Reliable Uncertainty Quantification in Machine Learning via Conformal Prediction

Yuanjie Shi

Washington State University, Pullman, WA 99163
yuanjie.shi@wsu.edu

Abstract

Deploying machine learning (ML) models in high-stakes domains such as healthcare and autonomous systems requires reliable uncertainty quantification (UQ) to ensure safe and accurate decision-making. Conformal prediction (CP) offers a robust, distribution-agnostic framework for UQ, providing valid prediction sets that guarantee a specified coverage probability. However, existing CP methods are often limited by assumptions that are violated in real-world scenarios, such as non-i.i.d. data, and by a lack of integration with modern machine learning workflows, particularly in large generative models. This research aims to address these limitations by advancing CP techniques to operate effectively in non-i.i.d. settings, improving predictive efficiency without sacrificing theoretical guarantees, and integrating CP directly into model training processes. These developments will enhance the practical applicability of CP for a wide range of ML tasks, enabling more reliable and interpretable models in high-stakes applications.

Research Questions

The primary questions guiding this research are:

- What techniques can be developed to maintain the validity of CP methods in real-world, non-i.i.d. (independent and identically distributed) data settings, where the assumptions of CP may be violated?
- How can CP methods be extended to large generative models, such as large language models (LLMs), to offer reliable uncertainty quantification in tasks like text generation, where predicting a set of possible outcomes is critical for handling ambiguous or open-ended queries?
- What strategies can be employed to improve the predictive efficiency (e.g., expected cardinality of prediction sets) of CP, while preserving its theoretical guarantees?
- How can CP be effectively integrated into machine learning models to ensure flexible and reliable uncertainty quantification across a wide range of high-stakes applications, such as healthcare and autonomous systems?

Related Work

Conformal prediction (CP) is a popular framework for uncertainty quantification (UQ) that measures the confidence of model in its predictions (Vovk, Gammerman, and Shafer 2005). CP converts model outputs (e.g., Softmax scores) into prediction sets that contain the true label with a specified probability, or coverage (e.g., 90%) (Romano, Sesia, and Candes 2020). This is achieved through a calibration step based on non-conformity scores, offering formal reliability guarantees, which is critical for high-stakes applications like medical diagnosis (Lu et al. 2022). Most works use split CP (Romano, Sesia, and Candes 2020), which calibrates on a held-out set with a pre-trained model, assuming i.i.d. data. Some CP works (Lu et al. 2023) extend CP to non-i.i.d. settings but faces challenges in real-world scenarios. Though CP has been applied in classification (Romano, Sesia, and Candes 2020) and regression (Romano, Patterson, and Candes 2019), its use in large generative models (e.g., LLMs) remains limited. Some studies (Mohri and Hashimoto 2024) have explored uncertainty quantification in text generation, but the general framework is unestablished. Improving prediction efficiency while ensuring valid coverage is an ongoing issue (Angelopoulos et al. 2020). Recent advances in conformity scores (Angelopoulos et al. 2020) and calibration methods (Ghosh et al. 2023) help, but there is no established framework for prediction efficiency. Some research integrates CP into model training to improve efficiency, known as conformal training (Stutz et al. 2021). However, the theoretical frameworks of learning bound and optimization analysis remain limited.

Research Plan and Proposed Contributions

The goal of this research is to extend CP for real-world scenarios, focusing on its use with non-i.i.d. data, large generative models, improving efficiency, and integrating CP into ML training. The expected contributions are:

- **CP in non-i.i.d. settings:** I will develop methods to adapt CP for non-i.i.d. data, particularly in federated learning (FL), using similarity measures and importance weighting to ensure valid coverage under distribution shifts.
- **Extending CP to generative models:** I plan to apply CP to large generative models (e.g., LLMs) for tasks like text

generation, providing reliable uncertainty quantification in ambiguous or open-ended cases.

- **Improving predictive efficiency:** I will explore nonconformity score distributions to build a theoretical framework that enhances CP's predictive efficiency without sacrificing validity.
- **Integrating CP into ML training:** I will incorporate CP principles directly into the ML training process as a regularization method, improving both accuracy and uncertainty quantification from the training phase onward.

These contributions aim to enhance CP's theoretical foundation and practical use in various machine learning tasks, including deep learning and generative models.

Research Progress and Timeline

Progress as of September 30, 2024

- **CP in non-i.i.d. settings:** I have completed a comprehensive literature review, designed the theoretical framework, and drafted the initial version of the manuscript.
- **Extending CP to generative models:** I have conducted a literature review and identified key challenges.
- **Improving predictive efficiency:** The literature review, framework design, and draft of the manuscript are complete. I have also begun implementing the code on the CIFAR100 dataset (Krizhevsky, Hinton et al. 2009).
- **Integrating CP into ML training:** I have completed the literature review, designed the framework, and written the first version of the manuscript. Theoretical analysis is also complete, and implementation is underway on the CIFAR-100 dataset.

Anticipated Progress by February 25-26, 2025

- **CP in non-i.i.d. settings:** By the workshop date, I plan to complete the code implementation, extend the approach to real-world datasets, and finalize the theoretical proof.
- **Extending CP to generative models:** I plan to complete the remaining tasks, including framework design, code implementation, and empirical validation. By early 2025, I aim to have initial results and a draft of the manuscript ready for submission.
- **Improving predictive efficiency:** I will extend the predictive efficiency methods to large-scale datasets and complete the theoretical analysis. By early 2025, I plan to have the manuscript ready for submission, showcasing the improvements in efficiency of CP.
- **Integrating CP into ML training:** I will finalize the implementation on larger datasets, revise the manuscript based on new findings, and submit the final version for publication.

Contributions of Collaborators

Thesis Supervisor: Dr. Yan Yan and Dr. Jana Doppa have identified key challenges in the research problems related to CP, assisted in formulating the research questions, and helped refine the manuscripts. **Collaborating PhD Student**: Seyed Hooman Shahrokhi contributed by proofreading the theoretical analysis for integrating CP into ML training. Subhankar Ghosh developed the first version of the code for the same task.

Conclusion and Future Directions

This research advances CP for non-i.i.d. settings, generative models, and improved predictive efficiency, while integrating CP into ML training. Key progress has been made in framework design, theoretical analysis, and initial implementation. Future work will focus on extending these methods to real-world datasets, refining efficiency, and applying CP to generative and reinforcement learning models. These efforts aim to enhance theoretical foundations of CP and expand its practical use in high-stakes applications.

Acknowledgements

I thank my supervisor, Dr. Yan Yan, for his invaluable guidance and support throughout this research. I also acknowledge Seyed Hooman Shahrokhi and Subhankar Ghosh for their contributions to the theoretical analysis and implementation. Lastly, I am grateful to the AAAI Doctoral Consortium for the opportunity to present and discuss my work.

References

Angelopoulos, A.; Bates, S.; Malik, J.; and Jordan, M. I. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Ghosh, S.; Belkhouja, T.; Yan, Y.; and Doppa, J. R. 2023. Improving Uncertainty Quantification of Deep Classifiers via Neighborhood Conformal Prediction: Novel Algorithm and Theoretical Analysis. *arXiv preprint arXiv:2303.10694*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

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

Lu, C.; Yu, Y.; Karimireddy, S. P.; Jordan, M.; and Raskar, R. 2023. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, 22942–22964. PMLR.

Mohri, C.; and Hashimoto, T. 2024. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*.

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

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33: 3581–3591.

Stutz, D.; Cemgil, A. T.; Doucet, A.; et al. 2021. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.