Filtering out noisy states for quantum neural networks

Albert Akhriev, Anton Dekusar IBM Research Europe – Dublin

In quantum machine learning (QML) noise is a significant challenge due to the fragile nature of quantum states and the complexity of maintaining quantum coherence. It includes decoherence, gate and measurement errors, cross-talk, thermal noise, etc. There are many sophisticated techniques developed to tackle this problem. However, quantum error mitigation does not come for free. It often takes significant amount of processing resources, *Cai et al., 2023*. Recently authors investigated quantum extension of the approach introduced in *Chen et al., 2018*, using Suzuki-Trotter approximation of time-evolution operator. While classical simulations showed encouraging results, the quantum implementation was far less impressive due to a full spectrum of limitations of NISQ devices.

In this study we proceed with the same extension but the goal is to find a way to mitigate quantum noise without incurring significant processing overhead. For simplicity, we consider a parametrized Trotter ansatz of a fixed structure whose outcome is applied to solve a binary classification problem.

Error mitigation strategy

We consider a parametrized ansatz and its operator $A(\Theta)$ acting on the initial state $|0\rangle$, where Θ is a vector of parameters to be learnt, Fig.1. Suzuki-Trotter approximation to the time-evolution operator, *Smith et al., 2019*, is particularly convenient for dynamic circuit adaptation. We use Suzuki-Trotter circuit, $A(\Theta)$, prepended with a layer of local $R_zR_yR_z$ parametrized gates to improve circuit trainability.

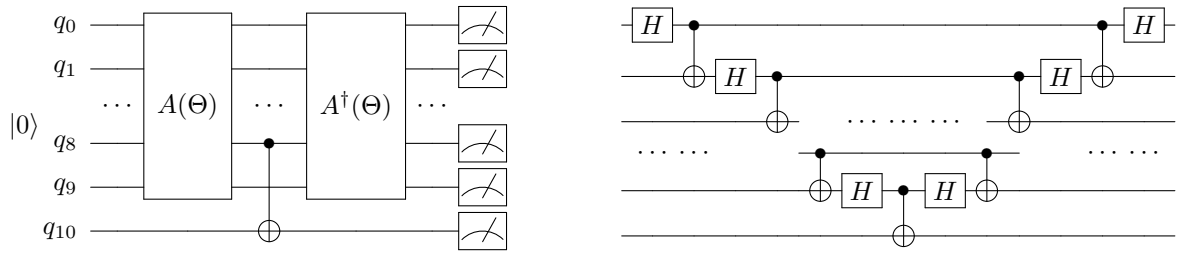

Figure 1: *Left.* The circuit used for binary classification. We extract measurements for qubits q_0 to q_9, and q_{10} separately. If the first 10 qubits are collectively in state $|0\rangle^{\otimes 10}$ we accept the observation at the qubit q_{10} for training, otherwise reject it. *Right.* For the model "stairs" we use a cascade of $CNOT$s interleaved with Hadamard gates instead of the middle $CNOT$, as depicted on the left.

The idea behind this circuit design (left picture on Fig.1) is based on the following claim.
Proposition 1: In the case there is no noise, and irrespective where control qubit of the middle $CNOT$ circuit is attached, the state $|0\rangle^{\otimes 10}$ is observed in the first 10 qubits with probability at least 0.5.

We postpone the proof to supplementary material, noting that it holds true for an arbitrary number of qubits in ansatz. Also, the ansatz does not require extra swap operations on IBM heavy-hex layout.

Since the state $|0\rangle^{\otimes 10}$ is observed in the majority of measurements, we assume that the observation in the last qubit q_{10} can be *trusted* once the state $|0\rangle^{\otimes 10}$ passed through the first qubits unaltered. If any other state has been observed instead of $|0\rangle^{\otimes 10}$, then we reject the measurement at q_{10} as unreliable. By filtering out less probable and potentially corrupted states, we aim to improve resilience of the training process against noise on the present days NISQ devices. As a variant, one can admit a single bit flip.

Dataset and binary classification

Practically, we build 5 similar circuits where ansatz is exactly the same but the control qubit of the middle $CNOT$ gates is connected to every second qubit between $A(\Theta)$ and $A(\Theta)^\dagger$. All the circuits are run independently. We then collect the average observation at the qubit q_{10} of all 5 circuits. Upon each shot, the state $|0\rangle_{q_{10}}$ contributes $+1$ (class **0**) and the state $|1\rangle_{q_{10}}$ contributes -1 (class **1**) to the mean value. Hence, on each iteration, after a number of shots fired (currently 1024) we get 5 scalar observations.

Here are 4 ways to aggregate the observations across different circuits. First, we just take a *mean* value. Second, on every iteration we retrain a small *SVM* classifier and the classifier obtained on the last iteration is used for testing. Third, similarly we train a trivial *linear* classifier. Finally, we use a variant of the circuit with aggregator as shown in right picture on Fig.1 instead of the middle $CNOT$ gate.

Whatever aggregation model is used — "linear", "mean", "svm" or "stairs" — the output of every training sample, which encodes parameters Θ in ansatz $A(\Theta)$, is computed as a mean value observed on qubit q_{10} after 1024 shots. These values are then collected across all the samples and fed into the *hinge-loss* objective function. In addition, we penalize the norm of parameter vector $\|\Theta\|$ in order to keep the accuracy of Suzuki-Trotter approximation within reasonable bounds.

We use the `scikit-learn` Digits dataset, which is a popular resource for practicing classification algorithms and image recognition tasks. It had been designed to facilitate the classification of handwritten digits (0 through 9) and contains $1,797$ 8x8 pixel images. At the beginning of a numerical experiment we pick up a couple of classes for binary classification, e.g., digits 1 vs 7, and map their labels to $\{-1, +1\}$.

A sample image is encoded into 27 parameters of a Trotter layer of `10-qubit` ansatz directly. We pick up 27 overlapping windows of 2×2 size in an 8×8 image. Any i-th Hamiltonian parameter is formed as linear combination of 4 pixel values of i-th window: $\Theta_i = \sum_{k=1}^{4} p_k x_k^{(i)} + p_5$, where $\{x_k^{(i)}\}$ are the pixel values and $\{p_k\}$ is a subset of *optimization* parameters. Parameters in the front layer of $R_z R_y R_z$ gates are *independent* on pixel values. We call them "free" parameters for initial state preparation.

Numerical experiments and discussion

We conducted two series of experiments using noiseless and noisy sampler implementations from the Qiskit framework. The latter one employs the default noise model for IBM Osaka quantum device. One of the most difficult binary classification cases — digits "1" vs "7" — was selected for presentation. We randomly picked up 80 training and 20 testing images. The small size of dataset was primarily dictated by the simulation time, which is quite long even for noiseless sampler (up to 6 hours on a single CPU).

In the case of noiseless simulation we can afford to run multiple numerical experiments with unique and random initial guesses. Fig. 2 and Table 1 provide more details. The noisy simulator is significantly (up to two orders) slower than the noiseless one, so we run only one experiment per aggregation model. Fig. 3 and Table 1 show results for every aggregation model obtained through the optimization with unique and random initial guesses. Fig. 4 gives some insight. In all experiments, `Nevergrad` gradient-free optimizer from Facebook (c) has been used with the maximum number of iterations set to 300.

As a baseline we trained a simple data re-uploading quantum model. This model is composed of alternating parameterized rotation gates and two-qubit entanglement units arranged in a checkerboard pattern. The rotation gates are defined by angles given by $a_i x_i + b_i$, where a_i and b_i are trainable parameters, x_i represents one pixel of data, and i denotes the index of the rotation gate. Each entanglement unit includes two trainable rotation gates and one $CNOT$ gate. Under the noiseless setup the model performed moderately accurate, but worse than "linear" and "svm" models, refer to Table 1 for the actual figures.

The goal of this study was to develop a practical solution for a typical machine learning task using a NISQ device (or its software analog), rather than creating a new mathematical algorithm. We proposed a binary classifier that is inherently resistant to the inevitable quantum noise present in current systems. Our experiments on a hardware simulator demonstrated promising results. We believe that combining quantum and classical approaches could be the way forward for the next generation of QML methods.

We envision a number of issues to be addressed in the future research. First of all the scalability. It would be interesting to extend the ansatz to the size when it can adopt the full-size `MNIST` 28×28 images. This will definitely decrease the chance to observe the state $|0\rangle$ in the first qubits. Second, for larger images decreasing the number of parameters would be crucial. Third, reducing the bias towards state $|0\rangle$ in observations on the last qubit can be important. Fourth, the dataset we are currently using is tiny due to computational limitations. One potential workaround would be a batch-based training.

Configuration	Linear	Mean	Stairs	SVM	Baseline
Train, noisy, single experiment	1.0	0.5	0.63	0.94	expecting ...
Test, noisy, single experiment	1.0	0.5	0.55	1.0	expecting ...
Train, noiseless, average	0.99	0.59	0.94	0.99	0.93
Test, noiseless, average	0.93	0.59	0.92	0.98	0.85

Table 1: Train and test scores from the experiments. One should take the high scores cautiously due to the small dataset size. Additionally, note that a strong noise is not necessary detrimental to final scores because it sometimes helps to avoid shallow local minima during optimization.

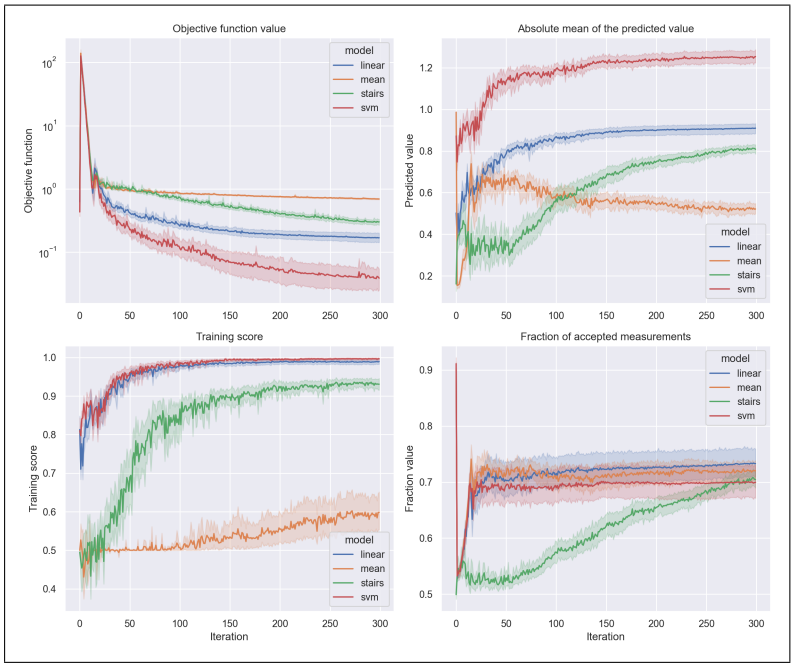

Figure 2: Digits "1" vs "7" classification. Here we expose the convergence profiles of all aggregation models collected across 22 independent simulations *without noise*. Each simulation started from a unique and random initial parameters. While the models "svm" and "linear" attain a low value of objective function, the other two — "stairs" and "mean" — do not progress well enough, and the "mean" one performs quite poorly. It can also be seen on the right-bottom panel that the fraction of accepted measurements exceeds 0.5 as predicted by Proposition 1.

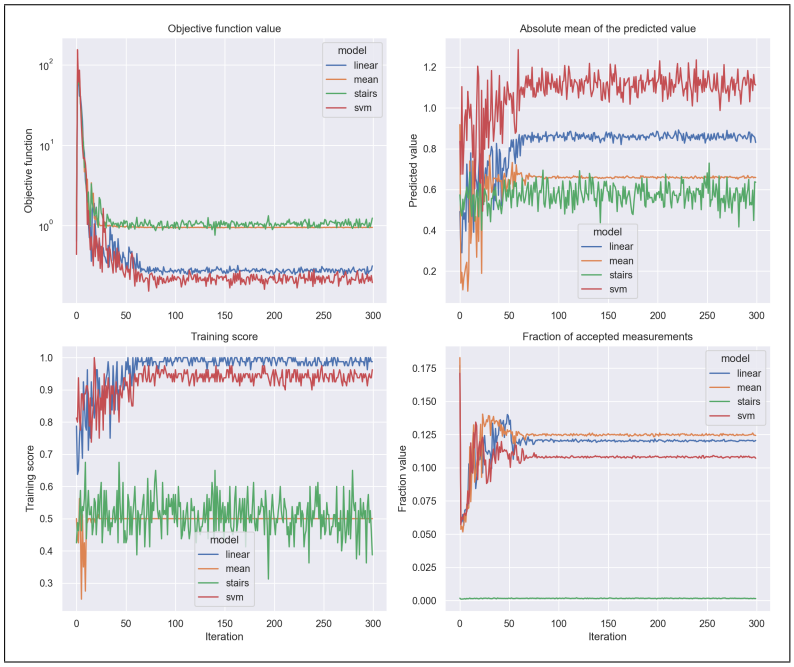

Figure 3: Digits "1" vs "7" classification. Here we demonstrate the convergence profiles of all aggregation models obtained from *noisy* sampler using the default noise model for IBM Osaka quantum device. We run a single simulation per aggregation model because every numerical experiment is very computationally expensive. There are two things to note. First, the perfomance of "stairs" and "mean" models is close to random guess (left-bottom panel). Second, the measurement acceptance rate is well below 0.5, which is somewhat expectable considering a strong disturbance generated by noisy simulator.

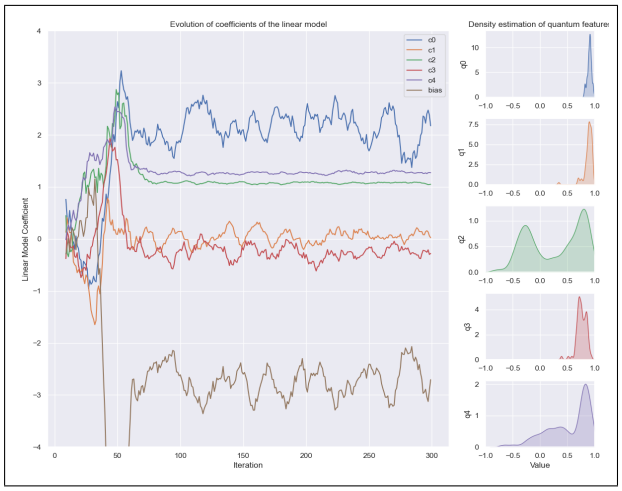

Figure 4: Left panel demonstrates how the coefficient of the linear model evolve over optimization iterations. The measurements have been obtained from *noisy* sampler. The profiles have been smoothed for better visibility. Right panel shows the distribution of corresponding predictions ("features") drawn from all 5 circuits. The circuits differ by the placement of the control qubit of the middle $CNOT$ gate. The 3nd and the 5th features (indices 2 and 4 respectively) are the most discriminative. They occupy more or less the entire interval $[-1 \ldots +1]$ and their complementary coefficients of the linear model are the least volatile.