Adversarial Quantum Machine Learning: An Information-Theoretic Generalization Analysis

Petros Georgiou, Sharu Theresa Jose and Osvaldo Simeone

In a manner analogous to their classical counterparts, quantum classifiers are vulnerable to adversarial attacks that perturb their inputs. A promising countermeasure is to train the quantum classifier by adopting an attack-aware, or adversarial, loss function. This paper studies the generalization properties of quantum classifiers that are adversarially trained against bounded-norm white-box attacks.

Problem Setting: As illustrated in Fig. 1(a), a classical input x is embedded into a quantum state $\rho(x)$ by a fixed and known *quantum embedding* map $x \mapsto \rho(x)$. Let $c \in \{1, \ldots, K\}$ denote the correct label assigned to input x. The classical tuple (x, c) is generated from an unknown data distribution $P(x, c)$. We assume x to be discrete-valued to avoid some technicalities, but the analysis can be extended to continuous-valued inputs x.

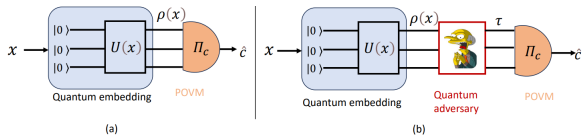

Fig. 1: Quantum machine learning model

The *quantum classifier* consists of a POVM applied to the quantum embedding $\rho(x)$. The POVM $\Pi = \{\Pi_c\}_{c=1}^K$ is defined by positive semi-definite matrices Π_c, for $c = 1, \ldots, K$, that satisfy the equality $\sum_{c=1}^K \Pi_c = I$, where I denotes the identity matrix. We use $\mathcal{M} = \{\Pi : \Pi_c \geq 0, \sum_{c=1}^K \Pi_c = I\}$ to denote the set of all POVMs. We consider as loss function

$$\ell(\Pi, \rho(x), c) = 1 - \text{Tr}(\Pi_c \rho(x)), \tag{1}$$

the probability of misclassifying $\rho(x)$ given its true label c.

In an *adversarial* setting, as illustrated in Fig. 1(b), a *quantum adversary* can perturb the input quantum state $\rho(x)$ with the goal of maximizing the classifier's loss. Specifically, we consider an adversary that perturbs the input state $\rho(x)$ into a state τ that is ϵ-close to the original state $\rho(x)$ in p-Schatten distance, i.e., $D_p(\rho(x), \tau) = \|\rho(x) - \tau\|_p \leq \epsilon$. Targeting a worst-case scenario, the adversary is assumed to know the quantum classifier Π and the loss function (1), resulting in *white-box* attacks. This results in the following *adversarial loss* of the classifier Π on data tuple $(\rho(x), c)$,

$$\ell_{p,\epsilon}(\Pi, \rho(x), c) = \max_{\tau : D_p(\tau, \rho(x)) \leq \epsilon)} \ell(\Pi, \tau, c). \tag{2}$$

In adversarial training, the quantum classifier is assumed to be aware of the presence of the adversary, and is trained by optimizing the *adversarial training risk*,

$$\widehat{L}_{p,\epsilon}(\Pi, \mathcal{T}) = \frac{1}{T} \sum_{n=1}^{T} \ell_{p,\epsilon}(\Pi, \rho(x_n), c_n),$$

the empirical average of the adversarial loss over the training set $\mathcal{T} = \{(x_n, c_n)\}_{n=1}^T$ of i.i.d tuples generated from distribution $P(x, c)$. The goal is to ensure that the adversarially trained quantum classifier incurs minimum *adversarial population risk*, $L_{p,\epsilon}(\Pi) = \mathbb{E}_{P(x,c)}[\ell_{p,\epsilon}(\Pi, \rho(x), c)]$, on new, previously unseen perturbed data. We define the *adversarial generalization error* of the classifier $\Pi \in \mathcal{M}$ as

$$\mathcal{G}_{p,\epsilon}(\Pi, \mathcal{T}) = L_{p,\epsilon}(\Pi) - \hat{L}_{p,\epsilon}(\Pi, \mathcal{T}). \tag{3}$$

Characterizing the Adversarial Generalization Error: To upper bound the adversarial generalization error in (3), we use the classical uniform convergence result from statistical learning theory. This gives that with probability at least $1 - \delta$, for $\delta \in (0, 1)$, over random draws of the dataset, the following inequality holds:

$$\mathcal{G}_{p,\epsilon}(\Pi, \mathcal{T}) \leq 2\mathcal{R}_{p,\epsilon}(\mathcal{M}) + \sqrt{2\log(2/\delta)/T}, \text{ where} \tag{4}$$

$$\mathcal{R}_{p,\epsilon}(\mathcal{M}) = \mathbb{E}_{\mathcal{T}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\Pi \in \mathcal{M}} \frac{1}{T} \sum_{n=1}^{T} \sigma_n \ell_{p,\epsilon}(\Pi, \rho(x_n), c_n) \right]$$

is the *adversarial Rademacher complexity* with $\{\sigma_n\}_{n=1}^T$ denoting i.i.d Rademacher variables.

Assumption: The quantum embedding map $x \mapsto \rho(x)$ is such that the minimum eigenvalue of $\rho(x)$ satisfies $\lambda_{\min}(\rho(x)) \geq \epsilon$.

Result 1: *Adversarial Rademacher complexity of binary classifiers is never smaller than in the non-adversarial setting.* Specifically, we have the following relationship

$$\mathcal{R}(\mathcal{M}) \leq \mathcal{R}_{p,\epsilon}(\mathcal{M}) \leq \mathcal{R}(\mathcal{M}) + \sqrt{2/T}\epsilon d^{1-1/p}. \tag{5}$$

where $\mathcal{R}(\mathcal{M}) = \mathcal{R}_{p,0}(\mathcal{M})$ is the non-adversarial Rademacher complexity and d denotes the dimension of the Hilbert space. Furthermore, contrary to the classical setting, the dimensional dependence of the upper bound (5) is a consequence of the choice of attack. In particular, only for $p = 1$ attack, the bound is independent of the dimension d.

Result 2: *Adversarial generalization error of K-class classifiers scale as* $O(\sqrt{K/T}(\sqrt{2^{I_2(X:Q)}} + \epsilon d^{1-1/p}))$. Specifically, the following inequality holds with probability at least $1 - \delta$,

$$\mathcal{G}_{p,\epsilon}(\Pi, \mathcal{T}) \leq \sqrt{\frac{K}{T} 2^{I_2(X:Q)}} + 2\sqrt{\frac{K}{T}} \epsilon d^{1-1/p} + \sqrt{\frac{2}{T} \log(2/\delta)}$$

where $I_2(X : Q) = \log_2 \left(\text{Tr} \sqrt{\sum_x P(x) \rho(x)^2} \right)^2$ is the 2-Rényi mutual information between the subsystems Q and X of the classical-quantum state $\rho_{CXQ} = \sum_x P(x, c) |cx\rangle\langle cx| \otimes \rho(x)$. The first term corresponds an upper bound on the non-adversarial Rademacher complexity, which is improved by a factor of 2 over previous works.

The first two authors are with the School of Computer Science, University of Birmingham, UK. They can be reached at pxg402@student.bham.ac.uk and s.t.jose@bham.ac.uk. Osvaldo Simeone (osvaldo.simeone@kcl.ac.uk) is with the Department of Engineering, King's College London.