Computable Model-Independent Bounds for Adversarial Quantum Machine Learning

Bacui Li[1], Tansu Alpcan[1], Chandra Thapa[2], and Udaya Parampalli[3]

[1]Department of Electrical and Electronic Engineering Department, University of Melbourne, Australia
[2]Data61, CSIRO, Sydney, Australia
[3]School of Computing and Information Systems, University of Melbourne, Australia

Abstract

Machine learning technologies, increasingly embedded in critical applications, raise significant security concerns, particularly due to their susceptibility to adversarial manipulations. This paper extends classical machine learning methodologies to quantum machine learning (QML) to establish computable lower bounds on adversarial risk. By assuming robust ground truth and worst-case attack scenarios, we provide bounds on experimental error rate against adversarial attacks utilizing the lower bound on adversarial risk, contributing to developing more secure quantum models.

We introduce a novel, model-independent method to evaluate the optimal robustness of quantum machine learning models under adversarial attacks. By leveraging concepts from classical machine learning, our computational approach efficiently estimates the lower bounds of adversarial errors for both classical and quantum machine learning models. We validate our bound-evaluation procedure through experiments that compare derived bounds with actual adversarial error rates in QML models, demonstrating the potential for high robustness in quantum systems.

Our contributions are threefold. First, we extend classical adversarial risk lower bound estimation methods to quantum contexts, specifically addressing quantum perturbation attacks. Second, we develop a new quantum attack strategy inspired by the classical Projected Gradient Descent (PGD) attack. Third, we validate our findings against empirical models and attacks, demonstrating a strong correlation between the derived bounds and observed adversarial error rates in quantum models. Experimental results using QML models on MNIST and Fashion-MNIST datasets consistently show that adversarial error rates are bounded by the estimated lower bounds, confirming the effectiveness of our proposed methodology.

The experiments also reveal a trade-off between non-adversarial accuracy and adversarial robustness, influenced by factors such as softmax temperature during training. Our findings suggest that quantum models exhibit a certain degree of inherent robustness.

In conclusion, our work provides a robust framework for evaluating and improving the security of QML models. The proposed adversarial risk lower bounds offer valuable references for developing resilient quantum machine learning systems and underscore the potential of QML in achieving high robustness against sophisticated adversarial attacks.

Keywords : Adversarial machine learning, quantum machine learning, robustness, Projected Gradient Descent (PGD), quantum perturbation attack.