Optimisation Problems in Constrained Machine Learning

Ruihan Zhang

Singapore Management University
81 Victoria St, Singapore, Singapore 188065
pxzhang@smu.edu.sg

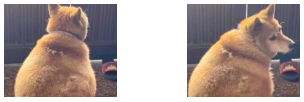
Abstract

Machine learning is widely used in various areas. However, the current machine learning framework remains vulnerable to issues such as adversarial attacks, fairness violations, and data leakage. These problems are not adequately captured by fitting models to collected data and focusing on test performance metrics alone, like accuracy or F1-score. In practice, machine learning tasks often involve additional quantities of interest, which turns an originally unconstrained optimisation problem (only optimising toward accuracy) into a constrained one. This thesis formally studies machine learning under different types of commonly concerning constraints, such as robustness, fairness, and privacy. I first focus on how the formal machine learning framework can be extended to incorporate robustness, which is a critical factor for safety. After that, I turn to more ethics-related aspects like fairness and privacy, to explore the possibility of formally fitting them into machine learning. My approach differs from empirically pushing up multiple metrics and instead emphasises fundamental ways to understand and address the underlying challenges.

Machine Learning Constrained by Perturbation Robustness

Robustness of machine learning (ML) is the first constraint I explore, because adversarial examples pose a security threat to many critical systems. This part contains 4 works.

I started from a well-known fact that truly improving robustness (deterministic robustness) comes at a cost of sacrificing model accuracy. I wonder if there is a certain fundamental limit on achieving robustness whilst maintaining accuracy. In my first work (Zhang and Sun 2024), I offer a Bayes error perspective to robustness analysis. Bayes error characterises the unavoidable error due to data distribution uncertainties (as illustrated in fig. 1). I use this concept to investigate the limit of certified robust accuracy and prove that the accuracy inevitably decreases in the pursuit of robustness due to the changed Bayes error in the altered data distribution. Further, a quantitative upper bound for certified robust accuracy is established, considering the distribution of individual classes and their boundaries. These theoretical results are then empirically evaluated on real-world datasets and are indeed consistent with the limited success of existing certified training results. For example, for CIFAR-10, the analysis estimates an upper bound (of certified robust accuracy) of 67.49%, meanwhile existing approaches are only able to increase it from 53.89% in 2017 to 62.84% in 2023, and have not advanced further since then.



(a) Is this a cat? (b) Eh, it's a dog.

Figure 1: Inherent uncertainty. The picture at left may look like a cat. In fact, it can be the back of a dog.

Since achieving deterministic robustness (*i.e.*, complete elimination of adversarial samples) always comes at an unacceptable accuracy cost, I continue to investigate in my second work (Zhang and Sun 2025a) whether probabilistic robustness (where a small portion (κ) of adversarial examples is tolerable in each vicinity) would suffer from the same issue. Also, I investigate the specific relationship between κ and this potential bound. I find that while Bayes uncertainty does affect probabilistic robustness, its impact is smaller than that on deterministic robustness. In other words, probabilistic robust accuracy can also never reach 1, but can get much closer than deterministic robust accuracy does. Moreover, the upper bound of probabilistic robust accuracy monotonically increases as κ grows, which is also empirically observed. Additional intriguing theoretical facts have been discovered. One, with optimal probabilistic robustness, each probabilistically robust input must also be deterministically robust in a smaller vicinity. Two, voting within the vicinity always improves probabilistic robust accuracy, which is intuitive but not previously proved.

Understanding robustness well, I continue to improve it, provably. While many methods already exist to build robust models, how to build certifiably robust yet accurate neural network models remains an open problem. For example, adversarial training only improves empirical robustness. Conversely, certified training provides certified robustness with the help of neural network verification techniques, but at the cost of a significant accuracy drop. In my third work (Zhang, Zhang, and Sun 2023), I propose a novel ap-

proach that aims to achieve both high accuracy and certified probabilistic robustness by integrating training and testing synergistically. This method has two parts, which together achieve the goal, *i.e.*, a probabilistic robust training method with minimising divergence variance in a given vicinity and a runtime inference method based on robustness testing. Experimentally, my approach significantly outperforms existing approaches in both certification rate and accuracy. Furthermore, it achieves the best defence against powerful attacking methods such as AutoAttack.

In my fourth work (Zhang and Sun 2025c), I explore the robustness evaluation, with a specific case study on face recognition. Face recognition is a widely used authentication technology in practice, where its robustness evaluation is especially necessary. Existing approaches to evaluating the robustness of face recognition systems are either based on empirical evaluation (*e.g.*, measuring attacking success rate using state-of-the-art attacking methods) or measuring the Lipschitz constant. While the former demands significant user efforts and expertise, the latter is extremely time-consuming. In pursuit of a comprehensive, efficient, easy-to-use and scalable estimation of the robustness of face recognition systems, I take an alternative approach and introduce RobFace, *i.e.*, evaluation using an optimised test suite. It contains different-level transferable adversarial face images that are designed to evaluate a face recognition system's robustness along a variety of dimensions. RobFace is system-agnostic and still consistent with system-specific empirical evaluation or formal analysis. This claim is supported by extensive experimental results with various perturbations on multiple face recognition systems. RobFace is also the first system-agnostic robustness estimation test suite.

ML with an Individual Fairness Constraint

Fairness in machine learning is more important than ever as ethical concerns continue to grow. Individual fairness demands that individuals differing only in sensitive attributes receive the same outcomes. However, commonly used machine learning algorithms often fail to achieve such fairness. To improve individual fairness, various training methods have been developed, such as incorporating fairness constraints as optimisation objectives. While these methods have demonstrated empirical effectiveness, they lack formal guarantees of fairness. Existing approaches that aim to provide fairness guarantees primarily rely on verification techniques, which can sometimes fail to produce definitive results. Moreover, verification alone does not actively enhance individual fairness during training. To address this limitation, I propose in my fifth work (Zhang and Sun 2025b) a novel framework that formally guarantees individual fairness throughout training. My approach consists of two parts, *i.e.*, (1) provably fair initialisation that ensures the model starts in a fair state, and (2) a fairness-preserving training algorithm that maintains fairness as the model learns. A key element of this method is the use of randomised response mechanisms, which protect sensitive attributes while maintaining fairness guarantees. I formally prove that this mechanism sustains individual fairness throughout the training process. Experimental evaluations also confirm that this approach is effective. *i.e.*, producing models that are empirically fair and accurate.

Unlearnability, a Constraint for Training Data

The recent success of machine learning models, especially large-scale models, relies heavily on training with massive data. These data are often collected from online sources. This raises serious concerns about the protection of user data, as individuals may not have given consent for their data to be used in training. To address this concern, recent studies introduce the concept of unlearnable examples, *i.e.*, data instances that appear natural but are intentionally altered to prevent models from effectively learning from them. While existing methods demonstrate empirical effectiveness, they typically rely on heuristic trials and lack formal guarantees. Besides, when unlearnable examples are mixed with clean data, as is often the case in practice, their unlearnability disappears. In my most recent work (Zhang et al. 2025), I propose a novel approach to constructing unlearnable examples by systematically maximising the Bayes error, a measurement of irreducible classification error. I develop an optimisation-based approach and provide an efficient solution using projected gradient ascent. The proposed method provably increases the Bayes error and remains effective when the unlearning examples are mixed with clean samples. Experimental results across multiple datasets and model architectures are consistent with my theoretical analysis and show that this approach can restrict data learnability, effectively in practice.

Future Works

I have so far studied each constrained machine learning problem with a single constraint. For the rest of my PhD, I plan to explore how multiple constraints jointly affect the ML framework, *e.g.*, both robustness and individual fairness.

References

Zhang, R.; and Sun, J. 2024. Certified Robust Accuracy of Neural Networks Are Bounded Due to Bayes Errors. In Gurfinkel, A.; and Ganesh, V., eds., *Computer Aided Verification*, 352–376. Cham: Springer Nature Switzerland. ISBN 978-3-031-65630-9.

Zhang, R.; and Sun, J. 2025a. Are Probabilistic Robust Accuracy Bounded.

Zhang, R.; and Sun, J. 2025b. Correct-by-Construction: Certified Individual Fairness through Neural Network Training. *Proc. ACM Program. Lang.*, 9(OOPSLA2).

Zhang, R.; and Sun, J. 2025c. RobFace: A Test Suite for Efficient Robustness Evaluation of Face Recognition Systems. *IEEE Transactions on Reliability*, 1–14.

Zhang, R.; Zhang, P.; Lim, E.-P.; and Sun, J. 2025. Towards Provably Unlearnable Examples via Bayes Error Optimisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, R.; Zhang, P.; and Sun, J. 2023. Towards Certified Probabilistic Robustness with High Accuracy. arXiv:2309.00879.