Conformal Prediction is Robust to Dispersive Label Noise

Shai Feldman* SHAI.FELDMAN@CS.TECHNION.AC.IL
Bat-Sheva Einbinder* BAT-SHEVAB@CAMPUS.TECHNION.AC.IL
Technion - Israel Institute of Technology

Stephen Bates STEPHENBATES@CS.BERKELEY.EDU
Anastasios N. Angelopoulos ANGELOPOULOS@BERKELEY.EDU
University of California, Berkeley

Asaf Gendler ASAFGENDLER@CAMPUS.TECHNION.AC.IL
Yaniv Romano YROMANO@TECHNION.AC.IL
Technion - Israel Institute of Technology

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

In most supervised classification and regression tasks, one would assume the provided labels reflect the ground truth. In reality, this assumption is often violated; see 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. 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. Quantifying the prediction uncertainty is crucial in high-stakes applications in general, and especially so in settings where the training data is inexact. Conformal prediction (Vovk et al., 2005) is a powerful tool for uncertainty quantification which generates prediction sets that represent the plausible outcome. In short, this paper outlines the fundamental conditions under which conformal prediction still works under label noise, and furthermore, studies its behavior with several common score functions and noise models.

Suppose we are given a pre-trained model $\hat{f}$, e.g., a classifier or a regressor, and a hold-out data $\{(X_i, Y_i)\}_{i=1}^n$, 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 dataset, 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 $\widehat{\mathcal{C}}(X_{\text{test}})$ that contains the unknown test response, Y_{test}, with high probability, e.g., 90%. That is, for a user-specified level $\alpha \in (0, 1)$, $\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}(X_{\text{test}})\right) \geq 1-\alpha$. This property is called *marginal coverage*, where the probability is defined over the calibration and test data. 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

*. Equal contribution.

and small otherwise. Abbreviate the scores on each calibration data point as $s_i = s(X_i, Y_i)$ for each $i = 1, ..., n$. Conformal prediction achieves the desired marginal coverage rate by setting $\hat{q}_{\text{clean}} = s_{(\lceil (n+1)(1-\alpha) \rceil)}$ as the $\lceil (n+1)(1-\alpha) \rceil$-smallest of the calibration scores and constructing the prediction sets as $\widehat{\mathcal{C}}(X_{\text{test}}) = \{y \in \mathcal{Y} : s(X_{\text{test}}, y) \leq \hat{q}_{\text{clean}}\}$.

In this paper, we suppose that the observed calibration labels, $\tilde{Y}_1, \ldots, \tilde{Y}_n$, are corrupted, while their clean versions are unavailable, so we cannot calculate $\hat{q}_{\text{clean}}$. In general, $\tilde{Y}_i = g(Y_i)$ for some *corruption function* $g : \mathcal{Y} \times [0, 1] \to \mathcal{Y}$, so the i.i.d. assumption and marginal coverage guarantee of conformal prediction are invalidated. 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 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. Formally, our objective is to find the conditions under which a valid coverage rate is obtained: $\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \geq 1 - \alpha$. We argue that prediction sets will produce valid coverage 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 $\hat{q}_{\text{clean}}$. Formally, we claim that if $\mathbb{P}(\tilde{s}_{\text{test}} \leq t) \leq \mathbb{P}(s_{\text{test}} \leq t)$ for all t, then the desired coverage rate is obtained on the unobserved clean labels $\mathbb{P}\left(Y_{\text{test}} \in \widehat{\mathcal{C}}_{\text{noisy}}(X_{\text{test}})\right) \geq 1 - \alpha$. The stochastic dominance requirement depends on the clean data distribution, the noise model, the ML model performance, and the score we use to construct the sets. In the paper, we present realistic experiments and mathematical analysis of cases in which this assumption holds.

In real-world applications, it is often desired to control metrics other than the miscoverage loss, as in segmentation or multi-label classification tasks. Examples of such alternative losses include the F1-score or the false negative rate, where the latter is especially relevant for high-dimensional responses. There have been developed extensions of the conformal framework that go beyond the 0-1 loss, providing a rigorous *risk control* guarantee over general loss functions (Bates et al., 2021; Angelopoulos et al., 2021, 2022). Formally, these techniques take a loss function L that measures prediction error and generate uncertainty sets with a risk controlled at pre-specified level α: $\mathbb{E}[L(Y_{\text{test}}, \widehat{\mathcal{C}}(X_{\text{test}})] \leq \alpha$. 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. Nonetheless, as outlined above, we argue that conservative sets are obtained when the distribution of the losses of noisy labels dominates the distribution of the losses of clean labels.

In this work, we present theoretical examples under which the stochastic dominance requirement holds, meaning that the conformal prediction and risk-controlling frameworks applied with noisy labels generate valid uncertainty sets with respect to the true, noiseless label. These examples include additive symmetric noise in the regression case, or dispersive corruptions in single-label or multi-label classification tasks. Furthermore, we conduct extensive synthetic and real-data experiments of various tasks, including classification, regression, segmentation, and more. In all these experiments, conformal prediction normally achieves conservative coverage/risk even with access only to noisy labels, unless the noise is adversarially designed. If such noise can be avoided, a user should feel safe deploying conformal prediction with noisy labels.

References

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 N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

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.

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.

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

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

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.

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