Benchmarking Data Encoding Methods in Quantum Machine Learning

Orlane Zang[1,2], Grégoire Barrué[2], Tony Quertier[2]

[1] ***LAMFA** - **L**aboratoire **A**miénois de **M**athématiques **F**ondamentales et **A**ppliquées,*
***Université de Picardie Jules Verne**, Amiens, France*

[2] ***Orange Innovation**, Rennes, France*

QUEST-IS'25, *EDF Lab, Paris Saclay, France*
1st to 4th December 2025

Table of contents

Motivations

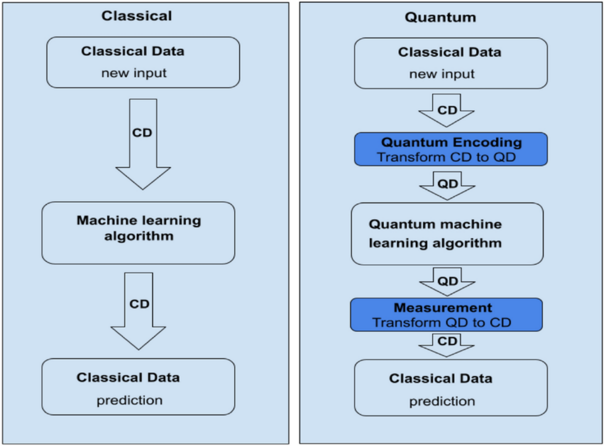
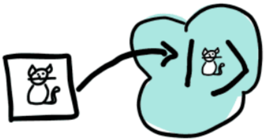


Figure: Classical and Quantum Machine Learning: general workflows

Quantum Data Encoding

The set of methods used to transform classical data into quantum data.

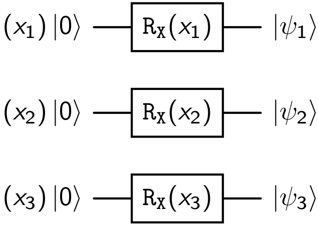


- The way in which data is encoded can significantly influence the results.[1]

[1] M. Schuld et al., "Effect of data encoding on the expressive power of variational quantum-machine-learning models", Physical Review A **103**, 10.1103/physreva.103.032430 (2021), http://dx.doi.org/10.1103/PhysRevA.103.032430.

- **Simple Angle Encoding**[2]

$$(x_i)_{i=1}^{N} \rightarrow |\psi\rangle = \bigotimes_{j=1}^{n} |\psi_j\rangle$$

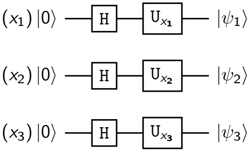


$$\begin{cases} |\psi_j\rangle = \mathrm{R_X}(x_j)\,|0\rangle = \cos\frac{x_j}{2}\,|0\rangle - i\sin\frac{x_j}{2}\,|1\rangle \\ |\psi_j\rangle = \mathrm{R_Y}(x_j)\,|0\rangle = \cos\frac{x_j}{2}\,|0\rangle + \sin\frac{x_j}{2}\,|1\rangle \end{cases} \quad (1)$$

n qubits encode n features

[2] M. Schuld and N. Killoran, "Quantum Machine Learning in Feature Hilbert Spaces", Phys. Rev. Lett. **122**, 040504 (2019), https://link.aps.org/doi/10.1103/PhysRevLett.122.040504.

- $\frac{\pi}{4}$ Angle Encoding[3]



$$\mathtt{U}(x) = \begin{pmatrix} \cos\left(\frac{\pi}{4} - x\right) & \sin\left(\frac{\pi}{4} - x\right) \\ -\sin\left(\frac{\pi}{4} - x\right) & \cos\left(\frac{\pi}{4} - x\right) \end{pmatrix}$$

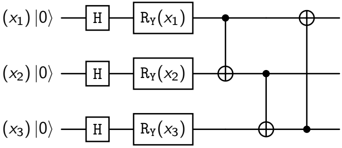
$$\mathtt{H}\,|0\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$\begin{aligned} |\psi_j\rangle &= \mathtt{U}(x)\mathtt{H}\,|0\rangle \\ &= \cos x_j\,|0\rangle + \sin x_j\,|1\rangle \end{aligned} \tag{2}$$

n qubits encode n features

[3] A. Chalumuri et al., "A hybrid classical-quantum approach for multi-class classification", Quantum Information Processing **20**, 119, 119 (2021).

- **Entangled Angle Encoding**[4]



CNOT gate

Input	$\lvert 00\rangle$	$\lvert 01\rangle$	$\lvert 10\rangle$	$\lvert 11\rangle$
Output	$\lvert 00\rangle$	$\lvert 01\rangle$	$\lvert 11\rangle$	$\lvert 10\rangle$

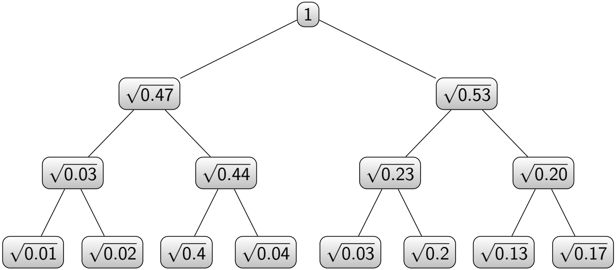
n qubits encode N features, $n = N$

[4] G. D. Luca et al., *Empirical Power of Quantum Encoding Methods for Binary Classification*, 2024, https://arxiv.org/abs/2408.13109.

Amplitude Encoding

$$\sum_{i=1}^{N} |x_i|^2 = 1, \quad |\psi\rangle = \sum_{i=1}^{N} x_i \, |i\rangle \,, \quad N = 2^n \quad \text{with} \quad n \in \mathbb{N}$$

Example[5]**.** $x = \left(\sqrt{0.01}, \sqrt{0.02}, \sqrt{0.4}, \sqrt{0.04}, \sqrt{0.03}, \sqrt{0.2}, \sqrt{0.13}, \sqrt{0.17}\right)$



[5] I. Araujo et al., "A divide-and-conquer algorithm for quantum state preparation", 10.48550/arXiv.2008.01511 (2020).

Amplitude Encoding

From the input x of dimension N, a new vector x_1 of dimension $\frac{N}{2}$ is created such that

$$x^N \to x_1^{\frac{N}{2}} = \left(\sqrt{\left| x[2k]^2 + x[2k+1] \right|^2} \right)_{k=0}^{\frac{N}{2}}.$$

As long as $\dim(x_1)$ is greater than 1, x is assigned x_1 and the procedure is repeated.

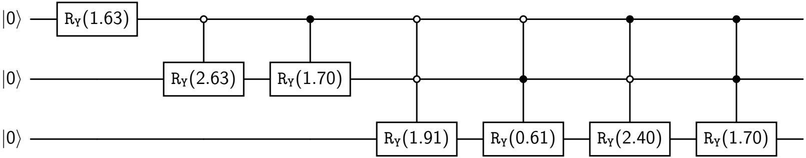
For each new vector x_1, the rotation angles are calculated as follows:

$$\alpha = \begin{cases} 2 \arcsin\left(\frac{x[2k+1]}{x_1[k]} \right), & \text{if} \quad x_1[k] > 0 \\ 2\pi - 2 \arcsin\left(\frac{x[2k+1]}{x_1[k]} \right), & \text{else.} \end{cases} \tag{3}$$

$$x^d = \left(\sqrt{0.01}, \sqrt{0.02}, \sqrt{0.4}, \sqrt{0.04}, \sqrt{0.03}, \sqrt{0.2}, \sqrt{0.13}, \sqrt{0.17}\right)$$

$$\downarrow$$

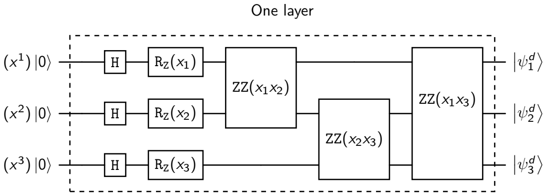
$$(1.63, 2.63, 1.70, 1.91, 0.61, 2.40, 1.70)$$



n qubits encode N data points, such that

$$\begin{cases} 2^{n-1} \leq N \leq 2^n, & \text{if } n = 1, \\ 2^{n-1} < N \leq 2^n, & \text{if } n > 1. \end{cases} \tag{4}$$

Instantaneous Quantum Polynomial (IQP) Encoding



- Originally introduced by[6], these circuits generate probability distributions that are believed to be classically inapproximable.

[6] D. Shepherd and M. J. Bremner, "Temporally unstructured quantum computation", Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **465**, 1413–1439 (2009), http://dx.doi.org/10.1098/rspa.2008.0443.

This method is capable of embedding N data points $(x_1, x_2, \cdots, x_j)$ into $n = N$ qubits such that

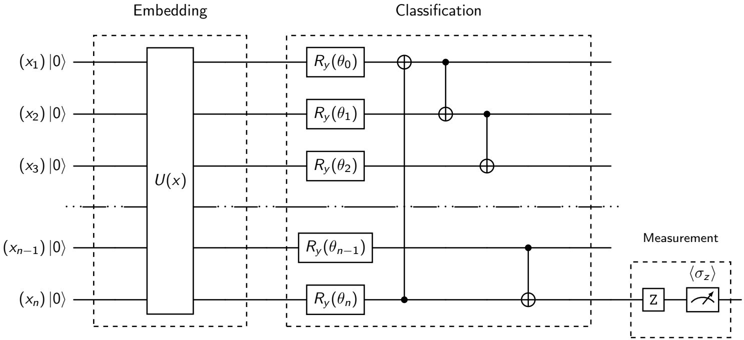
$$|\psi\rangle = \mathtt{U_Z}\ \mathtt{H}^{\otimes n}\ \mathtt{U_Z}\ \mathtt{H}^{\otimes n}\ |0\rangle^{\otimes n}, \tag{5}$$

where

$$\mathtt{U_Z} = \exp\left(\sum_{\substack{j=1\\k=1\\j\neq k}}^{n} -i\frac{x_j x_k}{2}\mathtt{Z}_j\mathtt{Z}_k + \sum_{j=1}^{n} -i\frac{x_j}{2}\mathtt{Z}_j \right), \tag{6}$$

and $\mathtt{Z}_i$ is the Pauli $\mathtt{Z}$ applied to the i^{th}-qubit.

Our model



- Cost function: **cross entropy**

Datasets and preprocessing

- **Wisconsin Diagnostic Breast Cancer (*WDBC*) dataset**: 569 total instances (30 features), which include 357 benign cases and 212 malignant ones.

- **Modified National Institute of Standards and Technology (MNIST) dataset**: 60,000 training images and 10,000 testing images of handwritten digits,
  - MNIST 01,
  - MNIST 08.

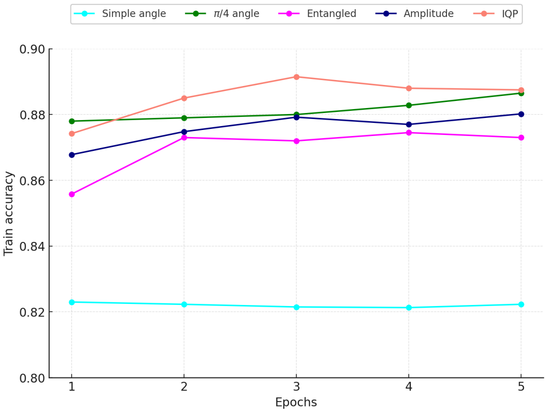
Datasets and preprocessing

- ***Malware dataset***: $163,279$ instances, of which $86,798$ are benign files

Datasets	Malware	MNIST	WDBC
Training	4000	4000	455
Test	2000	1954	114

- (Features, layers): **(NF, ML)**, $\quad N \in \{4,6,8\}$, $M \in \{2,4\}$
- We looked at training accuracy over **5 epochs** with a learning rate $\eta = 0.1$
- We used **Qiskit** simulators from *IBM*

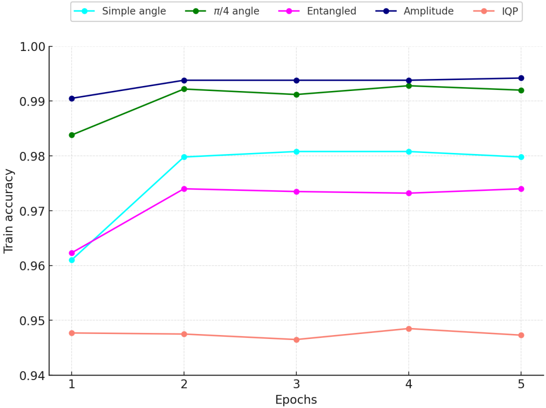
- Malware dataset



Encoding	Test	F1-Score
Simple A.	0.847	0.856
$\frac{\pi}{4}$-Angle	0.900	**0.892**
Entangled A.	0.892	**0.891**
Amplitude	0.883	0.876
IQP	0.893	0.883

Results - For 6 features, 4 layers

- MNIST01 dataset



Encoding	Test	F1-Score
Simple A.	0.987	0.987
$\frac{\pi}{4}$-Angle	0.991	**0.992**
Entangled A.	0.967	0.970
Amplitude	0.995	**0.995**
IQP	0.935	0.943

Conclusion and perspectives

- Encodings such as $\frac{\pi}{4}$-Angle and IQP need few features to return good accuracies
- Entangled Angle and $\frac{\pi}{4}$-Angle Encodings returned the best results compared with the Simple Angle Encoding, while the latter requires little epoch to achieve its best accuracies
- Amplitude Encoding was better on several models on the Malware and MNIST datasets, while on WDBC, it returned the lowest performance
- **Perspective:** interpret these results using Lie theory

THANK YOU!