Generating Universal Adversarial Perturbations for Quantum Classifiers

Gautham Anil[1*], Vishnu Vinod[1*] & Apurva Narayan[2,3,4]

[1] Indian Institute of Technology Madras, [2] University of Western Ontario [3] University of British Columbia, [4] University of Waterloo

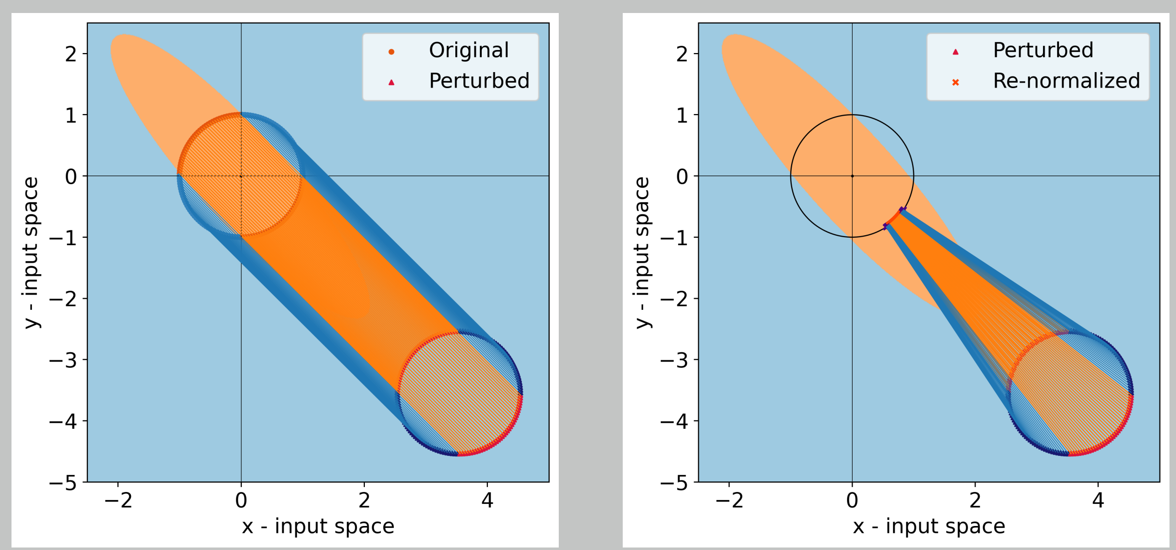
SCAN ME

Introduction

- Parametrized Quantum Circuit (PQC) based quantum classifiers are also vulnerable to adversarial attacks like their classical counterparts.
- **Universal Adversarial Perturbations** (UAPs) - perturbations which when applied to a batch of inputs can cause trained classifiers to misclassify most of the inputs in the batch.
- We introduce **QuGAP**: **Qu**antum **G**enerative **A**dversarial **P**erturbations.
- For classical data (QuGAP - A): Generates additive UAPs bounded by perturbation norm.
- For quantum data (QuGAP - U): Generates unitary UAPs controlled by quantum state fidelity constraints.
- First formulation of additive UAPs for quantum classifiers; We also establish the state-of-the-art for unitary UAPs.

Additive UAPs

- Input-agnostic perturbations on classical data prior to encoding, which cause quantum classifier to misclassify it.
- Using sufficiently large perturbation, we distort the samples such that they get projected onto a desired decision region of the quantum classifier.
- Theoretical and empirical analysis focused on amplitude encoded data.



(a) Displacement due to perturbation (b) Re-normalizing perturbed samples

Figure: Demonstration of our key idea using a simple XOR dataset with an elliptic decision boundary. All samples in the dataset get classified into a single label after perturbation.

Theorem: Existence of Additive UAPs

For an additive universal adversarial perturbation p applied on inputs of classifier $\mathcal{Q}$, a strength of perturbation $\|p\| \in \mathbb{R}$ will cause $\mathcal{Q}$ to classify all inputs as c (class to which $p/\|p\|$ belongs) if:

$$\boxed{\|p\| \geq \frac{2}{\left(\epsilon_c\sqrt{4-\epsilon_c^2}\right)}}$$

where ϵ_c is given by: $\epsilon_c = \sqrt{1 + \frac{1}{2d}\cdot\left(\hat{p}^T M^c \hat{p} - \hat{p}^T M^{c'} \hat{p}\right)} - 1$ *where $\hat{p} = p/\|p\|$ and c' is the class with highest output probability for $\hat{p}$ after c.*

QuGAP - A

- Perturbations are additive transformations applied to the classical data samples.
- Fooling loss to train the generative network for targeted and untargeted UAPs:

$$\boxed{\mathcal{L}_{fool,targeted} = \sum_{x\in\mathcal{D}} \mathcal{L}_{CE}(\hat{y}_x, t)} \qquad \boxed{\mathcal{L}_{fool,untargeted} = -\sum_{x\in\mathcal{D}} \mathcal{L}_{CE}(\hat{y}_x, c_x)}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, $\hat{y}_x$ gives prediction probabilities for x, c_x is the true label of x and t is the target label (for targeted UAPs).

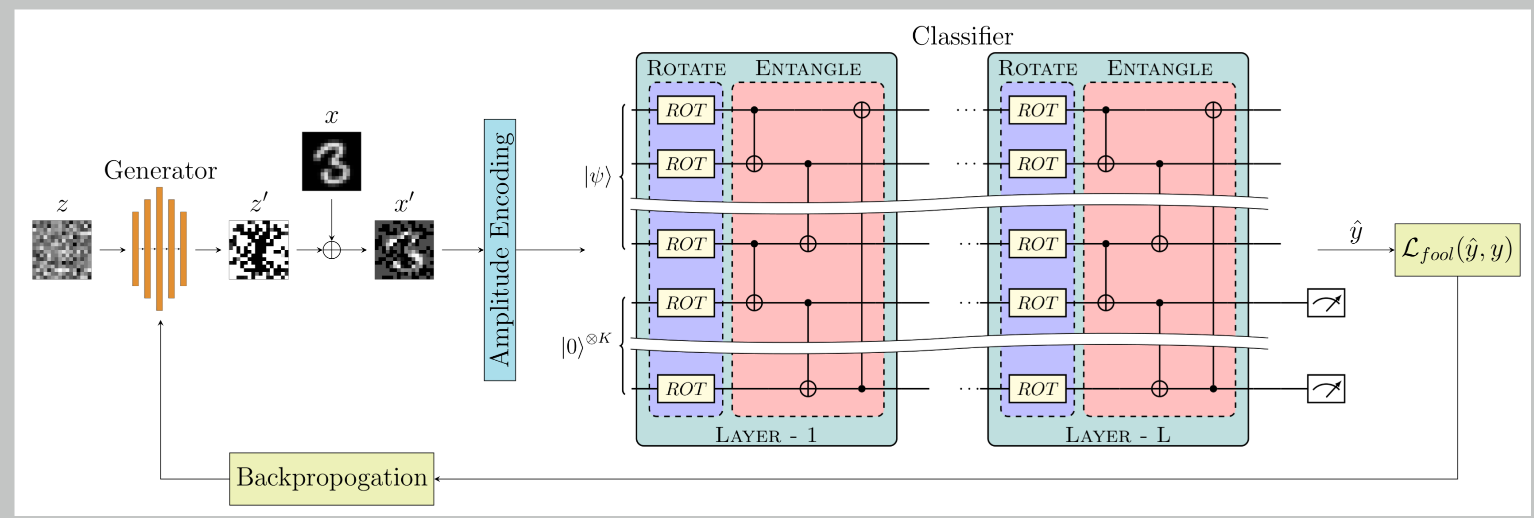


Figure: QuGAP-A: Framework for generating additive UAPs for quantum classifiers. Random vector z sampled from $\mathbb{R}^m$ is passed through a classical generative network. The generated perturbation z' is scaled to impose the norm constraint and added to an input sample x. The perturbed input sample x' is amplitude-encoded and passed through the trained quantum classifier $\mathcal{Q}$. Output predictions from $\mathcal{Q}$ are used to compute the fooling loss $\mathcal{L}_{fool}$. Gradients computed are backpropagated.

QuGAP - U

- Perturbations are unitary transformations applied directly on input quantum states.
- Applicable for both encoded (any encoding scheme) classical data and quantum data.
- A novel fidelity based loss ensures high fidelity between unperturbed and perturbed quantum states:

$$\boxed{\mathcal{L}_U = \mathcal{L}_{fool} + \alpha\mathcal{L}_{fid}}$$

where $\mathcal{L}_{fid} = (1 - \mathcal{F}(|\psi\rangle, |\phi\rangle))^2$; $\mathcal{F}(|\psi\rangle, |\phi\rangle) = |\langle\psi|\phi\rangle|^2$ is the quantum state fidelity.

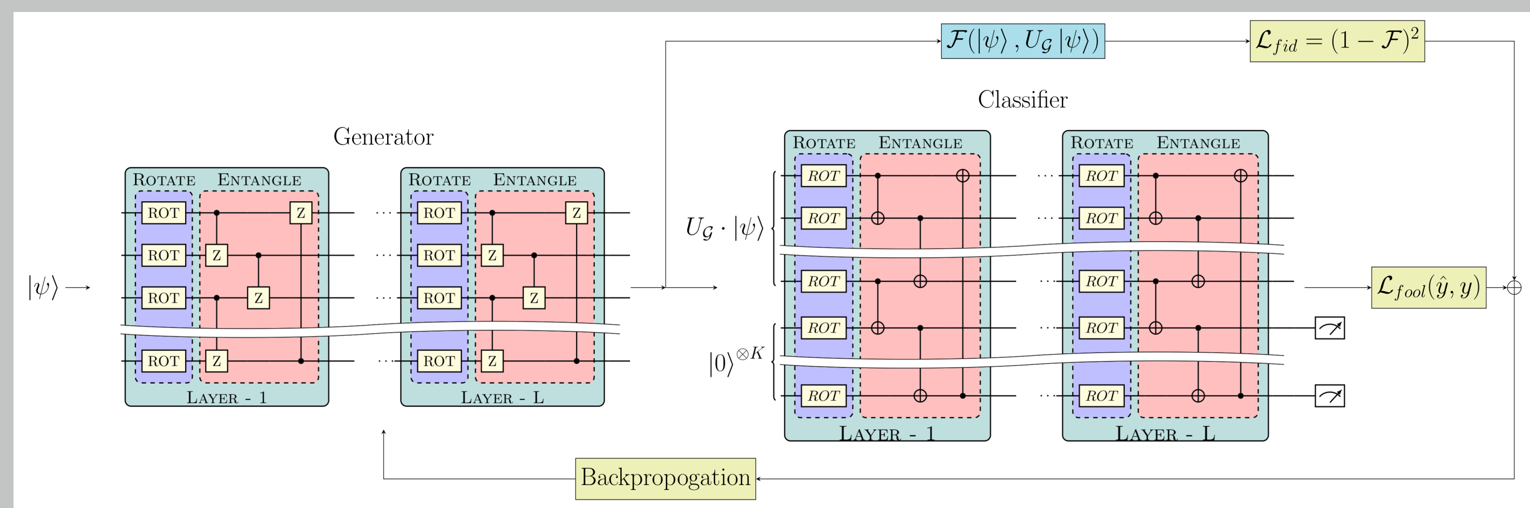


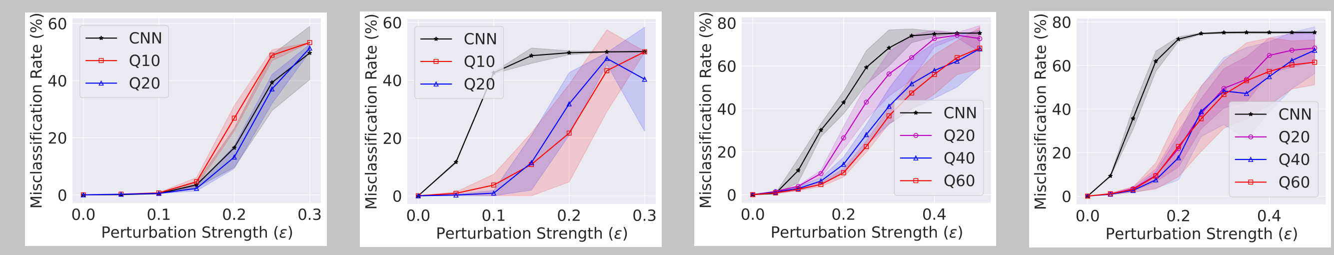
Figure: QuGAP-U: A framework for generating unitary UAPs for quantum classifiers. The quantum generator $\mathcal{G}_Q$ takes in an input state $|\psi_i\rangle$ and transforms it into a perturbed state $|\phi_i\rangle = U_{\mathcal{G}}|\psi_i\rangle$. The fidelity between $|\psi_i\rangle$ and $|\phi_i\rangle$ is computed from which $\mathcal{L}_{fid}$ is calculated. $|\phi_i\rangle$ is also passed through a trained quantum classifier $\mathcal{Q}$ to compute $\mathcal{L}_{fool}$. Gradients are computed using the total loss $\mathcal{L}_{fool} + \alpha\mathcal{L}_{fid}$ and used to update the parameters of $\mathcal{G}_Q$ over all training samples for multiple epochs.

Experimental Results

Misclassification rate: fraction of inputs misclassified by a perturbation.

Additive UAPs

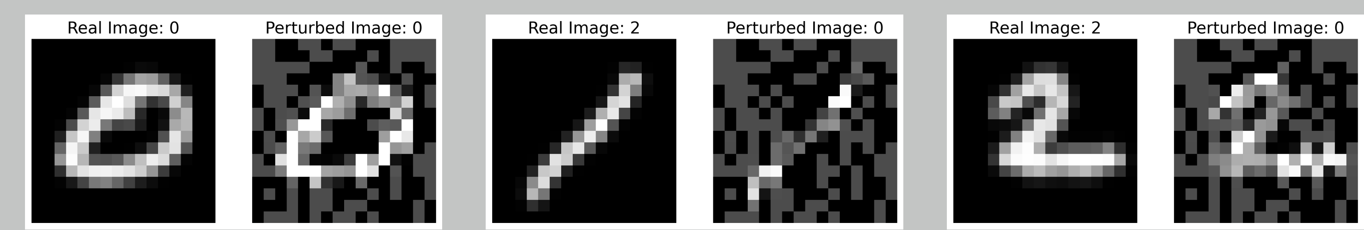
QuGAP-A evaluated on MNIST and FMNIST datasest. Misclassification rates plotted averaged over 10 runs for both binary (classes 0 and 1) and 4-class (classes 0, 1, 2 and 3) classification tasks.



(a) MNIST: 2-class (b) FMNIST: 2-class (c) MNIST: 4-class (d) FMNIST: 4-class

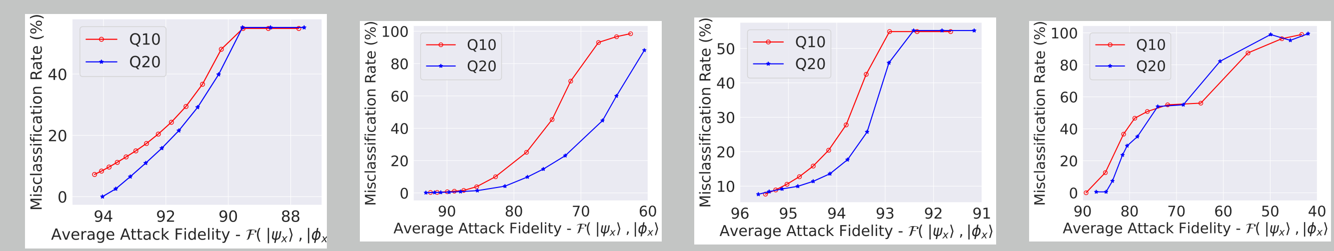
Visualization

Illustration of generated un-targeted UAPs for PQC20 along with predictions for real and perturbed images. The strength of perturbation is $\epsilon = 0.30$.



Unitary UAPs

QuGAP-U evaluated on Transverse Ising Model (TIM) and MNIST datasets. We first generate unitary UAPs with unitary $U_{\mathcal{G}_Q}$ acting as a proxy for the quantum generator $\mathcal{G}_Q$ ((a) & (b)). We then simulate QuGAP-U using a PQC as $\mathcal{G}_Q$ ((c) & (d)). QuGAP-U also outperforms the qBIM attack framework.



(a) TIM: Unitary (b) MNIST: Unitary (c) TIM: PQC Sim (d) MNIST: PQC Sim

Summary

- We theoretically show the existence of additive UAPs.
- We propose a framework (QuGAP-A) to generate additive UAPs.
- We propose state-of-the-art framework (QuGAP-U) to generate unitary UAPs.

Acknowledgements

We thank the Digital Research Alliance of Canada for access to computational resources. Our work was supported by the MITACS Globalink Research Internship award 2022 and by the NSERC Discovery Grant No. RGPIN-2019-05163.

Code Repository: https://github.com/Idsl-group/QuGAP

Supplementary: https://arxiv.org/abs/2402.08648