), which analyses the performance of conformal prediction under ambiguous ground truth, and Cauchois et al. (2022), that studies conformal prediction with weak supervision, which could be interpreted as a type of noisy label.

Lastly, a follow-up work to ours has recently been published (Sesia et al., 2023). Similarly to our work, it begins by studying the effect of label noise on the coverage achieved by standard conformal prediction. Then, an explicit factor that depicts the inflation or deflation of the coverage is estimated to adjust the desired coverage rate. The theoretical analysis requires no assumptions on the contamination process, but in order to achieve an applicable method that can automatically adapt to label noise, some mild assumptions on the relation between the clean and observable labels are used. Indeed, the presented experiments demonstrate less conservative coverage compared to the standard method. However, there are two key distinctions between this work and ours. It focuses on controlling the coverage rate in classification tasks, whilst our analysis extends to regression tasks, general risk control, and online settings. Additionally, it aims to modify the calibration algorithm to account for label noise, whereas our goal is to test the limits of conformal prediction in the label noise setting and reveal the conditions on the scores, noise models, and predictive models, under which the standard algorithm remains valid despite the presence of noisy labels.

5.2 Future Research Direction

Our work raises many new questions. First, one can try and define a score function that is more robust to label noise, continuing the line of Gendler et al. (2021); Frénay and Verleysen (2013); Cheng et al. (2022). Second, an important remaining question is how to achieve

exact risk control on the clean labels using minimal information about the noise model. Lastly, it would be interesting to analyze the robustness of alternative conformal methods such as cross-conformal and jackknife+ (Vovk, 2015; Barber et al., 2021) that do not require data-splitting.

Acknowledgments

Appendix A. Mathematical Proofs

A.1 General Analysis

We begin by proving Theorem 1.

Proof [Proof of Theorem 1] Our assumption states that

$$\mathbb{P}(\tilde{s}_{\text{test}} \leq t) \leq \mathbb{P}(s_{\text{test}} \leq t).$$

Note that the probability is only taken over $\tilde{s}_{\text{test}}$. Since $\hat{q}_{\text{noisy}}$ is constant (measurable) with respect to this probability, we have that, for any $\alpha \in (0, 1)$,

$$\mathbb{P}(s_{\text{test}} \leq \hat{q}_{\text{noisy}}) \geq \mathbb{P}(\tilde{s}_{\text{test}} \leq \hat{q}_{\text{noisy}}) \geq 1 - \alpha.$$

This implies that $Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})$ with probability at least $1 - \alpha$, completing the proof of the lower bound.

Regarding the upper bound, by the same argument,

$$\mathbb{P}(s_{\text{test}} \leq \hat{q}_{\text{noisy}}) \leq \mathbb{P}(\tilde{s}_{\text{test}} \leq \hat{q}_{\text{noisy}}) + u \leq 1 - \alpha + \frac{1}{n+1} + u.$$

∎

A.2 Regression

A.2.1 GENERAL REGRESSION RESULT

Here we provide an extension to Proposition 1 and prove it.

Theorem A.1 *Suppose an additive noise model g^{add} with a noise that has mean 0. Denote the prediction interval as $C(x) = [a_x, b_x]$. If for all $x \in \mathcal{X}$ density of $Y \mid X = x$ is peaked inside the interval:*

1. *$\forall \varepsilon \geq 0 : f_{Y|X=x}(b_x + \varepsilon) \leq f_{Y|X=x}(b_x - \varepsilon)$.*

2. *$\forall \varepsilon \geq 0 : f_{Y|X=x}(a_x - \varepsilon) \leq f_{Y|X=x}(a_x + \varepsilon)$.*

then, we obtain valid conditional coverage:

$$\mathbb{P}(Y \in C(x) \mid X = x) \geq \mathbb{P}(\tilde{Y} \in C(x) \mid X = x).$$

Proof For ease of notation, we omit the conditioning on $X = x$. In other words, we treat Y as $Y \mid X = x$ for some $x \in \mathcal{X}$. We begin by showing $\mathbb{P}(\tilde{Y} \leq b) \geq \mathbb{P}(Y \leq b)$.

$$\mathbb{P}(\tilde{Y} \leq b)$$
$$= \mathbb{P}(Y + \varepsilon \leq b)$$
$$= \mathbb{P}(Y + \varepsilon \leq b \mid \varepsilon \geq 0)\mathbb{P}(\varepsilon \geq 0) + \mathbb{P}(Y + \varepsilon \leq b \mid \varepsilon \leq 0)\mathbb{P}(\varepsilon \leq 0)$$
$$= \frac{1}{2}\mathbb{P}(Y + \varepsilon \leq b \mid \varepsilon \geq 0) + \frac{1}{2}\mathbb{P}(Y + \varepsilon \leq b \mid \varepsilon \leq 0)$$
$$= \frac{1}{2}\mathbb{P}(Y \leq b - \varepsilon \mid \varepsilon \geq 0) + \frac{1}{2}\mathbb{P}(Y \leq b + \varepsilon \mid \varepsilon \geq 0)$$
$$= \frac{1}{2}\mathbb{E}_{\varepsilon \geq 0}\left[\mathbb{P}(Y \leq b - \varepsilon) + \mathbb{P}(Y \leq b + \varepsilon)\right]$$
$$= \mathbb{P}(Y \leq b) + \frac{1}{2}\mathbb{E}_{\varepsilon \geq 0}\left[\mathbb{P}(Y \leq b - \varepsilon) - \mathbb{P}(Y \leq b) + \mathbb{P}(Y \leq b + \varepsilon) - \mathbb{P}(Y \leq b)\right]$$
$$= \mathbb{P}(Y \leq b) + \frac{1}{2}\mathbb{E}_{\varepsilon \geq 0}\left[\mathbb{P}(Y \leq b + \varepsilon) - \mathbb{P}(Y \leq b) - (\mathbb{P}(Y \leq b) - \mathbb{P}(Y \leq b - \varepsilon))\right]$$
$$= \mathbb{P}(Y \leq b) + \frac{1}{2}\mathbb{E}_{\varepsilon \geq 0}\left[\mathbb{P}(b \leq Y \leq b + \varepsilon) - \mathbb{P}(b - \varepsilon \leq Y \leq b)\right]$$
$$\leq \mathbb{P}(Y \leq b) + \frac{1}{2}\mathbb{E}_{\varepsilon \geq 0}\left[0\right]$$
$$= \mathbb{P}(Y \leq b).$$

The last inequality follows from the assumption that $\forall \varepsilon \geq 0 : f_Y(b + \varepsilon) \leq f_Y(b - \varepsilon)$. The proof for $\mathbb{P}(\tilde{Y} \geq a) \leq \mathbb{P}(Y \geq a)$ is similar and hence omitted. We get that:

$$\mathbb{P}(\tilde{Y} \in C(x))$$
$$= \mathbb{P}(a \leq \tilde{Y} \leq b)$$
$$= \mathbb{P}(\tilde{Y} \leq b) - \mathbb{P}(\tilde{Y} \geq a)$$
$$\leq \mathbb{P}(Y \leq b) - \mathbb{P}(Y \geq a) \text{ (follows from the above)}$$
$$= \mathbb{P}(a \leq Y \leq b)$$
$$= \mathbb{P}(Y \in C(x)).$$

∎

We now turn to prove Proposition 1 using Theorem A.1

Proof Since the density of $Y \mid X = x$ is symmetric and unimodal for all $x \in \mathcal{X}$ it is peaked inside any prediction interval that contains its median. Therefore, the prediction interval achieves valid conditional coverage:

$$\forall x \in \mathcal{X} : \mathbb{P}(\tilde{Y} \in C(x) \mid X = x) \geq \mathbb{P}(Y \in C(x) \mid X = x).$$

By taking the expectation over $X \sim P_X$ we obtain valid marginal coverage. ∎

A.3 Classification

A.3.1 General Classification Result

We commence by providing an extension to Proposition 2 and proving it.

Theorem A.2 *Suppose that $\widehat{\mathcal{C}}_{\mathrm{noisy}}(x) \subseteq \{1, ..., K\}$ is a prediction set. Denote $\beta := \mathbb{P}(\tilde{Y} \in \widehat{\mathcal{C}}_{\mathrm{noisy}}(x) \mid X = x)$. First, the coverage rate achieved over the clean labels is upper bounded by:*

$$\mathbb{P}(Y \in \widehat{\mathcal{C}}_{\mathrm{noisy}}(x) \mid X = x) \leq \beta + \frac{1}{2} \sum_{i=1}^{K} |\mathbb{P}(\tilde{Y} = i \mid X = x) - \mathbb{P}(Y = i \mid X = x)|,$$

Further suppose that for all $x \in \mathcal{X}$:

$$\sum_{i \in \widehat{\mathcal{C}}_{\mathrm{noisy}}(x)} \left(\mathbb{P}(Y = i \mid X = x) - \mathbb{P}(\tilde{Y} = i \mid X = x) \right) \geq 0.$$

Then, the coverage rate is lower bounded by:

$$\mathbb{P}(\tilde{Y} \in \widehat{\mathcal{C}}_{\mathrm{noisy}}(x) \mid X = x) \leq \mathbb{P}(Y \in \widehat{\mathcal{C}}_{\mathrm{noisy}}(x) \mid X = x).$$

Proof First, for ease of notation, we omit the conditioning on x and consider a prediction $C \subseteq \{1, ..., K\}$. Notice that if $C = \emptyset$ the proposition is trivially satisfied.

We begin by proving the following lower bound $\mathbb{P}(Y \in C) \geq \beta := \mathbb{P}(\tilde{Y} \in C)$. Denote: $\delta_i := \mathbb{P}(Y = i) - \mathbb{P}(\tilde{Y} = i)$. Since $\sum_{i \in C} \delta_i \geq 0$ we get that:

$$\mathbb{P}(Y \in C) = \sum_{i \in C} \mathbb{P}(Y = i) = \sum_{i \in C} \mathbb{P}(\tilde{Y} = i) + \sum_{i \in C} \delta_i \geq \sum_{i \in C} \mathbb{P}(\tilde{Y} = i) + 0 = \mathbb{P}(\tilde{Y} \in C) = \beta.$$

We now turn to prove the upper bound.

$$\sum_{i \in C} \delta_i = \frac{1}{2} \left(\sum_{i \in C} \delta_i + \sum_{i \in C} \delta_i \right)$$

$$= \frac{1}{2} \left(\sum_{i \in C} \delta_i + \sum_{i \in \{1, ..., K\} - C} -\delta_i \right)$$

$$\leq \frac{1}{2} \sum_{i=1}^{K} |\delta_i|$$

$$= \frac{1}{2} \sum_{i=1}^{K} |\mathbb{P}(\tilde{Y} = i) - \mathbb{P}(Y = i)|.$$

This gives us:

$$\mathbb{P}(Y \in C) = \mathbb{P}(\tilde{Y} \in C) + \sum_{i \in C} \delta_i \leq \mathbb{P}(\tilde{Y} \in C) + \frac{1}{2} \sum_{i=1}^{K} |\mathbb{P}(\tilde{Y} = i) - \mathbb{P}(Y = i)|.$$

And this concludes the proof.

∎

We now turn to prove Proposition 2 using Theorem A.2.

Proof We suppose without loss of generality that the labels are ranked from the most likely to the least likely, i.e., $\mathbb{P}(Y = i) \geq \mathbb{P}(Y = i+1)$. Since the prediction set contains the most likely labels, there exists some $1 \leq m \leq K$ such that the prediction set is $C = \{1, 2, ..., m\}$. We only need to show $\sum_{i=1}^{m} \left(\mathbb{P}(Y = i) - \mathbb{P}(\tilde{Y} = i) \right) \geq 0$ for $m \in \{1, ..., K\}$ and the result will follow directly from Theorem A.2. We follow the notations in the proof of Theorem A.2. First, we observe that under the assumed noise model:

$$\mathbb{P}(Y = i) \geq \frac{1}{K} \iff \mathbb{P}(Y = i) \geq \mathbb{P}(\tilde{Y} = i) \iff \delta_i \geq 0.$$

Denote by $1 \leq g \in \mathbb{N}$ the largest index for which $\mathbb{P}(Y = g) \geq \frac{1}{K}$. It follows that:

$$\forall i \in \{1, ..., K\} : i \leq g \iff \delta_i \geq 0.$$

If $m \leq g$ then all summands in $\sum_{i=1}^{m} \delta_i$ are positive since $\delta_i \geq 0$ for every $i \leq m \leq g$. Thus, $\sum_{i=1}^{m} \delta_i \geq 0$ as required. Otherwise,

$$\sum_{i=1}^{m} \delta_i = \sum_{i=1}^{g} \delta_i + \sum_{i=g+1}^{m} \delta_i \underset{\substack{i>g \iff \delta_i \leq 0}}{\geq} \sum_{i=1}^{g} \delta_i + \sum_{i=g+1}^{K} \delta_i = \sum_{i=1}^{K} \delta_i = 0.$$

Therefore, from Theorem A.2 we get:

$$\beta \leq \mathbb{P}(Y \in \widehat{\mathcal{C}}_{\text{noisy}}(x) \mid X = x) \leq \beta + \frac{1}{2} \sum_{i=1}^{K} |\mathbb{P}(\tilde{Y} = i \mid X = x) - \mathbb{P}(Y = i \mid X = x)|.$$

By taking the expectation over $X \sim P_X$ we obtain the desired marginal coverage bounds. The only non-trivial transition is marginalizing the TV-distance, which follows from the integral absolute value inequality:

$$\int_{x \in \mathcal{X}} |\mathbb{P}(\tilde{Y} = i \mid X = x) - \mathbb{P}(Y = i \mid X = x)| \geq$$
$$\left| \int_{x \in \mathcal{X}} \left(\mathbb{P}(\tilde{Y} = i \mid X = x) - \mathbb{P}(Y = i \mid X = x) \right) \right| =$$
$$\left| \mathbb{P}(\tilde{Y} = i) - \mathbb{P}(Y = i) \right|$$

∎

Finally, we turn to prove Corollary 2.

Proof We follow the notations and assumptions in Theorem A.2. We only need to show that $\frac{1}{2} \sum_{i=1}^{K} |\delta_i| \leq \varepsilon \frac{K-1}{K}$ and the result will follow directly from Theorem A.2.

In the random flip setting, $|\delta_i| = \left| \mathbb{P}(Y = i)\varepsilon - \frac{\varepsilon}{K} \right| = \varepsilon \left| \mathbb{P}(Y = i) - \frac{1}{K} \right|$. Therefore:

$$
\begin{aligned}
\frac{1}{2} \sum_{i=1}^{K} |\delta_i| &= \frac{1}{2} \sum_{i=1}^{K} \varepsilon \left| \mathbb{P}(Y = i) - \frac{1}{K} \right| \\
&= \frac{1}{2} \varepsilon \left(\sum_{i=1}^{g} \left| \mathbb{P}(Y = i) - \frac{1}{K} \right| + \sum_{i=g+1}^{K} \left| \mathbb{P}(Y = i) - \frac{1}{K} \right| \right) \\
&= \frac{1}{2} \varepsilon \left(\sum_{i=1}^{g} \left(\mathbb{P}(Y = i) - \frac{1}{K} \right) + \sum_{i=g+1}^{K} \left(\frac{1}{K} - \mathbb{P}(Y = i) \right) \right) \\
&= \frac{1}{2} \varepsilon \left(\sum_{i=1}^{g} \mathbb{P}(Y = i) - \sum_{i=1}^{g} \frac{1}{K} + \sum_{i=g+1}^{K} \frac{1}{K} - \sum_{i=g+1}^{K} \mathbb{P}(Y = i) \right) \\
&\leq \frac{1}{2} \varepsilon \left(1 - \frac{g}{K} + \frac{K - g}{K} - 0 \right) \\
&\leq \frac{1}{2} \varepsilon \left(\frac{K - g}{K} + \frac{K - g}{K} \right) \\
&\leq \varepsilon \frac{K - g}{K} \\
&\leq \varepsilon \frac{K - 1}{K}.
\end{aligned}
$$

Thus, we get:

$$
\mathbb{P}(\tilde{Y} \in C) \leq \mathbb{P}(Y \in C) \leq \mathbb{P}(\tilde{Y} \in C) + \frac{1}{2} \sum_{i=1}^{K} |\mathbb{P}(\tilde{Y} = i) - \mathbb{P}(Y = i)| \leq \mathbb{P}(\tilde{Y} \in C) + \varepsilon \frac{K - 1}{K},
$$

which concludes the proof. ∎

A.3.2 CONFUSION MATRIX

The `confusion matrix` noise model is more realistic than the `random flip`. However, there exists a score function that causes conformal prediction to fail for any non-identity confusion matrix. We define the corruption model as follows: consider a matrix T in which $(T)_{i,j} = \mathbb{P}(\tilde{Y} = j \mid Y = i)$.

$$
g^{\text{confusion}}(y) = \begin{cases} 1 & \text{w.p. } T_{1,y} \\ \dots & \\ K & \text{w.p. } T_{K,y}. \end{cases}
$$

Proposition A.1 *Let $\widehat{\mathcal{C}}_{\text{noisy}}$ be constructed as in Recipe 1 with any score function s and the corruption function $g^{\text{confusion}}$. Then,*

$$
\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}}) \right) \geq 1 - \alpha.
$$

if and only if for all classes $j \in \{1, \ldots, K\}$,

$$\sum_{j'=1}^{K} \mathbb{P}(\tilde{Y} = j' \mid Y = j)\mathbb{P}(\tilde{s} \leq t \mid \tilde{Y} = j') \geq \mathbb{P}(s \leq t \mid Y = j).$$

The proof is below.

Proof By law of total probability, $\mathbb{P}(s \leq t) = \mathbb{E}\left[\mathbb{P}(s \leq t \mid Y = j)\right] = \sum_{j=1}^{K} w_j \mathbb{P}(s \leq t \mid Y = j)$. But under the noisy model, we have instead that

$$\mathbb{P}(\tilde{s} \leq t) = \mathbb{E}\left[\mathbb{P}(\tilde{s} \leq t \mid \tilde{Y} = j')\right] = \sum_{j=1}^{K}\sum_{j'=1}^{K} w_j \mathbb{P}(\tilde{Y} = j' \mid Y = j)\mathbb{P}(\tilde{s} \leq t \mid \tilde{Y} = j').$$

We can write

$$\mathbb{P}(s \leq t) - \mathbb{P}(\tilde{s} \leq t) = \sum_{j=1}^{K}\sum_{j'=1}^{K} w_j \mathbb{P}(\tilde{Y} = j' \mid Y = j)\mathbb{P}(\tilde{s} \leq t \mid \tilde{Y} = j') - \sum_{j=1}^{K} w_j \mathbb{P}(s \leq t \mid Y = j).$$

Combining the sums and factoring, the above display equals

$$\sum_{j=1}^{K} w_j \left(\sum_{j'=1}^{K}\left(\mathbb{P}(\tilde{Y} = j' \mid Y = j)\mathbb{P}(\tilde{s} \leq t \mid \tilde{Y} = j')\right) - \mathbb{P}(s \leq t \mid Y = j)\right).$$

We can factor this expression as

$$\sum_{j=1}^{K} w_j \mathbb{P}(s \leq t \mid Y = j) \sum_{j'=1}^{K}\left(\mathbb{P}(\tilde{Y} = j' \mid Y = j)\frac{\mathbb{P}(\tilde{s} \leq t \mid \tilde{Y} = j')}{\mathbb{P}(s \leq t \mid Y = j)} - \frac{1}{K}\right).$$

The stochastic dominance condition holds uniformly over all choices of base probabilities w_j if and only if for all $j \in [K]$,

$$\sum_{j'=1}^{K} \mathbb{P}(\tilde{Y} = j' \mid Y = j)\mathbb{P}(\tilde{s} \leq t \mid \tilde{Y} = j') \geq \mathbb{P}(s \leq t \mid Y = j).$$

∎

Notice that the left-hand side of the above display is a convex mixture of the quantiles $\mathbb{P}(\tilde{s} \leq t \mid \tilde{Y} = j')$ for $j' \in [K]$. Thus, the necessary and sufficient condition is for the noise distribution $\mathbb{P}(\tilde{Y} = j' \mid Y = j)$ to place sufficient mass on the classes j' whose quantiles are larger than $\mathbb{P}(s \leq t \mid Y = j)$. But of course, without assumptions on the model and score, the latter is unknown, so it is impossible to say which noise distributions will preserve coverage.

A.3.3 Prediction Set Size Analysis

When using the non-random APS scores with the oracle model, the prediction set size for a given x can be expressed as

$$k^* = \left\{ \min k : \sum_{j=1}^{k} \pi_{(j)}(x) \geq 1 - \alpha \right\},$$

where $\pi_y(x) = \mathbb{P}(Y \mid X)$, the desired coverage level is $1 - \alpha$, and $\pi_{(1)}(x) \geq \pi_{(2)}(x) \geq \cdots \geq \pi_{(K)}(x)$ are the order statistics of $\pi_1(x), \pi_2(x), \ldots, \pi_K(x)$. Under the random flip noise, the noisy conditional class probabilities are given by

$$\tilde{\pi}_y(x) = (1 - \epsilon)\pi_y(x) + \frac{\epsilon}{K},$$

where K is the number of labels and ϵ is the fraction of flipped labels. Therefore, the noisy set size is given by:

$$k^{*,\text{noisy}} = \left\{ \min k : \sum_{j=1}^{k} \tilde{\pi}_{(j)}(x) \geq 1 - \alpha \right\}$$

$$= \left\{ \min k : \sum_{j=1}^{k} \pi_{(j)}(x) + \epsilon \left(\frac{k}{K} - \sum_{j=1}^{k} \pi_{(j)}(x) \right) \geq 1 - \alpha \right\} \geq k^*.$$

As a result, we get $k^{*,\text{noisy}} \geq k^*$, where the term that controls the inflation of the noisy set size $\epsilon \left(\frac{k}{K} - \sum_{j=1}^{k} \pi_{(j)}(x) \right)$ is non-positive, and is a function of the noise level and the oracle conditional class probabilities.

A.4 Distribution-Free Results

We begin by proving Proposition 3.

Proof [Proof of Proposition 3] For convenience, assume the existence of probability density functions $\tilde{p}$ and p for $\tilde{Y}$ and Y respectively (these can be taken to be probability mass functions if $\mathcal{Y}$ is discrete). Also define the multiset of Y values $E = \{Y_1, ..., Y_n\}$ and the corresponding multiset of $\tilde{Y}$ values $\tilde{E}$. Take the set

$$\mathcal{A} = \{y : \tilde{p}(y) > p(y)\}.$$

Since $Y \overset{d}{\neq} \tilde{Y}$, we know that the set $\mathcal{A}$ is nonempty and $\mathbb{P}(\tilde{Y} \in \mathcal{A}) = \delta_1 > \mathbb{P}(Y \in \mathcal{A}) = \delta_2 \geq 0$. The adversarial choice of score function will be $s(x, y) = \mathbb{1}\{y \in \mathcal{A}^c\}$; it puts high mass wherever the ground truth label is more likely than the noisy label. The crux of the argument is that this design makes the quantile smaller when it is computed on the noisy data than when it is computed on clean data, as we next show.

Begin by noticing that, because $s(x, y)$ is binary, $\hat{q}_{\text{clean}}$ is also binary, and therefore $\hat{q}_{\text{clean}} > t \iff \hat{q}_{\text{clean}} = 1$. Furthermore, $\hat{q}_{\text{clean}} = 1$ if and only if $|E \cap \mathcal{A}|| < \lceil (n+1)(1-\alpha) \rceil$. Thus, these events are the same, and for any $t \in (0, 1]$,

$$\mathbb{P}\left(\hat{q}_{\text{clean}} \geq t\right) = \mathbb{P}\left(|E \cap \mathcal{A}| < \lceil (n+1)(1-\alpha) \rceil\right).$$

By the definition of $\mathcal{A}$, we have that $\mathbb{P}\left(|E \cap \mathcal{A}| < \lceil (n+1)(1-\alpha) \rceil\right) > \mathbb{P}\left(|\tilde{E} \cap \mathcal{A}| < \lceil (n+1)(1-\alpha) \rceil\right)$. Chaining the inequalities, we get

$$\mathbb{P}\left(\hat{q}_{\text{clean}} \geq t\right) > \mathbb{P}\left(|\tilde{E} \cap \mathcal{A}| < \lceil (n+1)(1-\alpha) \rceil\right) = \mathbb{P}\left(\hat{q} \geq t\right).$$

Since s_{n+1} is measurable with respect to E and $\tilde{E}$, we can plug it in for t, yielding the conclusion. $\blacksquare$

Remark A.3 *In the above argument, if one further assumes continuity of the (ground truth) score function and $\mathbb{P}(\tilde{Y} \in \mathcal{A}) = \mathbb{P}(Y \in \mathcal{A}) + \rho$ for*

$$\rho = \inf\left\{\rho' > 0 \ : \text{BinomCDF}(n, \delta_1, \lceil (n+1)(1-\alpha) \rceil - 1) + \frac{1}{n} < \right.$$
$$\left. \text{BinomCDF}(n, \delta_2 + \rho', \lceil (n+1)(1-\alpha) \rceil - 1)\right\},$$

then

$$\mathbb{P}(s_{n+1} \leq \hat{q}) < 1 - \alpha.$$

In other words, the noise must have some sufficient magnitude in order to disrupt coverage.

Next, we provide an additional impossibility result, similar to Proposition 3.

Proposition A.2 *Given a score $s(x, y) = |\hat{f}(x) - y|$ and $\tilde{Y} = \beta Y$, for $\beta \in (0, 1)$. Assume further that $\hat{f}(x) = \mathbb{E}[\tilde{Y} \mid X = x]$, then there exists $\alpha_0 \in [0, 1]$ such that for any α satisfying $1 - \alpha \geq 1 - \alpha_0 : \mathbb{P}(Y \in \hat{\mathcal{C}}_{\text{noisy}}(X)) \leq 1 - \alpha$.*

The proof is provided below:

Proof Denote $\mu_x := \mathbb{E}[Y \mid X = x]$. Then

$$\hat{f}(x) = \mathbb{E}[\tilde{Y} \mid X = x] = \mathbb{E}[\beta Y \mid X = x] = \beta \mathbb{E}[Y \mid X = x] = \beta \mu_x.$$

$$\tilde{s} = |\hat{f}(x) - \tilde{Y}| = |\beta \mu_x - \beta Y| = \beta |\mu_x - Y|$$

$$s = |\hat{f}(x) - Y| = |\beta \mu_x - Y|$$

We will show that $F_{\tilde{s}}(t) \geq F_s(t)$, $\forall t \in \mathbb{R}$. Notice that this is true conditional on X (we do not write this explicitly to enhance clarity.)

Denote $G = \mu_x - Y, H = |G|, J = \beta\mu_x - Y$ and then $\tilde{s} = \beta H, s = |J|$.

$$F_G(g) = \mathbb{P}(G \leq g) = \mathbb{P}(\mu_x - Y \leq g) = 1 - \mathbb{P}(Y \leq \mu_x - g) = 1 - F_Y(\mu_x - g).$$

$$F_H(h) = \mathbb{P}(H \leq h) = \mathbb{P}(|G| \leq h) = \mathbb{P}(-h \leq G \leq h) = F_G(h) - F_G(-h).$$

$$F_J(j) = \mathbb{P}(J \leq j) = \mathbb{P}(\beta\mu_x - Y \leq j) = 1 - \mathbb{P}(Y \leq \beta\mu_x - j) = 1 - F_Y(\beta\mu_x - j).$$

$$F_{\tilde{s}}(t) = F_H(\frac{t}{\beta}) = F_G(\frac{t}{\beta}) - F_G(-\frac{t}{\beta}) = F_Y(\mu_x + \frac{t}{\beta}) - F_Y(\mu_x - \frac{t}{\beta})$$

$$F_s(t) = F_J(t) - F_J(-t) = F_Y(\beta\mu_x + t) - F_Y(\beta\mu_x - t)$$

Let $t \geq 0$, then:

$$F_{\tilde{s}}(t) - F_s(t) = F_Y(\mu_x + \frac{t}{\beta}) - F_Y(\mu_x - \frac{t}{\beta}) - (F_Y(\beta\mu_x + t) - F_Y(\beta\mu_x - t))$$

$$= F_Y(\mu_x + \frac{t}{\beta}) - F_Y(\mu_x - \frac{t}{\beta}) - \left(F_Y(\beta(\mu_x + \frac{t}{\beta})) - F_Y(\beta(\mu_x - \frac{t}{\beta}))\right)$$

For $t \geq |\beta\mu_x|$ from the monotonicity of F_Y we get: $F_Y(\mu_x + \frac{t}{\beta}) \geq F_Y(\beta(\mu_x + \frac{t}{\beta}))$ and $F_Y(\beta(\mu_x - \frac{t}{\beta})) \geq F_Y(\mu_x - \frac{t}{\beta})$. Therefore: $F_{\tilde{s}}(t) \geq F_s(t)$ or equivalently $\mathbb{P}(\tilde{s} \leq t) \geq \mathbb{P}(s \leq t)$. Denote $\alpha_0(x) := 1 - \mathbb{P}(\tilde{s} \leq |\beta\mu_x|)$. Suppose that α satisfies $1 - \alpha \geq 1 - \alpha_0(x)$. Denote by $\hat{q}_{\mathrm{noisy}}(x)$ the $1 - \alpha$ quantile of the noisy scores $\tilde{s}$, conditional on $X = x$. Then, we get that $\hat{q}_{\mathrm{noisy}}(x) \geq |\beta\mu_x|$, and thus:

$$1 - \alpha = \mathbb{P}(\tilde{s} \leq \hat{q}_{\mathrm{noisy}}(x)) \geq \mathbb{P}(s \leq \hat{q}_{\mathrm{noisy}}(x)).$$

Lastly, we return to consider the notations of conditioning on X. Denote $\alpha_0 = \inf_{x \in \mathcal{X}}(\alpha_0(x))$. Suppose that $\alpha \in [0, 1]$ satisfies $1 - \alpha \geq 1 - \alpha_0$. Denote by $\hat{q}_{\mathrm{noisy}}$ the $1 - \alpha$ quantile of the noisy scores $\tilde{s}$, marginally on $x \in \mathcal{X}$. Then, we obtain:

$$1 - \alpha = \mathbb{P}(\tilde{Y} \in \mathcal{C}_{\mathrm{noisy}}(X)) = \mathbb{P}(\tilde{s} \leq \hat{q}_{\mathrm{noisy}}) \geq \mathbb{P}(s \leq \hat{q}_{\mathrm{noisy}}) = \mathbb{P}(Y \in \mathcal{C}_{\mathrm{noisy}}(X)),$$

where above, the probability is taken marginally over $x \in \mathcal{X}$.

■

We commence by proving Corollary 3.

Proof [Proof of Corollary 3] This a consequence of the TV bound from Barber et al. (2023) with weights identically equal to 1. ■

Unfortunately, getting such a TV bound requires a case by case analysis. It's not even straightforward to get a TV bound under strong Gaussian assumptions.

Proposition A.3 (No general TV bound) *Assume* $Y \sim \mathcal{N}(0, \tau^2)$ *and* $\tilde{Y} = y + Z$, *where* $Z \sim \mathcal{N}(0, \sigma^2)$. *Then* $D_{\mathrm{TV}}(Y, \tilde{Y}) \overset{\tau \to 0}{\to} 1$.

Proof

$$\text{TV}(\mathcal{N}(0, \tau^2), \mathcal{N}(0, \tau^2 + \sigma^2)) = \int_{-\infty}^{\infty} \left| e^{-x^2/\tau^2} - e^{-x^2/(\tau^2+\sigma^2)} \right| dx \overset{\tau \to 0}{\to} 1.$$

■

A.5 False-Negative Risk

In this section, we prove all FNR robustness propositions. Here, we suppose that the response $Y \in [0, 1]^n$ is a binary vector of size $n \in \mathbb{N}$, where $Y_i = 1$ indicates that the i-th label is present. We further suppose a vector-flip noise model from (8), where ε is a binary random vector of size n as well. These notations apply for segmentation tasks as well, by flattening the response matrix into a vector. The prediction set $C(X) \subseteq \{1, ..., n\}$ contains a subset of the labels. We begin by providing additional theoretical results and then turn to the proofs.

A.5.1 ADDITIONAL THEORETICAL RESULTS

Proposition A.4 *Let $\widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})$ be a prediction set that controls the FNR risk of the noisy labels at level α. Suppose that*

1. *The prediction set contains the most likely labels in the sense that for all $x \in \mathcal{X}$, $k \in \widehat{\mathcal{C}}_{\text{noisy}}(x)$, $i \notin \widehat{\mathcal{C}}_{\text{noisy}}(x)$ and $m \in \mathbb{N}$:*

$$\mathbb{P}\left(Y_k = 1 \middle| \sum_{j \neq i,k} Y_j = m, X = x\right) \geq \mathbb{P}\left(Y_i = 1 \middle| \sum_{j \neq i,k} Y_j = m, X = x\right).$$

2. *For a given input $X = x$, the noise level of all response elements is the same, i.e., for all $i \neq j \in \{1 \ldots K\}$: $\mathbb{P}(\epsilon_i = 1 \mid X = x) = \mathbb{P}(\epsilon_j = 1 \mid X = x)$.*

3. *The noise is independent of $Y \mid X = x$ in the sense that $\epsilon \perp\!\!\!\perp Y \mid X = x$ for all $x \in \mathcal{X}$.*

4. *The noises of different labels are independent of each other given X, i.e., $\epsilon_i \perp\!\!\!\perp \epsilon_j \mid X = x$ for all $i \neq j \in \{1, \ldots K\}$ and $x \in \mathcal{X}$.*

Then

$$\mathbb{E}\left[L^{\text{FNR}}\left(Y_{\text{test}}, \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right)\right] \leq \alpha.$$

A.5.2 GENERAL DERIVATIONS

In this section, we formulate and prove a general lemma that is used to prove label-noise robustness with the false-negative rate loss.

Lemma A.4 *Suppose that $x \in \mathcal{X}$ is an input variable and $C(x)$ is a prediction set. Denote by β_i the noise level at the i-th element: $\beta_i := \mathbb{P}(\varepsilon_i = 1 \mid X = x)$. We define:*

$$e_{k,i}(y) := \beta_i y_k \mathbb{E}\left[\frac{1}{\sum_{j=1}^n \tilde{Y}_j} \middle| \varepsilon_i = 1, Y = y \right].$$

If for all $k \in C(x)$ and $i \notin C(x)$:

$$\mathbb{E}\left[\frac{e_{k,i}(Y) - e_{i,k}(Y)}{\sum_{j=1}^n Y_j} \middle| X = x \right] \geq 0$$

Then $C(x)$ achieves conservative conditional risk:

$$\mathbb{E}\left[L^{FNR}(Y, C(X)) \mid X = x \right] \leq \mathbb{E}\left[L^{FNR}(\tilde{Y}, C(X)) \mid X = x \right].$$

Furthermore, $C(X)$ achieves valid marginal risk:

$$\mathbb{E}\left[L^{FNR}(Y, C(X)) \right] \leq \mathbb{E}\left[L^{FNR}(\tilde{Y}, C(X)) \right].$$

Proof For ease of notation, we omit the conditioning on $X = x$. That, we take some $x \in \mathcal{X}$ and treat Y as $Y \mid X = x$. We also denote the prediction set as $C = C(x)$. Denote:

$$\delta := L^{\mathrm{FNR}}(\tilde{Y}, C) - L^{\mathrm{FNR}}(Y, C).$$

Our goal is to show that:
$$\mathbb{E}\left[\delta \right] \geq 0.$$

If $C = \emptyset$ or $C = \{1, ..., n\}$ then the proposition is trivially satisfied. Therefore, for the rest of this proof, we assume that $C \neq \emptyset, \{1, ..., n\}$. We begin by developing δ as follows.

$$
\begin{aligned}
\delta &= L^{\mathrm{FNR}}(\tilde{Y}, C) - L^{\mathrm{FNR}}(Y, C) \\
&= \frac{\sum_{j \in C} Y_j}{\sum_j Y_j} - \frac{\sum_{j \in C}[Y_j - 2\varepsilon_j Y_j + \varepsilon_j]}{\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j} \\
&= \frac{\sum_i \sum_{j \in C} Y_j[Y_i - 2\varepsilon_i Y_i + \varepsilon_i] - \sum_i \sum_{j \in C} Y_i[Y_j - 2\varepsilon_j Y_j + \varepsilon_j]}{(\sum_j Y_j)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)} \\
&= \frac{\sum_i \sum_{j \in C} Y_j Y_i - 2Y_j \varepsilon_i Y_i + Y_j \varepsilon_i - \sum_i \sum_{j \in C} Y_i Y_j - 2Y_i \varepsilon_j Y_j + Y_i \varepsilon_j}{(\sum_j Y_j)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)} \\
&= \frac{\sum_i \sum_{j \in C} Y_j Y_i - 2Y_j \varepsilon_i Y_i + Y_j \varepsilon_i - [Y_i Y_j - 2Y_i \varepsilon_j Y_j + Y_i \varepsilon_j]}{(\sum_j Y_j)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)} \\
&= \frac{\sum_i \sum_{j \in C} Y_j \varepsilon_i - Y_i \varepsilon_j}{(\sum_j Y_j)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)} \\
&= \frac{(\sum_i \varepsilon_i)(\sum_{j \in C} Y_j) - (\sum_i Y_i)(\sum_{j \in C} \varepsilon_j)}{(\sum_j Y_j)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)}.
\end{aligned}
$$

Without loss of generality, we assume that $C = \{1, ..., p\}$. We now compute the expectation of each term separately.

$$
\mathbb{E}\left[\frac{(\sum_i \varepsilon_i)(\sum_{k=1}^p Y_k)}{(\sum_j Y_j)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)}\Bigg| Y = y\right]
$$

$$
= \sum_i \mathbb{E}\left[\frac{\varepsilon_i(\sum_{k=1}^p Y_k)}{(\sum_j Y_j)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)}\Bigg| Y = y\right]
$$

$$
= \sum_i \frac{(\sum_{k=1}^p y_k)}{\sum_j y_j}\mathbb{E}\left[\frac{\varepsilon_i}{\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Bigg| Y = y\right]
$$

$$
= \sum_i \frac{(\sum_{k=1}^p y_k)}{\sum_j y_j}\mathbb{E}\left[\frac{\varepsilon_i}{\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Bigg| \varepsilon_i = 1, Y = y\right]\mathbb{P}(\varepsilon_i = 1)
$$

$$
= \frac{1}{\sum_j y_j}\sum_i \beta_i\left(\sum_{k=1}^p y_k\right)\mathbb{E}\left[\frac{1}{(1 - Y_i) + \sum_{j\neq i} Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Bigg| \varepsilon_i = 1, Y = y\right]
$$

$$
= \frac{1}{\sum_j y_j}\sum_i \sum_{k=1}^p \beta_i y_k \mathbb{E}\left[\frac{1}{(1 - Y_i) + \sum_{j\neq i} Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Bigg| \varepsilon_i = 1, Y = y\right]
$$

$$
\mathbb{E}\left[\frac{(\sum_i Y_i)(\sum_{k=1}^p \varepsilon_k)}{(\sum_j Y_j)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)}\Bigg| Y = y\right]
$$

$$
= \sum_i \mathbb{E}\left[\frac{Y_i(\sum_{k=1}^p \varepsilon_k)}{(\sum_j Y_j)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)}\Bigg| Y = y\right]
$$

$$
= \sum_i \frac{y_i}{\sum_j y_j}\mathbb{E}\left[\frac{(\sum_{k=1}^p \varepsilon_k)}{\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Bigg| Y = y\right]
$$

$$
= \sum_i \frac{y_i}{\sum_j y_j}\sum_{k=1}^p \mathbb{E}\left[\frac{\varepsilon_k}{\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Bigg| Y = y\right]
$$

$$
= \sum_i \frac{y_i}{\sum_j y_j}\sum_{k=1}^p \beta_k\mathbb{E}\left[\frac{\varepsilon_k}{\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Bigg| \varepsilon_k = 1, Y = y\right]
$$

$$
= \sum_i \frac{y_i}{\sum_j y_j}\sum_{k=1}^p \beta_k\mathbb{E}\left[\frac{1}{(1 - Y_k) + \sum_{j\neq k} Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Bigg| \varepsilon_k = 1, Y = y\right]
$$

$$
= \frac{1}{\sum_j y_j}\sum_i \sum_{k=1}^p \beta_k y_i\mathbb{E}\left[\frac{1}{(1 - Y_k) + \sum_{j\neq k} Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Bigg| \varepsilon_k = 1, Y = y\right]
$$

We now compute the expected value of δ for a given vector of labels $Y = y$.

$$
\begin{aligned}
\mathbb{E}\left[\delta | Y = y\right] &= \mathbb{E}\left[\left.\frac{(\sum_i \varepsilon_i)(\sum_{k=1}^p Y_k)}{(\sum_i Y_i)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)}\right| Y = y\right] - \mathbb{E}\left[\left.\frac{(\sum_i Y_i)(\sum_{k=1}^p \varepsilon_k)}{(\sum_i Y_i)(\sum_j Y_j - 2\varepsilon_j Y_j + \varepsilon_j)}\right| Y = y\right] \\
&= \frac{1}{\sum_j y_j}\left(\sum_i \sum_{k=1}^p e_{k,i}(y) - e_{i,k}(y)\right) \\
&= \frac{1}{\sum_j y_j}\left[\left(\sum_{i=1}^p \sum_{k=1}^p e_{k,i}(y) + \sum_{i=p+1}^n \sum_{k=1}^p e_{k,i}(y)\right) - \left(\sum_{i=1}^p \sum_{k=1}^p e_{i,k}(y) + \sum_{i=p+1}^n \sum_{k=1}^p e_{i,k}(y)\right)\right] \\
&= \frac{1}{\sum_j y_j}\left[\sum_{i=p+1}^n \sum_{k=1}^p e_{k,i}(y) - \sum_{i=p+1}^n \sum_{k=1}^p e_{i,k}(y)\right] \\
&= \frac{1}{\sum_j y_j}\left[\sum_{i=p+1}^n \sum_{k=1}^p e_{k,i}(y) - e_{i,k}(y)\right] \\
&= \sum_{i=p+1}^n \sum_{k=1}^p \frac{e_{k,i}(y) - e_{i,k}(y)}{\sum_j y_j}
\end{aligned}
$$

Finally, we get valid risk conditioned on $X = x$:

$$
\mathbb{E}\left[\delta\right] = \mathbb{E}\left[\sum_{i=p+1}^n \sum_{k=1}^p \frac{e_{k,i}(Y) - e_{i,k}(Y)}{\sum_j Y_j}\right] = \sum_{i=p+1}^n \sum_{k=1}^p \mathbb{E}\left[\frac{e_{k,i}(Y) - e_{i,k}(Y)}{\sum_j Y_j}\right] \geq 0.
$$

Above, the last inequality follows from the assumption of this lemma. We now marginalize the above result to obtain valid marginal risk:

$$
\mathbb{E}\left[\delta\right] = \mathbb{E}_X\left[\mathbb{E}_{\tilde{Y}, Y | X = x}\left[\delta \mid X = x\right]\right] \geq \mathbb{E}_X\left[0\right] \geq 0.
$$

$\blacksquare$

A.5.3 PROOF OF PROPOSITION A.4

Proof As in Lemma A.4, we omit the conditioning on X and treat Y as $Y \mid X = x$. Without loss of generality, we suppose that $C = \{1, ..., p\}$, where $\mathbb{P}(\tilde{Y}_i = 1) \geq \mathbb{P}(\tilde{Y}_{i+1} = 1)$ for all $i \in \{1, ..., n-1\}$. Since $\forall i \in \{1, ..., n-1\} : \mathbb{P}(\varepsilon_i = 1) \leq 0.5$, the order of class probabilities is preserved under the corruption, meaning that

$$
\forall i \in \{1, ..., n-1\} : \mathbb{P}(Y_i = 1) \geq \mathbb{P}(Y_{i+1} = 1).
$$

We now compute $\mathbb{E}\left[\frac{e_{k,i}(Y)-e_{i,k}(Y)}{\sum_j Y_j}\right]$ for $k < i$:

$$\mathbb{E}\left[\frac{e_{k,i}(Y) - e_{i,k}(Y)}{\sum_i Y_i}\right]$$

$$= \mathbb{E}\left[\frac{\beta_i Y_k \mathbb{E}\left[\frac{1}{(1-Y_i)+\sum_{j\neq i} Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\Big| Y = Y\right] - \beta_k Y_i \mathbb{E}\left[\frac{1}{(1-Y_k)+\sum_{j\neq k} Y_j - 2\varepsilon_j Y_j + \varepsilon_j}\mid Y = Y\right]}{\sum_j Y_j}\right]$$

$$= \sum_{y\in\mathcal{Y}} \mathbb{P}(Y = y)\left[\frac{\beta_i y_k}{\sum_j y_j}\mathbb{E}\left[\frac{1}{(1-y_i)+\sum_{j\neq i} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right]\right.$$

$$\left. - \frac{\beta_k y_i}{\sum_j y_j}\mathbb{E}\left[\frac{1}{(1-y_k)+\sum_{j\neq k} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right]\right]$$

$$= \sum_{\{y\in\mathcal{Y}:y_i\neq y_k\}} \mathbb{P}(Y = y)\left[\frac{\beta_i y_k}{\sum_j y_j}\mathbb{E}\left[\frac{1}{(1-y_i)+\sum_{j\neq i} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right]\right]$$

$$- \sum_{\{y\in\mathcal{Y}:y_i\neq y_k\}} \mathbb{P}(Y = y)\left[\frac{\beta_k y_i}{\sum_j y_j}\mathbb{E}\left[\frac{1}{(1-y_k)+\sum_{j\neq k} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right]\right]$$

$$= \sum_{\{y\in\mathcal{Y}:y_k=1,y_i=0\}} \mathbb{P}(Y = y)\frac{\beta_i}{\sum_j y_j}\mathbb{E}\left[\frac{1}{2-\varepsilon_k+\sum_{j\neq i,k} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right]$$

$$- \sum_{\{y\in\mathcal{Y}:y_i=1,y_k=0\}} \mathbb{P}(Y = y)\frac{\beta_k}{\sum_j y_j}\mathbb{E}\left[\frac{1}{2-\varepsilon_i+\sum_{j\neq k,i} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right]$$

To simplify the equation, denote by y^* the vector y with indexes i, k swapped, that is:

$$y_j^* = \begin{cases} y_j & j \neq i, k, \\ y_i & j = k, \\ y_k & j = i. \end{cases}$$

Notice that $\beta_k \mathbb{E}\left[\frac{1}{2-\varepsilon_i+\sum_{j\neq k,i} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right] = \beta_i \mathbb{E}\left[\frac{1}{2-\varepsilon_k+\sum_{j\neq k,i} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right]$ since $\varepsilon_i \overset{d}{=} \varepsilon_k$ and these variables are independent of Y and of ε_j for $j \neq i, k$. Further denote $\gamma_m = \beta\left(\sum_j y_j\right)\mathbb{E}\left[\frac{1}{2-\varepsilon_k+\sum_{j\neq i,k} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right]$ for $\sum_{j\neq i,k} y_j = m$ and $y_i \neq y_k$. Notice that γ_m is well defined as $\beta\mathbb{E}\left[\frac{1}{2-\varepsilon_k+\sum_{j\neq i,k} y_j - 2\varepsilon_j y_j + \varepsilon_j}\Big| Y = y\right]$ has the same value for every y such that $\sum_{j\neq i,k} y_j = m$ since y and ε are independent. Also $\sum_j y_j = m+1$ if $\sum_{j\neq i,k} y_j = m$ and $y_i \neq y_k$. Lastly, we denote $\gamma'_m = \gamma_m \mathbb{P}\left(\sum_{j\neq i,k} Y_j = m\right)$. We continue computing

$\mathbb{E}\left[\frac{e_{k,i}(Y)-e_{i,k}(Y)}{\sum_j Y_j}\right]$ for $k < i$:

$$\mathbb{E}\left[\frac{e_{k,i}(Y)-e_{i,k}(Y)}{\sum_j Y_j}\right]$$

$$= \sum_{\{y\in\mathcal{Y}:y_k=1,y_i=0\}} (\mathbb{P}(Y=y)-\mathbb{P}(Y=y^*))\frac{\beta}{\sum_j y_j}\mathbb{E}\left[\frac{1}{2-\varepsilon_k+\sum_{j\neq i,k} y_j - 2\varepsilon_j y_j + \varepsilon_j}\middle| Y=y\right]$$

$$= \sum_m \sum_{\{y\in\mathcal{Y},y_k=1,y_i=0,\sum_{j\neq i,k} y_j=m\}} (\mathbb{P}(Y=y)-\mathbb{P}(Y=y^*))\gamma_m$$

$$= \sum_m \gamma_m \sum_{\{y\in\mathcal{Y},y_k=1,y_i=0,\sum_{j\neq i,k} y_j=m\}} [\mathbb{P}(Y=y)-\mathbb{P}(Y=y^*)]$$

$$= \sum_m \gamma_m \left[\sum_{\{y\in\mathcal{Y},y_k=1,y_i=0,\sum_{j\neq i,k} y_j=m\}} \mathbb{P}(Y=y) - \sum_{\{y\in\mathcal{Y},y_i=1,y_k=0,\sum_{j\neq i,k} y_j=m\}} \mathbb{P}(Y=y)\right]$$

$$= \sum_m \gamma_m \left[\mathbb{P}\left(Y_k=1,Y_i=0,\sum_{j\neq i,k} Y_j=m\right) - \mathbb{P}\left(Y_k=0,Y_i=1,\sum_{j\neq i,k} Y_j=m\right)\right]$$

$$= \sum_m \gamma'_m \left[\mathbb{P}\left(Y_k=1,Y_i=0 \mid \sum_{j\neq i,k} Y_j=m\right) - \mathbb{P}\left(Y_k=0,Y_i=1 \mid \sum_{j\neq i,k} Y_j=m\right)\right]$$

$$= \sum_m \gamma'_m \left[\mathbb{P}\left(Y_k=1,Y_i=0\middle|\sum_{j\neq i,k} Y_j=m\right) + \mathbb{P}\left(Y_k=1,Y_i=1\middle|\sum_{j\neq i,k} Y_j=m\right)\right.$$

$$\left. - \mathbb{P}\left(Y_k=1,Y_i=1\middle|\sum_{j\neq i,k} Y_j=m\right) - \mathbb{P}\left(Y_k=0,Y_i=1\middle|\sum_{j\neq i,k} Y_j=m\right)\right]$$

$$= \sum_m \gamma'_m \left[\mathbb{P}\left(Y_k=1\middle|\sum_{j\neq i,k} Y_j=m\right) - \mathbb{P}\left(Y_i=1\middle|\sum_{j\neq i,k} Y_j=m\right)\right]$$

We assume that $\forall m : \mathbb{P}\left(Y_k=1\middle|\sum_{j\neq i,k} Y_j=m\right) \geq \mathbb{P}\left(Y_i=1\middle|\sum_{j\neq i,k} Y_j=m\right)$ and therefore:

$$\mathbb{E}\left[\frac{e_{k,i}(Y)-e_{i,k}(Y)}{\sum_j Y_j}\right]$$

$$= \sum_m \gamma'_m \left[\mathbb{P}\left(Y_k=1\middle|\sum_{j\neq i,k} Y_j=m\right) - \mathbb{P}\left(Y_i=1\middle|\sum_{j\neq i,k} Y_j=m\right)\right]$$

$$\geq 0.$$

According to Lemma A.4, the above concludes the proof. ∎

A.5.4 MULTI-LABEL CLASSIFICATION

Here we prove Proposition 4.

Proof [Proof of Proposition 4] As in Lemma A.4, we omit the conditioning on X and treat Y as $Y \mid X = x$. Without loss of generality, we suppose that $C = \{1, ..., p\}$, where $\mathbb{P}(\tilde{Y}_i = 1) \geq \mathbb{P}(\tilde{Y}_{i+1} = 1)$ for all $i \in \{1, ..., n-1\}$. Since $\forall i \in \{1, ..., n-1\} : \mathbb{P}(\varepsilon_i = 1) \leq 0.5$, the order of class probabilities is preserved under the corruption, meaning that

$$\forall i \in \{1, ..., n-1\} : \mathbb{P}(Y_i = 1) \geq \mathbb{P}(Y_{i+1} = 1).$$

Since Y is a deterministic function of X, we get that Y is some non-increasing constant binary vector. Suppose that $k \in C$ and $i \notin C$. Since C contains the most likely labels, these indexes satisfy $k < i$. We now compare $e_{k,i}(y)$ to $e_{i,k}$. According to the definition of $e_{k,i}$:

$$e_{k,i} := \beta_i y_k \mathbb{E}\left[\frac{1}{\sum_{j=1}^n \tilde{Y}_j}\middle| \varepsilon_i = 1, Y = y\right].$$

Since $\sum_{j=1}^n \tilde{Y}_j$ is a assumed to be a constant, $e_{k,i}(y)$ and $e_{i,k}$ may differ only in the values of y_k, y_i and β_k, β_i. We go over all four combinations of y_k and y_i.

1. If $y_k = 0, y_i = 0$, then $e_{k,i}(y) = e_{i,k}(y) = 0$.

2. If $y_k = 1, y_i = 0$, then $e_{k,i}(y) \geq 0$ and $e_{i,k}(y) = 0$.

3. If $y_k = 0, y_i = 1$, then we get a contradiction to the property of Y being non-increasing, and thus this case is not possible.

4. If $y_k = 1, y_i = 1$, then $e_{k,i}(y) \geq e_{i,k}(y) = 0$ since $\beta_i \geq \beta_k$:

$$\mathbb{P}(\varepsilon_k = 0) = \mathbb{P}(\tilde{Y}_k = 1) \geq \mathbb{P}(\tilde{Y}_i = 1) = \mathbb{P}(\varepsilon_i = 0)$$
$$\Rightarrow 1 - \mathbb{P}(\varepsilon_k = 1) \geq 1 - \mathbb{P}(\varepsilon_i = 1)$$
$$\Rightarrow \beta_k = \mathbb{P}(\varepsilon_k = 1) \leq \mathbb{P}(\varepsilon_i = 1) = \beta_i.$$

Therefore, under all cases, we get that $e_{k,i}(y) \geq e_{i,k}(y)$ for all $k \in C, i \notin C$, and thus:

$$\mathbb{E}\left[\frac{e_{k,i}(Y) - e_{i,k}(Y)}{\sum_{j=1}^n Y_j}\middle| X = x\right] \geq 0.$$

Finally, by applying Lemma A.4, we obtain valid conditional risk.

∎

A.5.5 SEGMENTATION

Proof [Proof of Proposition 5] This proposition is a special case of Proposition A.4 and thus valid risk follows directly from this result. ∎

A.6 Risk Control for Regression Tasks

A.6.1 General Risk Bound for Regression Tasks

Here we prove Proposition 6.

Proof [Proof of Proposition 6] For ease of notation, we omit the conditioning on $X = x$. In other words, we treat Y as $Y \mid X = x$ for some $x \in \mathcal{X}$. Given a prediction set $\hat{C}(x)$, we consider the loss as a function of Y, where $\hat{C}(x)$ is fixed, and denote it by:

$$L(Y) := L(Y, \hat{C}(x)).$$

We expand $L(Y + \varepsilon)$ using Taylor's expansion:

$$L(Y + \varepsilon) = L(Y) + \varepsilon L'(Y) + \frac{1}{2}\varepsilon^2 L''(\xi),$$

where ξ is some real number between Y and $Y + \varepsilon$.

$$
\begin{aligned}
\delta &:= L(\tilde{Y}) - L(Y) \\
&= L(Y + \varepsilon) - L(Y) \\
&= \varepsilon L'(Y) + \frac{1}{2}\varepsilon^2 L''(\xi).
\end{aligned}
$$

We now develop each term separately. Since $\varepsilon \perp\!\!\!\perp Y$ it follows that $\varepsilon \perp\!\!\!\perp L'(Y)$.

$$\mathbb{E}\left[\varepsilon L'(Y)\right] = \mathbb{E}\left[\varepsilon\right]\mathbb{E}\left[L'(Y)\right] = 0\mathbb{E}\left[L'(Y)\right] = 0$$

Since $L''(y)$ is bounded by $q \leq L''(y) \leq Q$, we get the following:

$$\frac{1}{2}q\mathrm{Var}\left[\varepsilon\right] = \frac{1}{2}q\mathbb{E}\left[\varepsilon^2\right] = \mathbb{E}\left[\frac{1}{2}\varepsilon^2 q_x\right] \leq \mathbb{E}\left[\frac{1}{2}\varepsilon^2 L''(\xi)\right] \leq \mathbb{E}\left[\frac{1}{2}\varepsilon^2 Q_x\right] = \frac{1}{2}Q\mathbb{E}\left[\varepsilon^2\right] = \frac{1}{2}Q\mathrm{Var}\left[\varepsilon\right]$$

We get that:

$$\mathbb{E}[\delta] = \mathbb{E}\left[\varepsilon L'(Y) + \frac{1}{2}\varepsilon^2 L''(\xi)\right] = 0 + \mathbb{E}\left[\frac{1}{2}\varepsilon^2 L''(\xi)\right].$$

Therefore:

$$\frac{1}{2}q\mathrm{Var}\left[\varepsilon\right] \leq \mathbb{E}[\delta] \leq \frac{1}{2}Q\mathrm{Var}\left[\varepsilon\right].$$

$$\Rightarrow \frac{1}{2}q\mathrm{Var}\left[\varepsilon\right] \leq \mathbb{E}\left[L(\tilde{Y}, \hat{C}(X)) - L(Y, \hat{C}(X))\right] \leq \frac{1}{2}Q\mathrm{Var}\left[\varepsilon\right].$$

We now turn to consider the conditioning on $X = x$ and obtain marginalized bounds by taking the expectation over all X:

$$\alpha - \frac{1}{2}Q\mathrm{Var}\left[\varepsilon\right] \leq \mathbb{E}\left[L(Y, \hat{C}(X))\right] \leq \alpha - \frac{1}{2}q\mathrm{Var}\left[\varepsilon\right]$$

where $\alpha := \mathbb{E}\left[L(Y, \hat{C}(X))\right]$, $q := \mathbb{E}_X[q_X]$, and $Q := \mathbb{E}_X[Q_X]$. Additionally, if $L(y, C(x))$ is convex for all $x \in \mathcal{X}$, then $q_x \geq 0$, and we get conservative coverage:

$$\mathbb{E}\left[L(Y, \hat{C}(X))\right] \leq \mathbb{E}\left[L(\tilde{Y}, \hat{C}(X))\right] = \alpha.$$

■

A.6.2 DERIVING A MISCOVERAGE BOUND FROM THE GENERAL RISK BOUND IN REGRESSION TASKS

In this section we prove Proposition 7 and show how to obtain tight coverage bounds. First, we define a parameterized smoothed miscoverage loss:

$$L_{d,c}^{sm}(y, [a, b]) = \frac{2}{1 + e^{-d\left(\left(2\frac{y-a}{b-a}-1\right)^2\right)^c}}.$$

Above, $c, d \in \mathbb{R}$ are parameters that affect the loss function. We first connect between the smooth miscoverage and the standard miscoverage functions:

$$L(y, [a, b]) = \mathbb{1}\{L_{d,c}^{sm}(y, [a, b]) \geq h\}. \tag{13}$$

We now invert the above equation to find h:

$$
\begin{aligned}
h &= L_{d,c}^{sm}\left(\min_{L(y,[a,b])=0} y, [a, b]\right) \\
&= L_{d,c}^{sm}\left(\max_{L(y,[a,b])=0} y, [a, b]\right) \\
&= L_{d,c}^{sm}\left(a, [a, b]\right) \\
&= L_{d,c}^{sm}\left(b, [a, b]\right) \\
&= \frac{2}{1 + e^{-d\left(\left(2\frac{a-a}{b-a}-1\right)^2\right)^c}} \\
&= \frac{2}{1 + e^{-d\left((-1)^2\right)^c}} \\
&= \frac{2}{1 + e^{-d}}.
\end{aligned}
$$

Therefore, $h(d)$ is a function that depends only on d. We now denote the second derivative of the smoothed loss by:

$$q_x(c, d) = \min_y \frac{\partial^2}{\partial y^2} L_{c,d}^{sm}(y, C(x)).$$

Importantly, $q_x(c, d)$ can be empirically computed by sweeping over all $y \in \mathbb{R}$ and computing the second derivative of $L_{c,d}^{sm}$ for each of them.

We obtain an upper bound for the miscoverage of $C(x)$ by applying Markov's inequality using (13):

$$\mathbb{P}(Y \notin C(X) \mid X = x) = \mathbb{P}(L_{d,c}^{sm}(Y, C(X)) \geq h(d) \mid X = x) \leq \frac{\mathbb{E}\left[L_{c,d}^{sm}(Y, C(X)) \mid X = x\right]}{h(d)}. \tag{14}$$

Next, we employ smoothed miscoverage upper bound provided by Proposition 7:

$$\mathbb{E}\left[L_{c,d}^{sm}(Y, C(X)) \mid X = x\right] \leq \mathbb{E}\left[L_{c,d}^{sm}(\tilde{Y}, C(X)) \mid X = x\right] - \frac{1}{2}q_x(c, d)\mathrm{Var}\left[\varepsilon\right] \tag{15}$$

Finally, we combine (14) and (15) and derive the following miscoverage bound:

$$\mathbb{P}(Y \notin C(X) \mid X = x) \leq \frac{\mathbb{E}\left[L_{c,d}^{sm}(\tilde{Y}, C(X)) \mid X = x\right] - \frac{1}{2}q_x(c, d)\mathrm{Var}\left[\varepsilon\right]}{h(d)},$$

which can be restated to:

$$\mathbb{P}(Y \in C(X) \mid X = x) \geq 1 - \frac{\mathbb{E}\left[L_{c,d}^{sm}(\tilde{Y}, C(X)) \mid X = x\right] - \frac{1}{2}q_x(c, d)\mathrm{Var}\left[\varepsilon\right]}{h(d)}.$$

Finally, we take the expectation over all X to obtain marginal coverage bound:

$$\mathbb{P}(Y \in C(X)) \geq 1 - \frac{\mathbb{E}\left[L_{c,d}^{sm}(\tilde{Y}, C(X))\right] - \frac{1}{2}\mathbb{E}_X[q_x(c, d)]\mathrm{Var}\left[\varepsilon\right]}{h(d)}. \tag{16}$$

Crucially, all variables in (16) are empirically computable so the above lower bound can is known in practice. Additionally, the parameters c, d can be tuned over a validation set to obtain tighter bounds.

A.6.3 DERIVING A MISCOVERAGE BOUND FROM THE DENSITY'S SMOOTHNESS

We say that the PDF of $Y \mid X = x$ is K_x Lipschitz if

$$|f_{Y|X=x}(y + u) - f_{Y|X=x}(y)| \leq K_x u,$$

where $f_{Y|X=x}$ is the PDF of $Y \mid X = x$ and $K_x \in \mathbb{R}$ is a constant that depends only on x.

Proposition A.5 *Suppose that $C(x)$ is a prediction interval. Under the additive noise model g^{add} from (3), if the PDF of $Y \mid X = x$ is K_x Lipschitz then:*

$$\mathbb{P}(Y \in C(X)) \geq \mathbb{P}(\tilde{Y} \in C(X)) - \mathbb{E}\left[|C(X)|K_X\right]\mathbb{E}\left[|Z|\right].$$

Proof First, by the definition of the noise model we get that:

$$
\begin{aligned}
f_{\tilde{Y}|X=x}(y) &= \int_{-\infty}^{\infty} f_{Y|X=x}(y-z) f_Z(z)\, dz \\
&\leq \int_{-\infty}^{\infty} (f_{Y|X=x}(y) + K_x|z|) f_Z(z)\, dz \\
&= f_{Y|X=x}(y) \int_{-\infty}^{\infty} f_Z(z)\, dz + K_x \int_{-\infty}^{\infty} |z| f_Z(z)\, dz \\
&= f_{Y|X=x}(y) + K_x \mathbb{E}[|Z|].
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\mathbb{P}(Y \in C(X) \mid X = x) &= \int_{y \in C(x)} f_{Y|X=x}(y)\, dy \\
&\geq \int_{y \in C(x)} (f_{\tilde{Y}|X=x}(y) - K_x \mathbb{E}[|Z|])\, dy \\
&= \mathbb{P}(\tilde{Y} \in C(X) \mid X = x) - |C(x)| K_x \mathbb{E}[|Z|].
\end{aligned}
$$

We marginalize the above to obtain the marginal coverage bound:

$$
\mathbb{P}(Y \in C(X)) = \mathbb{P}(\tilde{Y} \in C(X)) - \mathbb{E}[|C(X)|K_X]\mathbb{E}[|Z|].
$$

■

A.6.4 The Image Miscoverage Loss

In this section, we analyze the setting where the response variable is a matrix $\mathcal{Y} = \mathbb{R}^{W \times H}$. Here, the uncertainty is represented by a prediction interval $C^{i,j}(X)$ for each pixel i,j in the response image. Here, our goal is to control the *image miscoverage* loss (12), defined as:

$$
L^{\mathrm{im}}(Y, C(X)) = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} 1\{Y^{i,j} \notin C^{i,j}(X)\}.
$$

While this loss can be controlled under an i.i.d assumption by applying the methods proposed by Angelopoulos et al. (2021, 2024), these techniques may produce invalid uncertainty sets in the presence of label noise. We now show that conservative image miscoverage risk is obtained under the assumptions of Theorem A.1.

Proposition A.6 *Suppose that each element of the response matrix is corrupted according to an additive noise model g^{add} with a noise that has mean 0. Suppose that for every pixel i,j of the response matrix, the prediction interval $C^{i,j}(X)$ and the conditional distribution of $Y^{i,j} \mid X = x$ satisfy the assumptions of Theorem A.1 for all $x \in \mathcal{X}$. Then, we obtain valid conditional image miscoverage:*

$$
\forall x \in \mathcal{X} : \mathbb{E}[L^{im}(Y, C(X)) \mid X = x] \leq \mathbb{E}[L^{im}(\tilde{Y}, C(X)) \mid X = x].
$$

Proof For ease of notation, we suppose that Y is a random vector of length k and C^i is the prediction interval constructed for the i-th element in Y. Suppose that $x \in \mathcal{X}$. Under the assumptions of Theorem A.1, we get that for all $i \in \{1, .., , k\}$:

$$\mathbb{P}(Y^i \notin C^i(x) \mid X = x) \leq \mathbb{P}(\tilde{Y}^i \notin C^i(x) \mid X = x).$$

Therefore:

$$
\begin{aligned}
\mathbb{E}[L^{\mathrm{im}}(Y, C(X)) \mid X = x] &= \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} 1\{Y^i \notin C^i(X)\} \mid X = x\right] \\
&= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left[1\{Y^i \notin C^i(X)\} \mid X = x\right] \\
&= \frac{1}{k}\sum_{i=1}^{k} \mathbb{P}\left[Y^i \notin C^i(X) \mid X = x\right] \\
&\leq \frac{1}{k}\sum_{i=1}^{k} \mathbb{P}\left[\tilde{Y}^i \notin C^i(X) \mid X = x\right] \\
&= \frac{1}{k}\sum_{i=1}^{k} \mathbb{P}\left[\tilde{Y}^i \notin C^i(X) \mid X = x\right] \\
&= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left[1\{\tilde{Y}^i \notin C^i(X)\} \mid X = x\right] \\
&= \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} 1\{\tilde{Y}^i \notin C^i(X)\} \mid X = x\right] \\
&= \mathbb{E}[L^{\mathrm{im}}(\tilde{Y}, C(X)) \mid X = x].
\end{aligned}
$$

$\blacksquare$

A.7 Online Learning

A.7.1 LABEL-NOISE IN ONLINE LEARNING SETTINGS

Here we provide the proof to Proposition 8.

Proof [Proof of Proposition 8]

Suppose that for every $t \in \mathbb{N}$:

$$\mathbb{E}_{Y_t \mid X_t = x}\left[L_t(Y_t, \hat{C}_t(X_t)) \mid X_t = x\right] \leq \alpha_t.$$

Draw T uniformly from $[0, 1, ..., \infty]$. Then, from the law of total expectation, it follows that:

$$
\begin{aligned}
\mathbb{E}_T\left[L_T(Y_T, \hat{C}_T(X_T))\right] &= \mathbb{E}_T\left[\mathbb{E}_{X_T|T=t}\left[\mathbb{E}_{Y_T|X_T=x}\left[L_T(Y_T, \hat{C}_T(X_T)) \mid X_T = x\right] \mid T = t\right]\right] \\
&\leq \mathbb{E}_T\left[\mathbb{E}_{X_T|T=t}\left[\alpha_T \mid T = t\right]\right] \\
&= \mathbb{E}_T\left[\alpha_T\right] \\
&= \alpha,
\end{aligned}
$$

where $\alpha := \mathbb{E}_T[\alpha_T]$. Finally, we get:

$$
\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} L_t(Y_t.\hat{C}_t(X_t)) = \mathbb{E}_T\left[L_T(Y_T, \hat{C}_T(X_T))\right] \leq \alpha
$$

∎

A.7.2 LABEL-NOISE ROBUSTNESS WITH THE MISCOVERAGE COUNTER LOSS

In this section, we suppose an online learning setting, where the data $\{(x_t, y_t)\}_{t=1}^{\infty}$ is given as a stream. The miscoverage counter loss (Feldman et al., 2023b) counts the number of consecutive miscoverage events that occurred until the timestamp t. Formally, given a series of prediction sets: $\{\hat{C}_t(x_t)\}_{t=1}^{T}$, and a series of labels: $\{y_t\}_{t=1}^{T}$, the miscoverage counter at timestamp t is defined as:

$$
L_t^{\mathrm{MC}}(y_t, \hat{C}_t(x_t)) = \begin{cases} L_{t-1}^{\mathrm{MC}}(y_{t-1}, \hat{C}_{t-1}(x_{t-1})) + 1, & y_t \notin \hat{C}_t(x_t) \\ 0, & \text{otherwise.} \end{cases}
$$

We now show that conservative miscoverage counter risk is obtained under the presence of label noise.

Proposition A.7 *Suppose that for all $0 < k \in \mathbb{N}$, $1 < t \in \mathbb{N}$, and $x_t \in \mathcal{X}$:*

$$
\begin{aligned}
\mathbb{P}(Y_t \in \widehat{\mathcal{C}}_{\mathrm{noisy}}^t(X_t) \mid X_t = x_t, L_{t-1}^{MC}(Y_{t-1}, \hat{C}_{t-1}(X_{t-1})) = k) &\geq \\
\mathbb{P}(\tilde{Y}_t \in \widehat{\mathcal{C}}_{\mathrm{noisy}}^t(X_t) \mid X_t = x_t, L_{t-1}^{MC}(\tilde{Y}_{t-1}, \hat{C}_{t-1}(X_{t-1})) = k).
\end{aligned}
$$

Also assume that:

$$
\mathbb{P}(Y_1 \in \widehat{\mathcal{C}}_{\mathrm{noisy}}^1(X_1) \mid X_1 = x_1) \geq \mathbb{P}(\tilde{Y}_1 \in \widehat{\mathcal{C}}_{\mathrm{noisy}}^1(X_1) \mid X_1 = x_1).
$$

If the miscoverage counter risk of the noisy labels is controlled at level α, then the miscoverage counter of the clean labels is controlled at level α:

$$
\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[L_t^{MC}(Y_t, \hat{C}_t(X_t))] \leq \alpha.
$$

Notice that the conditions of Proposition A.7 follow from Theorem A.2 in classification tasks or from Theorem A.1 in regression tasks. In other words, we are guaranteed to obtain valid miscoverage counter risk when the requirements of these theorems are satisfied. We now demonstrate this result in a regression setting.

Corollary A.8 *Suppose that the noise model and the conditional distributions of the clean $Y_t \mid X_t, X_{t-1}, Y_{t-1}$ and noisy $\tilde{Y}_t \mid X_t, X_{t-1}, \tilde{Y}_{t-1}$ labels satisfy the assumptions of Theorem A.1 for all $t \in \mathbb{N}$. Then, the miscoverage counter of the clean labels is more conservative than the risk of the noisy labels.*

We now turn to prove Proposition A.7.

Proof [Proof of Proposition A.7] Suppose that $x_t \in \mathcal{X}$. Our objective is to prove that any series of intervals $\{\hat{C}_t(X_t)\}_{t=1}^{\infty}$ satisfies:

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t))] \leq \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[L_t^{\mathrm{MC}}(\tilde{Y}_t, \hat{C}_t(X_t))]$$

First, we show by induction over $0 < t, k \in \mathbb{N}$ that:

$$\mathbb{P}[L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t)) = k] \leq \mathbb{P}[L_t^{\mathrm{MC}}(\tilde{Y}_t, \hat{C}_t(X_t)) = k].$$

Base: for $k = 1$ and $t = 1$:

$$\begin{aligned}
\mathbb{P}[L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t)) = k \mid X_t = x_t] &= \mathbb{P}[Y_t \notin, \hat{C}_t(X_t) \mid X_t = x_t] \\
&\leq \mathbb{P}[\tilde{Y}_t \notin, \hat{C}_t(X_t) \mid X_t = x_t] \\
&= \mathbb{P}[L_t^{\mathrm{MC}}(\tilde{Y}_t, \hat{C}_t(X_t)) = k \mid X_t = x_t].
\end{aligned}$$

Inductive step: suppose that the statement is correct for t, k. We now show for $k + 1$:

$$\begin{aligned}
&\mathbb{P}[L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t)) = k + 1] \\
&= \mathbb{P}[L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t)) = k + 1 \mid L_{t-1}^{\mathrm{MC}}(Y_{t-1}, \hat{C}_{t-1}(X_{t-1})) = k]\mathbb{P}[L_{t-1}^{\mathrm{MC}}(Y_{t-1}, \hat{C}_{t-1}(X_{t-1})) = k] \\
&= \mathbb{P}[Y_t \notin \hat{C}_t(X_t) \mid L_{t-1}^{\mathrm{MC}}(Y_{t-1}, \hat{C}_{t-1}(X_{t-1})) = k]\mathbb{P}[L_{t-1}^{\mathrm{MC}}(Y_{t-1}, \hat{C}_{t-1}(X_{t-1})) = k] \\
&\leq \mathbb{P}[\tilde{Y}_t \notin \hat{C}_t(X_t) \mid L_{t-1}^{\mathrm{MC}}(\tilde{Y}_{t-1}, \hat{C}_{t-1}(X_{t-1})) = k]\mathbb{P}[L_{t-1}^{\mathrm{MC}}(\tilde{Y}_{t-1}, \hat{C}_{t-1}(X_{t-1})) = k] \\
&= \mathbb{P}[L_t^{\mathrm{MC}}(\tilde{Y}_t, \hat{C}_t(X_t)) = k + 1].
\end{aligned}$$

Inductive step 2: suppose that the statement is correct for t, k. We now show for $t + 1$:

$$\begin{aligned}
&\mathbb{P}[L_{t+1}^{\mathrm{MC}}(Y_{t+1}, \hat{C}_{t+1}(X_{t+1})) = k] \\
&= \mathbb{P}[L_{t+1}^{\mathrm{MC}}(Y_{t+1}, \hat{C}_{t+1}(X_{t+1})) = k \mid L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t)) = k - 1]\mathbb{P}[L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t)) = k - 1] \\
&= \mathbb{P}[Y_{t+1} \notin \hat{C}_{t+1}(X_{t+1}) \mid L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t)) = k - 1]\mathbb{P}[L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t)) = k - 1] \\
&\leq \mathbb{P}[\tilde{Y}_{t+1} \notin \hat{C}_{t+1}(X_{t+1}) \mid L_t^{\mathrm{MC}}(\tilde{Y}_t, \hat{C}_t(X_t)) = k - 1]\mathbb{P}[L_t^{\mathrm{MC}}(\tilde{Y}_t, \hat{C}_t(X_t)) = k - 1] \\
&= \mathbb{P}[L_{t+1}^{\mathrm{MC}}(\tilde{Y}_{t+1}, \hat{C}_{t+1}(X_{t+1})) = k]
\end{aligned}$$

Finally, we compute the miscoverage counter risk over the time horizon:

$$
\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[L^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t))] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{\infty} k \mathbb{P}[L_t^{\mathrm{MC}}(Y_t, \hat{C}_t(X_t)) = k]
$$

$$
\leq \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{\infty} k \mathbb{P}[L_t^{\mathrm{MC}}(\tilde{Y}_t, \hat{C}_t(X_t)) = k]
$$

$$
= \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[L_t^{\mathrm{MC}}(\tilde{Y}_t, \hat{C}_t(X_t))]
$$

$$
\leq \alpha.
$$

∎

Appendix B. Additional Experimental Details and Results

B.1 Classification: Object Recognition Experiment

Here we present additional results of the classification experiment with CIFAR-10H explained in Section 2.2, but first provide further details about the data set and training procedure. The CIFAR-10H data set contains the same 10,000 images as CIFAR-10, but with labels from a single annotator instead of a majority vote of 50 annotators. We fine-tune a ResNet18 model pre-trained on the clean training set of CIFAR-10, which contains 50,000 samples. Then we randomly select 2,000 observations from CIFAR-10H for calibration. The test set contains the remaining 8,000 samples, but with CIFAR-10 labels. We apply conformal prediction with the APS score. The marginal coverage achieved when using noisy and clean calibration sets are depicted in Figure 1. This figure shows that (i) we obtain the exact desired coverage when using the clean calibration set; and (ii) when calibrating on noisy data, the constructed prediction sets over-cover the clean test labels. Figure 15 illustrates the average prediction set sizes that are larger when using noisy data for calibration and thus lead to higher coverage levels.

B.2 Regression: Aesthetic Visual Rating

Herein, we provide additional details regarding the training of the predictive models for the real-world regression task. As explained in Section 2.2, we use a VGG-16 model—pre-trained on the ImageNet data set—whose last (deepest) fully connected layer is removed. Then, we feed the output of the VGG-16 model to a linear fully connected layer to predict the response. We train two different models: a quantile regression model for CQR and a classic regression model for conformal with residual magnitude score. Both models are trained on $34,000$ noisy samples, calibrated on $7,778$ noisy holdout points, and tested on $7,778$ clean samples. We train the quantile regression model for 70 epochs using 'SGD' optimizer with a batch size of 128 and an initial learning rate of 0.001 decayed every
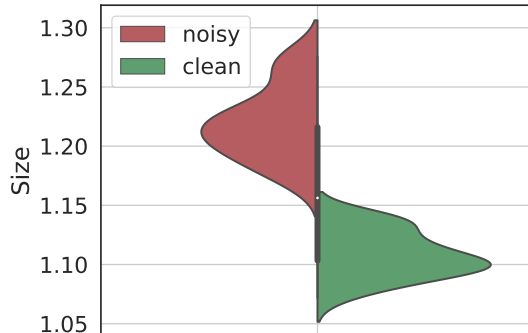
Figure 15: **Effect of label noise on CIFAR-10**. Distribution of average prediction set sizes over 30 independent experiments evaluated on CIFAR-10H test data using noisy and clean labels for calibration. Other details are as in Figure 1

20 epochs exponentially with a rate of 0.95 and a frequency of 10. We apply dropout regularization to avoid overfitting with a rate of 0.2. We train the classic regression model for 70 epochs using 'Adam' optimizer with a batch size of 128 and an initial learning rate of 0.00005 decayed every 10 epochs exponentially with a rate of 0.95 and a frequency of 10. The dropout rate in this case is 0.5.

B.3 Synthetic Classification: Adversarial Noise Models

In contrast with the noise distributions presented in Section 4.1, here we construct adversarial noise models to intentionally reduce the coverage rate.

1. `Most frequent confusion`: we extract from the confusion matrix the pair of classes with the highest probability to be confused between each other, and switch their labels until reaching a total probability of ϵ. In cases where switching between the most common pair is not enough to reach ϵ, we proceed by flipping the labels of the second most confused pairs of labels, and so on.

2. `Wrong to right`: wrong predictions during calibration cause larger prediction sets during test time. Hence making the model think it makes fewer mistakes than it actually does during calibration can lead to under-coverage during test time. Here, we first observe the model predictions over the calibration set, and then switch the labels only of points that were misclassified. We switch the label to the class that is most likely to be the correct class according to the model, hence making the model think it was correct. We switch a suitable amount of labels in order to reach a total switching probability of ϵ (this noise model assumes there are enough wrong predictions in order to do so).

3. `Optimal adversarial`: we describe here an algorithm for building the worst possible label noise for a specific model using a specific non-conformity score. This noise will decrease the calibration threshold at most and as a result, will cause significant
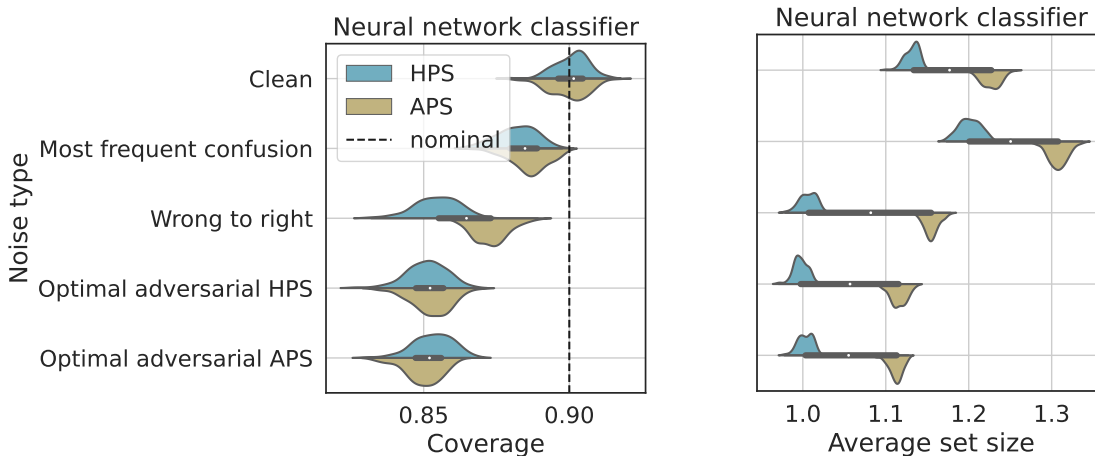
Figure 16: **Effect of label noise on synthetic multi-class classification data**. Performance of conformal prediction sets with target coverage $1 - \alpha = 90\%$, using a noisy training set and a noisy calibration set with adversarial noise models. Left: Marginal coverage; Right: Average size of predicted sets. The results are evaluated over 100 independent experiments.

under-coverage during test time. To do this, we perform an iterative process. In each iteration, we calculate the non-conformity scores of all of the calibration points with their current labels. We calculate the calibration threshold as in regular conformal prediction and then, from the points that have a score above the threshold, we search for the one that switching its label can reduce its score by most. We switch the label of this point to the label that gives the lowest score and then repeat the iterative process with the new set of labels. Basically, at every step, we make the label swap that will decrease the threshold by most.

In these experiments, we apply the same settings as described in Section 4.1 and present the results in Figure 16. We can see that the `optimal adversarial` noise causes the largest decrease in coverage as one would expect. The `most-frequent-confusion` noise decreases the neural network coverage to approximately 89%. The `wrong-to-right` noise decreases the coverage to around 85% with the HPS score and to around 87% with the APS score. This gap is expected as this noise directly reduces the HPS score. We can see that the optimal worst-case noise for each score function reduces the coverage to around 85% when using that score. This is in fact the maximal decrease in coverage possible theoretically, hence it strengthens the optimally of our iterative algorithm.

B.4 Synthetic Regression: Additional Results

Here we first illustrate in Figure 17 the data we generate in the synthetic regression experiment from Section 4.2 and the different corruptions we apply.
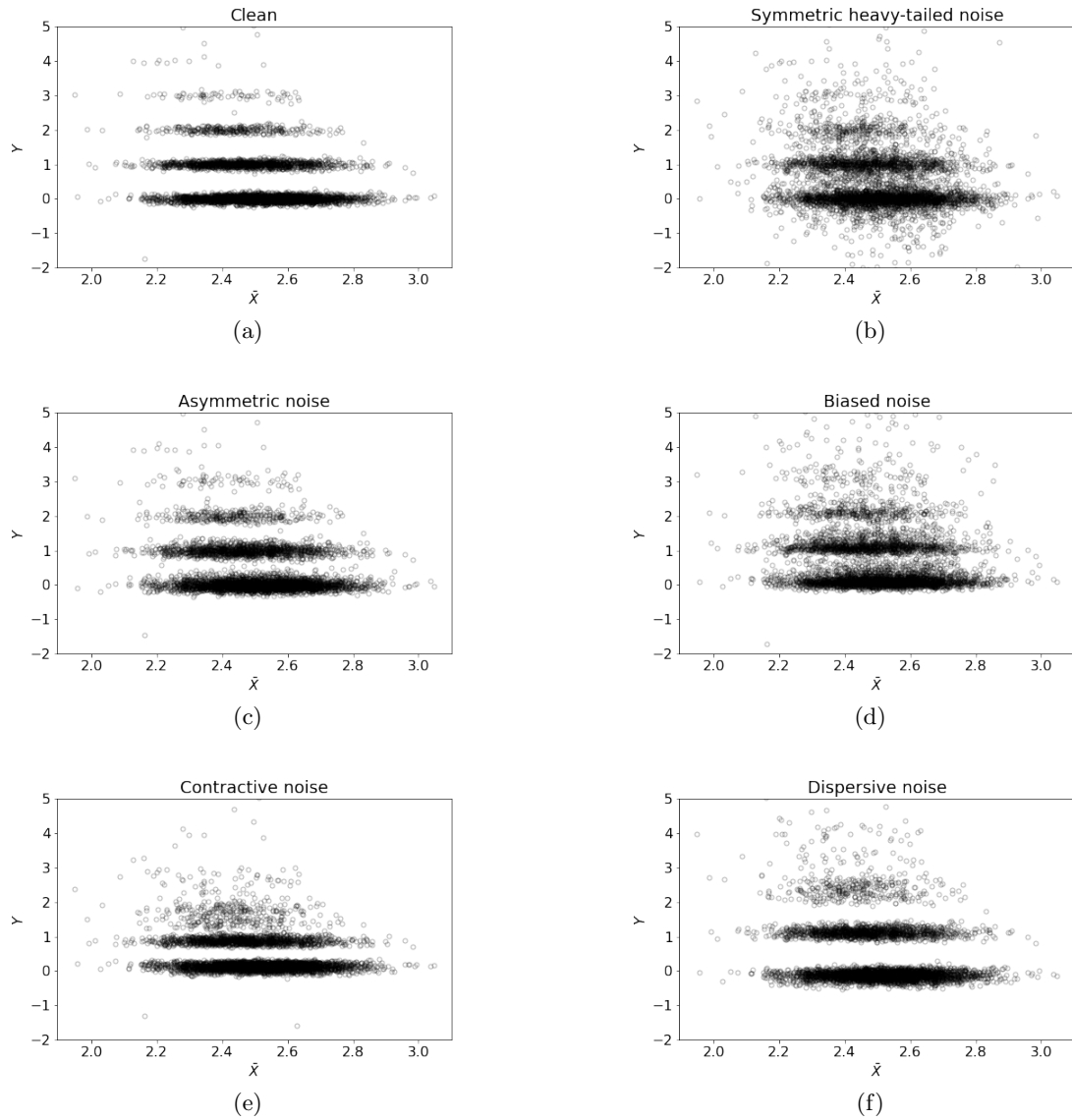
Figure 17: **Illustration of the generated data with different corruptions.** (a): Clean samples. (b): Samples with symmetric heavy-tailed noise. (c): Samples with asymmetric noise. (d): Samples with biased noise. Noise magnitude is set to 0.1. (e): Samples with contractive noise. (f): Samples with dispersive noise.

In Section 4.2 we apply some realistic noise models and examine the performance of conformal prediction using CQR score with noisy training and calibration sets. Here we construct some more experiments using the same settings, however we train the models using clean data instead of noisy data. Moreover, we apply an additional adversarial noise model that differs from those presented in Section 4.2 in the sense that it is designed to intentionally reduce the coverage level.

Wrong to right: an adversarial noise that depends on the underlying trained regression model. In order to construct the noisy calibration set we switch 7% of the responses as follows: we randomly swap between outputs that are not included in the interval predicted by the model and outputs that are included.

Figures 18 and 19 depict the marginal coverage and interval length achieved when applying the different noise models. We see that the adversarial `wrong to right` noise model reduces the coverage rate to approximately 83%. Moreover, these results are similar to those achieved in Section 4.2, except for the conservative coverage attained using `biased` noise, which can be explained by the more accurate low and high estimated quantiles.
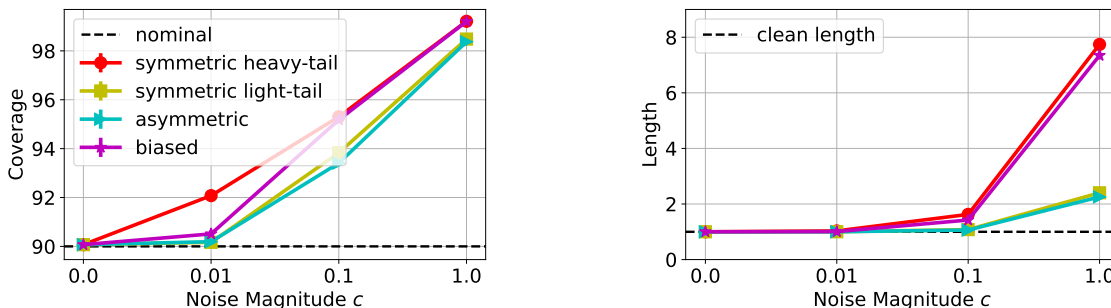

Figure 18: **Response-independent noise.** Performance of conformal prediction intervals with target coverage $1 - \alpha = 90\%$, using a clean training set and a noisy calibration set. Left: Marginal coverage; Right: Length of predicted intervals (divided by the average clean length) using symmetric, asymmetric and biased noise with a varying magnitude. Other details are as in Figure 7.
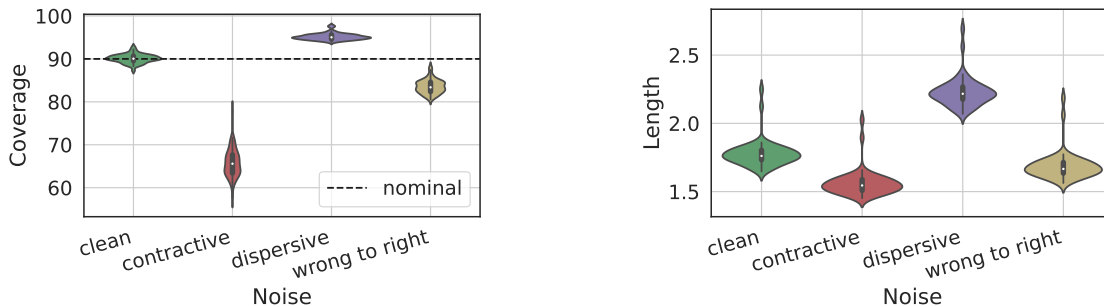

Figure 19: **Response-dependent noise.** Performance of conformal prediction intervals with target coverage $1 - \alpha = 90\%$, using a clean training set and a noisy calibration set. Left: Marginal coverage; Right: Length of predicted intervals. The results are evaluated over 50 independent experiments.

Lastly, in order to explain the over-coverage or under-coverage achieved for some of the different noise models, as depicted in Figures 7 and 8, we present in Figure 20 the CQR scores and their 90%'th empirical quantile. Over-coverage is achieved when the noisy scores

are larger than the clean ones, for example, in the `symmetric heavy tailed` case, and under-coverage is achieved when the noisy scores are smaller.
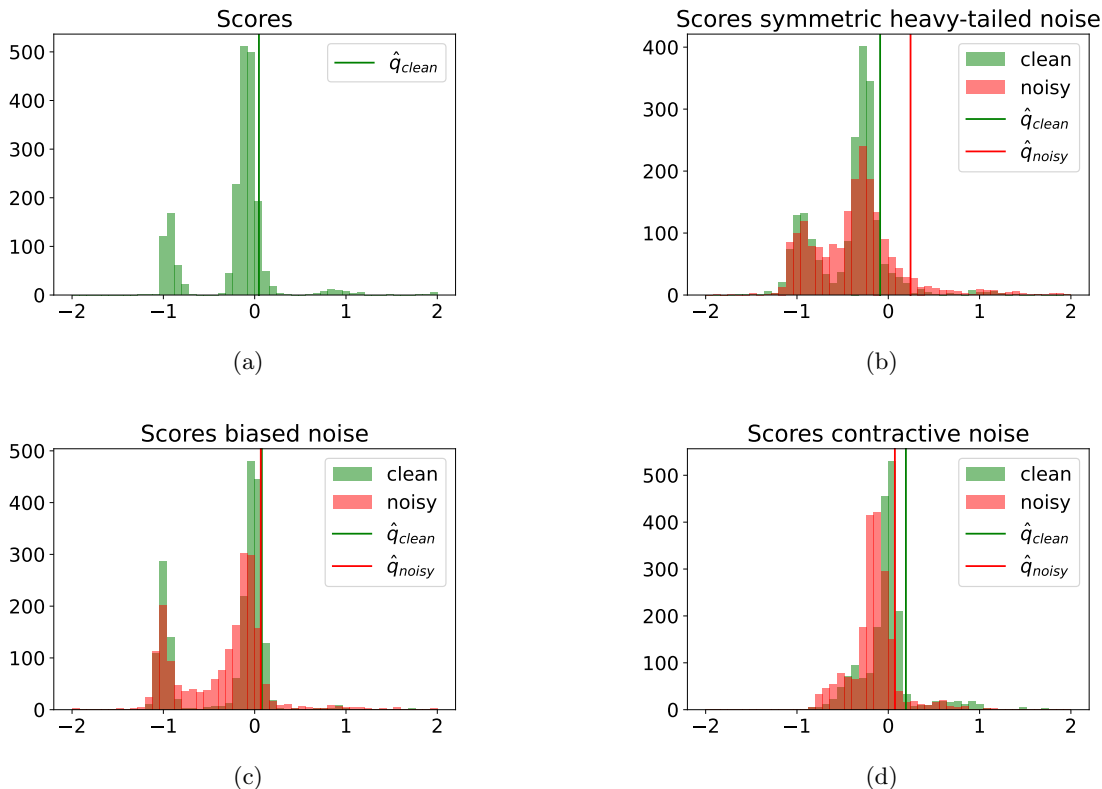

Figure 20: **Illustration of the CQR scores**. (a): Clean training and calibration sets. (b): Symmetric heavy-tailed noise. (c): Biased noise. Noise magnitude is set to 0.1. (d): Contractive noise. Other details are as in Figure 7.

B.5 The Multi-label Classification Experiment with the COCO Data Set

Here, we provide the full details about the experimental setup of the real multi-label corruptions from Section 3.2. We asked 9 annotators to annotate 117 images. Each annotator labeled approximately 15 images separately, except for two annotators who labeled 15 images as a couple. The annotators were asked to label each image under 30 seconds, although this request was not enforced. Figure 21 presents the number of labels that are missed or mistakenly added to each image. On average, 1.25 of the labels were missed and 0.5 were mistakenly added, meaning that each image contains a total of 1.75 label mistakes on average.
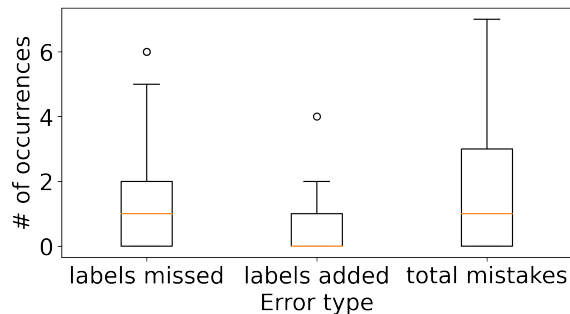
Figure 21: Analysis of human-made label corruptions in MS COCO data set.

B.6 Multi-label Classification with Artificial Corruptions

In this section, we study three types of corruptions that demonstrate Proposition 4 in the sense that the number of positive labels, $|\tilde{Y}|$, is a deterministic function of X. The data set we use in the following experiments is the MS COCO data set (Lin et al., 2014), in which the input image may contain up to $K = 80$ positive labels. We consider the given annotations as ground-truth labels and artificially corrupt them to generate noisy ones, as conducted in Zhao and Gomes (2021); Kumar et al. (2020). The first noise model applies dependent corruptions by taking β of the positive labels and turning them into different ones, according to a pre-defined transition map that was randomly chosen. The second one applies independent corruptions: β of the positive labels turn into different ones, chosen independently and uniformly. Finally, the partial noise model corrupts the labels as the independent one, except for 30 pre-defined labels that remain unchanged. In all experiments, we set $\beta = 0.45$.

We fit a TResNet (Ridnik et al., 2021) model on 100k clean samples and calibrate it using 10K noisy samples with conformal risk control, as outlined in (Angelopoulos et al., 2024, Section 3.2). We control the false-negative rate (FNR), defined in (7), at different labels and measure the FNR obtained over clean and noisy versions of the test set which contains 30k samples. Figure 22 displays the results, showing that the risk obtained over the clean labels is more conservative than the risk of the clean labels. This is not a surprise, as it is supported by our theory, in Proposition 4.

B.7 Regression Miscoverage Bounds Experiment

We repeat the experimental protocol detailed in Section 4.5 with the same data sets and analyze the miscoverage bounds formulated in Section 3.3.3. Here, we apply conformalized quantile regression to control the miscoverage rate at different levels using a noisy calibration set. Furthermore, we choose the miscoverage bound hyperparameters, c, d, from (16) by a grid search over the calibration set with the objective of tightening the miscoverage bound. Figure 23 displays the miscoverage rate achieved on the clean and noisy versions of the test set, along with the miscoverage bound. Importantly, in contrast to Theorem A.1, this bound requires no assumptions on the distribution of $Y \mid X = x$ or on the distribution of
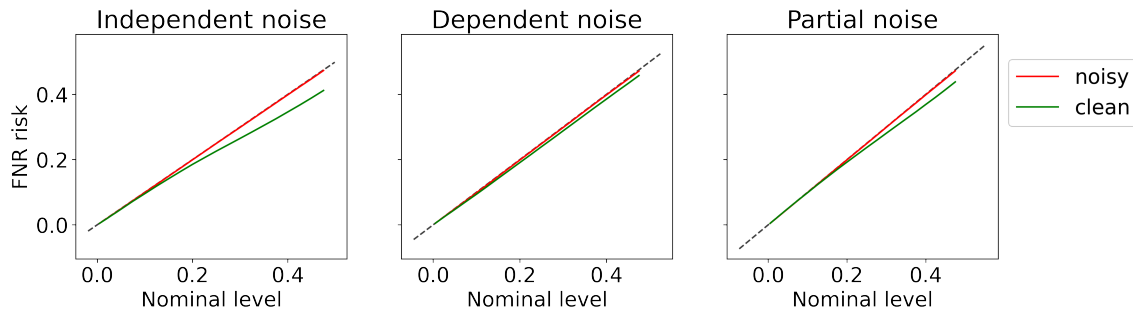
Figure 22: FNR on MS COCO data set, achieved over noisy (red) and clean (green) test sets. The calibration scheme is applied with noisy annotations. Results are averaged over 50 trials.

the noise ε. The only requirement is that the noise is independent of the response variable and has mean 0. This great advantage covers up for the looseness of the bound.
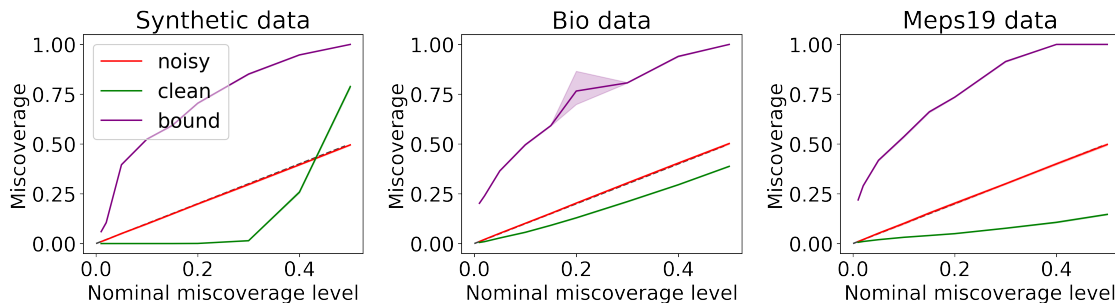

Figure 23: Miscoverage rate achieved over noisy (red) and clean (green) test sets. The calibration scheme is applied using noisy annotations to control the miscoverage level. Results are averaged over 10 random splits of the calibration and test sets.

Appendix C. Related Algorithms

C.1 Conformal Risk Control

Here we provide a pseudo code of the conformal risk control algorithm, following (Angelopoulos et al., 2024).

This prediction set $\widehat{\mathcal{C}}_{\hat{\lambda}}(X_{\text{test}})$ produced by Algorithm 1 satisfies $\mathbb{E}[L(Y_{\text{test}}, \widehat{\mathcal{C}}_{\hat{\lambda}}(X_{\text{test}}))] \leq \alpha$, for proof see (Angelopoulos et al., 2024).

Algorithm 1: Conformal risk control

Input: Exchangeable non-increasing losses, $L_i : \Lambda \to (-\infty, B], i = 1, \ldots, n$, where $L_i(\lambda) = L(Y_i, \widehat{\mathcal{C}}_\lambda(X_i))$, desired risk level α.

Process:

$\hat{R}_n(\lambda) = \frac{1}{n} \cdot (L_1(\lambda) + \cdots + L_n(\lambda))$

$\hat{\lambda} = \inf \left\{ \lambda : \frac{n}{n+1} \hat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha \right\}$

Output: Uncertainty set for a new test point, $\widehat{\mathcal{C}}_{\hat{\lambda}}(X_{\text{test}})$, applied with $\hat{\lambda}$.

C.2 Adaptive Conformal Inference

Below, we provide a pseudo-code of `ACI`, following (Gibbs and Candes, 2021, Algorithm 1).

Algorithm 2: Adaptive Conformal Inference

Input: Data $\{(X_t, Y_t)\}_{t=1}^T \subseteq \mathcal{X} \times \mathcal{Y}$, given as a stream, miscoverage level $\alpha \in (0,1)$, a score function S, a calibration set size n_2, a step size $\gamma > 0$, and a learning model $\mathcal{M}$.

Process: Initialize $\alpha_0 = \alpha$.

for $t = n_2, \ldots, T$ **do**

    Split the observed data $\{(X_i, Y_i)\}_{i=1}^{t-1}$ into a training set of size $t - 1 - n_2$, indexed by $\mathcal{I}_1$ and a calibration set of size n_2, indexed by $\mathcal{I}_2$.

    Obtain a predictive model $\mathcal{M}_t$ by fitting the initial one on the training set $\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}$.

    Construct a prediction set for the new point X_t:

$$\widehat{\mathcal{C}}_t(X_t) = \{y \in \mathcal{Y} : S(\mathcal{M}_t(X_t), y) \leq Q_{1-\alpha_t}(\{S(\mathcal{M}_t(X_i), Y_i)\}_{i \in \mathcal{I}_2})\}$$

    Obtain Y_t.

    Compute $\text{err}_t = \mathbb{1}\{Y_t \notin \widehat{\mathcal{C}}_t(X_t)\}$.

    Update $\alpha_{t+1} = \alpha_t + \gamma(\alpha - \text{err}_t)$.

Output: Uncertainty sets $\widehat{\mathcal{C}}_t(X_t)$ for each time step $t = n_2, \ldots, T$.

C.3 Rolling Risk Control

Here we provide a pseudo code of `Rolling RC`, first developed by (Feldman et al., 2023b).

References

Mohamed Abdalla and Benjamin Fine. Hurdles to artificial intelligence deployment: Noise in schemas and "gold" labels. *Radiology: Artificial Intelligence*, 5(2):e220056, 2023.

Algorithm 3: Rolling Risk Control

Input: Data $\{(X_t, Y_t)\}_{t=1}^T \subseteq \mathcal{X} \times \mathcal{Y}$, given as a stream, desired risk level $r \in \mathbb{R}$, a step size $\gamma > 0$, a set constructing function $f : (\mathcal{X}, \mathbb{R}, \mathbb{M}) \to 2^{\mathcal{Y}}$ and an online learning model $\mathcal{M}$.

Process:

Initialize $\theta_0 = 0$.

for $t = 1$ **to** T **do**

> Construct a prediction set for the new point X_t: $\widehat{\mathcal{C}}_t(X_t) = f(X_t, \theta_t, \mathcal{M}_t)$.
> Obtain Y_t.
> Compute $l_t = L(Y_t, \widehat{\mathcal{C}}_t(X_t))$.
> Update $\theta_{t+1} = \theta_t + \gamma(l_t - r)$.
> Fit the model $\mathcal{M}_t$ on (X_t, Y_t) and obtain the updated model $\mathcal{M}_{t+1}$.

Output: Uncertainty sets $\widehat{\mathcal{C}}_t(X_t)$ for each time step $t \in \{1, ...T\}$.

Görkem Algan and Ilkay Ulusoy. Label noise types and their effects on deep learning. *arXiv preprint arXiv:2003.10471*, 2020.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023. ISSN 1935-8237.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint*, 2021. arXiv:2110.01052.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4): 343–370, 1988.

Javed A Aslam and Scott E Decatur. On the sample complexity of noise-tolerant learning. *Information Processing Letters*, 57(4):189–195, 1996.

Rina Foygel Barber. Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14(2):3487 – 3524, 2020.

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

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816 – 845, 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), September 2021. ISSN 0004-5411.

Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, 11(1):1–14, 2020.

bio. Physicochemical properties of protein tertiary structure data set. `https://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure`. Accessed: January, 2019.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John Duchi. Predictive inference with weak supervision. *arXiv preprint arXiv:2201.08315*, 2022.

Chen Cheng, Hilal Asi, and John Duchi. How many labelers do you have? a closer look at gold-standard labels. *arXiv preprint arXiv:2206.12041*, 2022.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

António Farinhas, Chrysoula Zerva, Dennis Thomas Ulmer, and Andre Martins. Non-exchangeable conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=j511LaqEeP`.

Shai Feldman, Bat-Sheva Einbinder, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Conformal prediction is robust to dispersive label noise. In *Conformal and Probabilistic Prediction with Applications*, pages 624–626. PMLR, 2023a.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023b. ISSN 2835-8856.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2013.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 2013.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Simon Jenni and Paolo Favaro. Deep bilevel learning. In *Proceedings of the European Conference on Computer Vision*, pages 618–633, 2018.

Ishan Jindal, Matthew Nokleby, and Xuewen Chen. Learning deep networks from noisy labels with dropout regularization. In *2016 IEEE 16th International Conference on Data Mining*, pages 967–972. IEEE, 2016.

Yueying Kao, Chong Wang, and Kaiqi Huang. Visual aesthetic quality assessment with a regression model. In *2015 IEEE International Conference on Image Processing*, pages 1583–1587. IEEE, 2015.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

Himanshu Kumar, Naresh Manwani, and PS Sastry. Robust learning of multi-label classifiers under label noise. In *Proceedings of the ACM IKDD CoDS and COMAD*, pages 90–97. 2020.

Yonghoon Lee and Rina Foygel Barber. Binary classification with corrupted labels. *Electronic Journal of Statistics*, 16(1):1367 – 1392, 2022.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

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

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

Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, pages 3355–3364. PMLR, 2018.

meps_19. Medical expenditure panel survey, panel 19. `https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181`. Accessed: January, 2019.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.

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

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021.

Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1400–1409, 2021.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 3543–3553. 2019.

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

Matteo Sesia, YX Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *arXiv preprint arXiv:2309.05092*, 2023.

Pulkit Singh, Joshua C Peterson, Ruairidh M Battleday, and Thomas L Griffiths. End-to-end deep prototype and exemplar models for predicting human behavior. *arXiv preprint arXiv:2007.08723*, 2020.

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

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 2530–2540. 2019.

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

Vladimir Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*, pages 444–453, 1999.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World.* Springer, New York, NY, USA, 2005.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. *Advances in Neural Information Processing Systems*, 32, 2019.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021.

Bodi Yuan, Jianyu Chen, Weidong Zhang, Hung-Shuo Tai, and Sara McMains. Iterative cross learning on noisy labels. In *IEEE Winter Conference on Applications of Computer Vision*, pages 757–765. IEEE, 2018.

Wenting Zhao and Carla Gomes. Evaluating multi-label classifiers with noisy labels. *arXiv preprint arXiv:2102.08427*, 2021.