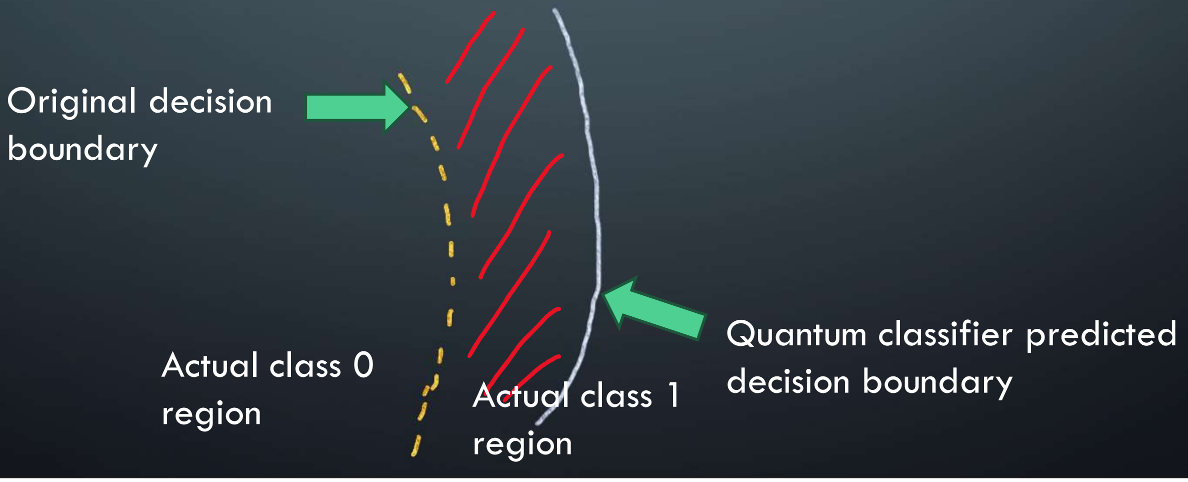
OPTIMIZATION ALGORITHM FOR BLACK BOX METHOD OF ADVERSARIAL ATTACKS

- First pass a set of queries to the oracle function. It would question the model to formulate an approximate decision boundary.
- Lemma 1: Highest probability of getting adversarial examples is near the decision boundary
- Lemma 2: Adversarial examples must lie in the red region



Original decision boundary

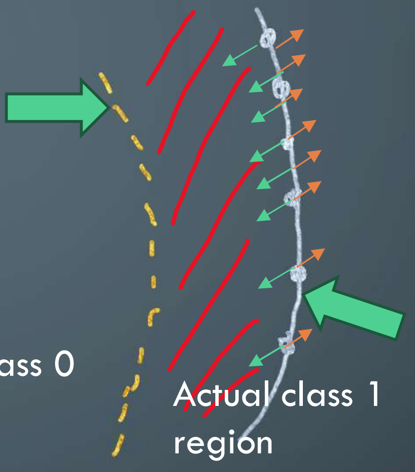
Actual class 0 region

Actual class 1 region

Quantum classifier predicted decision boundary

PROPOSED ALGO 1

- Initialize n points on the decision boundary.
- Move each point both left and right, normal to the surface.
- Identify the direction on which moving creates adversarial examples
- Generate adversarial examples on that direction
- It can be further analyzed that an intersecting decision boundary, also works here with a set of points going towards left or right.



Original decision boundary

Actual class 0 region

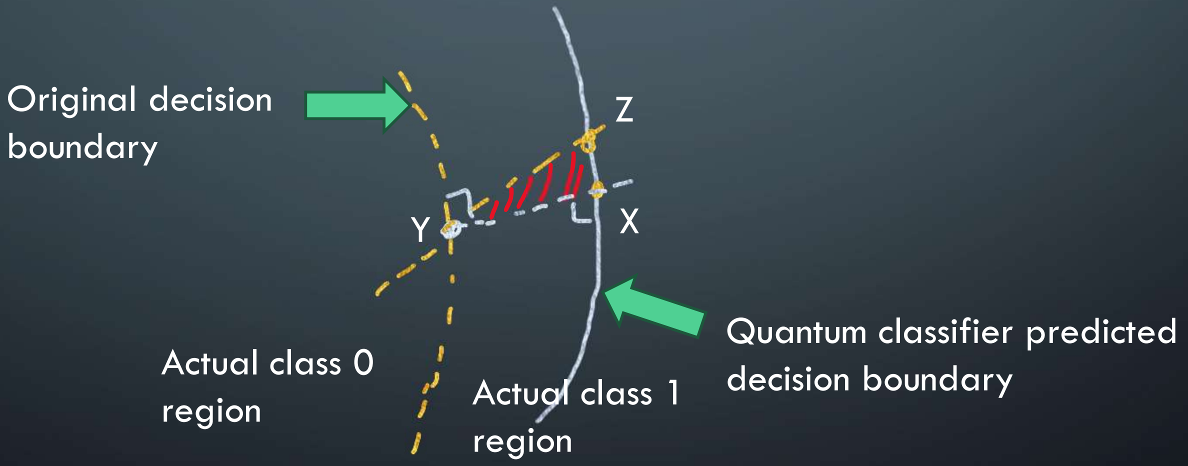
Actual class 1 region

Quantum classifier predicted decision boundary

Try to Devise quantum approaches like Grover-based search to generate Adversarial examples, quadratically faster

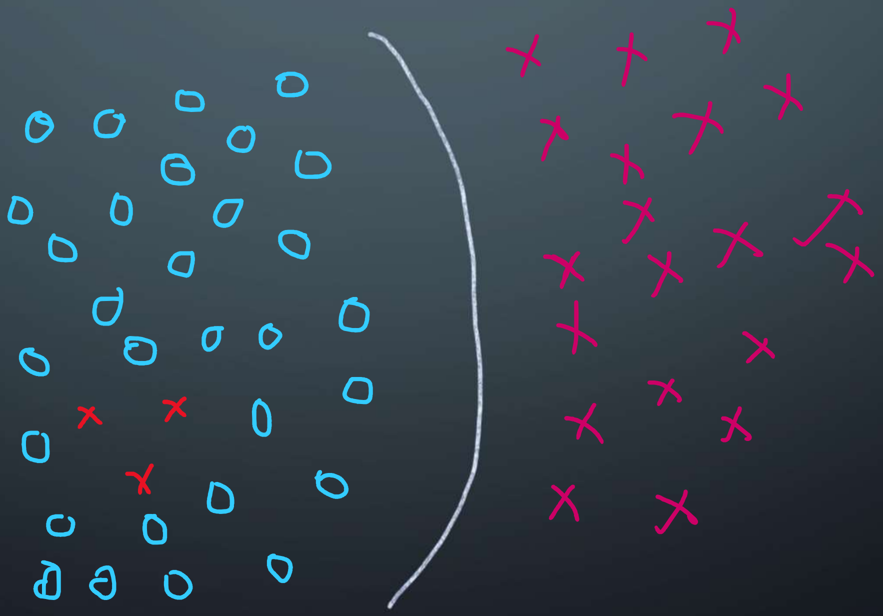
FOLLOW UP ASSUMPTION [NOT GOOD]

- Assumption : we have access to original decision boundary, we can get rid of the trial approaches of determining direction by drawing 2 normals and stating high probability of finding adversarial examples in this triangle.


Original decision boundary
Z
Y
X
Quantum classifier predicted decision boundary
Actual class 0 region
Actual class 1 region

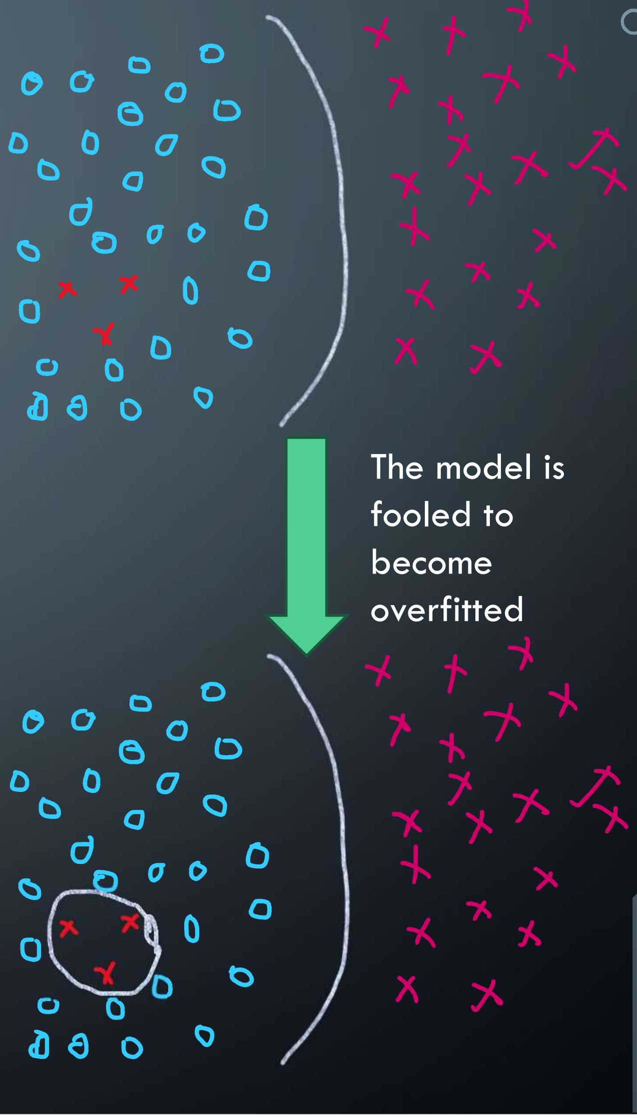
OVERFITTING ATTACK [BY FOOLING] ON THE MODEL IS POSSIBLE [IN CASE THERE ALREADY EXISTS MISCLASSIFIED EXAMPLES WHICH IS] NOT NEAR TO THE DECISION BOUNDARY

- The correctness of a model can be changed in this case



PROPOSED ALGO 2

- Do a fast quantum search [maybe grover-based] to find such misclassified points.
- Generate X > N adversarial examples near that region by providing small perturbation, N = A threshold number of points [we can try determine that mathematically]
- The model is now confused and becomes overfitted by changing its decision boundary
- More feature dependency can be devised in quantum models by adding extra CNOT gates for entanglements



The model is fooled to become overfitted