Regularized Robustly Reliable Learning

As machine learning systems are increasingly deployed in high-stakes and interactive environments, it becomes important not only that they make accurate predictions on average, but that they know when an individual prediction can be trusted. This concern is especially acute under instance-targeted data poisoning attacks, where an adversary corrupts the training data with the specific goal of inducing an error on a particular test point. Prior work of Balcan, Blum, Hanneke, and Sharma [1] introduced the framework of robustly reliable learning for this setting: a learner may abstain, but whenever it does predict, it provides a per-instance certificate that the prediction is correct so long as the target belongs to the assumed hypothesis class and the adversary has not exceeded a stated corruption budget. This gives a stronger guarantee than stability certificates, which only certify that the prediction would not change under a bounded attack; robust reliability instead certifies correctness. That work characterized the optimal certifiable region, gave nearly matching upper and lower bounds, and provided efficient algorithms for important cases such as linear separators over log-concave distributions.

In this work [2], we address two challenges left open by that framework. First, for highly expressive hypothesis classes, the original definition can become vacuous: if two classifiers both fit the training data perfectly but disagree on a test point, then the learner must abstain, even when one explanation is much simpler or more natural than the other. We introduce regularized robustly reliable learners, which incorporate a complexity or "unnaturalness" measure into the certificate. The learner now outputs a prediction together with complexity levels certifying that the prediction is correct whenever the true target is sufficiently simple and the poisoning budget is sufficiently small. This allows nontrivial guarantees even for flexible hypothesis classes by distinguishing plausible low-complexity explanations from arbitrary high-complexity alternatives.

Second, the generic robustly reliable learner from prior work is computationally impractical in many settings because it may require re-solving an empirical risk minimization problem, essentially retraining, for each test point. We study when reliability certificates can instead be produced substantially faster than retraining. Using ideas from dynamic algorithms, we give efficient regularized robustly reliable learners in several settings, including algorithms based on bidirectional dynamic programming and dynamic maximum matching. These results suggest a broader principle for reliable AI systems: in some cases, it is possible to preserve strong per-instance correctness guarantees while shifting part of the work from full retraining to faster test-time certification.

Overall, the paper develops a framework for learning systems that can make trustworthy predictions under instance-targeted adversarial attacks, while also accounting for model complexity and computational efficiency. This connects naturally to the study of reliable and adaptive learning in agentic environments, where systems must operate under feedback, strategic behavior, and uncertainty, and where knowing when and why a prediction is trustworthy is as important as the prediction itself.

References

[1] Maria-Florina Balcan, Avrim Blum, Steve Hanneke, and Dravyansh Sharma. Robustly-reliable learners under poisoning attacks. In *Proceedings of the 35th Annual Conference on Learning Theory (COLT)*, PMLR 178:4498-4534, 2022.

[2] Avrim Blum and Donya Saless. Regularized robustly reliable learners. In *Proceedings of The 37th International Conference on Algorithmic Learning Theory (ALT)*, PMLR 313:1-35, 2026.