Class-Conditional Robust Conformal Prediction for Structured Perturbations

Luis Marchante Arjona LUIS.MARCHANTEARJONA@DLR.DE

Protim Bhattacharjee PROTIM.BHATTACHARJEE@DLR.DE
German Aerospace Center (DLR), Berlin, Germany

Peter Jung PETER.JUNG@DLR.DE
German Aerospace Center (DLR), Berlin, Germany
Technical University of Berlin (TUB), Berlin, Germany

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

1. Introduction and Motivation

We introduce a conformal prediction (CP) Vovk et al. (2005) method that leverages class-conditional randomized smoothing with localized perturbations to address structured corruption in AI-driven multispectral image classification. This setting reflects realistic corruptions such as those found in remote sensing where specific object categories may be disproportionately affected by environmental or hardware-induced fluctuations, e. g. specifically only in red spectral channel or near-infrared channels. The Randomized Smoothed Conformal Prediction (RSCP) framework Gendler et al. (2022) makes use of global uniform noise to construct valid prediction sets. Since real-world perturbation are rarely uniform, RSCP would lead to an increased prediction set size for uncorrupted classes, reducing informativeness and efficiency of the conformal method. For such asymmetric perturbations, we propose a class-conditional RSCP framework in which perturbations are applied only to certain target classes. Our approach allows for risk-stratified robustness, providing more nuanced uncertainty estimates to noise-prone classes without sacrificing coverage for unaffected categories. Class conditional coverage guarantees for smoothed scores with class-dependent noise is guaranteed via (Angelopoulos et al., 2025, Theorem 4.9).

2. Experiments

We evaluate our method using a ResNet50 model trained on high-resolution RGB satellite images from an augmented version Dahiya (2020) of the UC Merced Land Use Dataset Yang and Newsam (2010), using only 5 classes with 500 images per class. The training set contained 2000 samples. A hold-out evaluation set of 250 samples was used for both validation and testing due to the limited dataset size. The conformal calibration set consisted of another 250 samples. For adding structured corruptions at a specific class y, we selectively inject l_2-norm bounded additive noise, δ^y, into a specific spectral channel (e.g. red channel). Let $(X_i, Y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ denote n calibration samples. Let $\mathcal{K} \subseteq \mathcal{Y}$ be the set of classes subject to spectral perturbations. Smoothed calibration scores per class are calculated according to the procedure in Gendler et al. (2022). The non-conformity score for each class with a

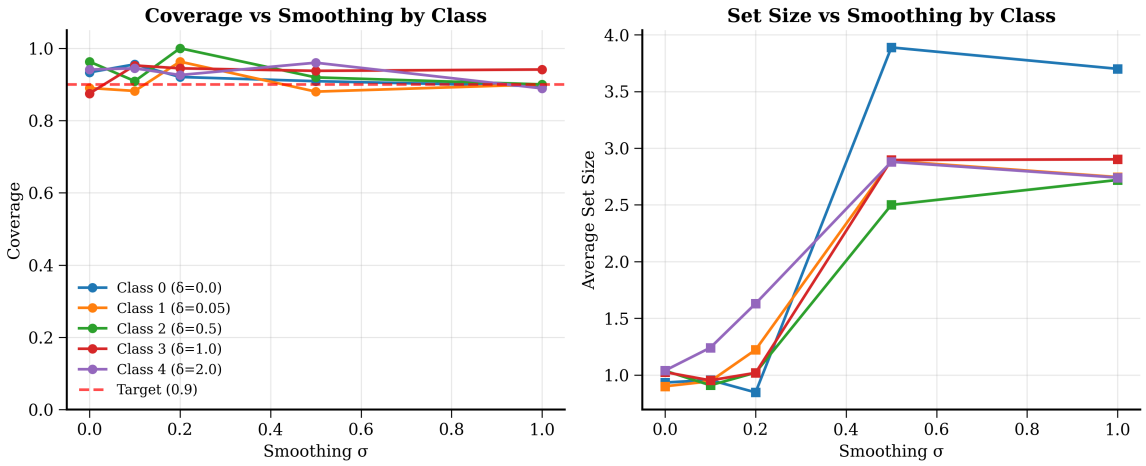

Figure 1: Empirical coverage and average prediction set size as a function of red-channel perturbation strength (δ) and smoothing level (σ). Coverage is maintained across classes despite increasing corruption levels, with smoothing helping to counteract the effect of perturbation. Set sizes increase sharply beyond a critical σ, suggesting a trade-off between robustness and efficiency.

pre-trained neural network with a softmax output, $f(\cdot)$ is defined as $s(f(x), y) = 1 - f^y(x)$. Let $\mathcal{D}_y = \{s_i : Y_i = y\}$ denote the set of calibration indices for class $y \in \mathcal{Y}$. Define the class-conditional $1-\alpha$ confidence quantile, $q^y_{1-\alpha} = \text{Quantile}_{1-\alpha}(\mathcal{D}_y)$. The class-conditional prediction set for a test sample X_{n+1} with σ as the smoothing factor is constructed as:

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(f(\tilde{X}_{n+1}), y) \leq q^y_{1-\alpha} + \frac{\delta^y}{\sigma} \right\}, \text{where} \begin{cases} \tilde{X}_{n+1} = X_{n+1} + \delta^y & \text{if } y \in \mathcal{K} \\ \tilde{X}_{n+1} = X_{n+1} & \text{otherwise.} \end{cases}$$

In real-world applications, such perturbation strength, δ^y, can be estimated different ways, through noise modelling, Rasti et al. (2018); Cerra et al. (2014), and more recently via machine learning methods Wischow et al. (2024).

3. Conclusion

Our study demonstrates that conformal prediction methods can be adapted to handle realistic, structured perturbations. The proposed approach preserves theoretical coverage guarantees under conditional exchangeability and enables selective robustness for critical classes. This is beneficial in applications like remote sensing and autonomous systems where structured noise is prevalent. The RSCP framework of smoothed conformal scores is combined with class-conditional calibration. Smoothing enables to maintain coverage, however, efficiency is dependent on σ. A critical σ is observed, suggesting a trade-off between robustness and efficiency. As expected the unperturbed class remains efficient in comparison to the δ-perturbed classes below this σ value. This framework can be naturally extended to Mondrian conformal prediction, enabling finer-grained calibration and risk control across groups or taxonomies.

References

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction, 2025. URL https://arxiv.org/abs/2411.11824.

Daniele Cerra, Rupert Müller, and Peter Reinartz. Noise reduction in hyperspectral images through spectral unmixing. *IEEE Geoscience and Remote Sensing Letters*, 11(1):109–113, 2014. doi: 10.1109/LGRS.2013.2247562.

Gotam Dahiya. UC Merced Land-Use Scene Classification Dataset. https://www.kaggle.com/datasets/apollo2506/landuse-scene-classification, 2020.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9L1BsI4wP1H.

Behnood Rasti, Paul Scheunders, Pedram Ghamisi, Giorgio Licciardi, and Jocelyn Chanussot. Noise reduction in hyperspectral imagery: Overview and application. *Remote Sensing*, 10(3), 2018. ISSN 2072-4292. doi: 10.3390/rs10030482.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, 2005. ISBN 0-387-00152-2. doi: 10.1007/b106715.

Maik Wischow, Patrick Irmisch, Anko Börner, and Guillermo Gallego. Real-time noise source estimation of a camera system from an image and metadata. *Advanced Intelligent Systems*, (230047), April 2024.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.