Understanding Generalization in Quantum Machine Learning with Margins

Tak Hur [1] **Daniel K. Park** [1 2]

1. Introduction

Quantum machine learning (QML) stands out as an innovative application of quantum computation. The success of QML algorithm does not solely depend on how well the model fits the training data but, more importantly, on their ability to accurately predict the outcomes of previously unseen data. This crucial capability, known as generalization, has been extensively explored and analyzed through the lens of statistical learning theory. However, recent studies have highlighted the limitations of current understandings of generalization based on uniform bounds in both classical and quantum machine learning frameworks (Zhang et al., 2021; Gil-Fuster et al., 2024). In this work, we propose a *complexity measures* based on margin distribution, which can accurately capture the generalization performance of QML models.

2. Rethinking Generalization

Suppose there is an unknown joint probability distribution $\mathcal{D}$ governing the quantum state ρ and its corresponding label y. In this section, for simplicity, we will consider $\rho \in \mathbb{C}^{2^n \times 2^n}$ and $y \in \{-1, +1\}$, namely n-qubit binary classification task. With m independent and identically distributed (i.i.d) samples $S = \{\rho_i, y_i\}_{i=1}^m$, the goal is to find a hypothesis h^* with small true error $R(h^*) = \mathbb{E}_{\rho,y\sim\mathcal{D}}[\mathbb{1}_{\mathrm{sgn}(h^*(\rho))\neq y}]$. Since the true distribution $\mathcal{D}$ is unknown, we alternatively find h (from a hypothesis class $\mathcal{H}$) with small empirical risk, $\hat{R}_S(h) = 1/|S| \sum_{\rho,y\in S} \mathbb{1}_{\mathrm{sgn}(h(\rho))\neq y}$. For a hypothesis h, we define a generalization gap as a difference between true and empirical risk, $g(h) = R(h) - \hat{R}_S(h)$. A common way to understand generalization is to upper bound $g(h)$ by a *complexity measure* of the hypothesis class $\mathcal{H}$. For example, the hypothesis class of Quantum Neural Networks (QNN) with parameterized quantum circuit $U(\theta)$ and observable O can be expressed as $\mathcal{H}_{\mathrm{QNN}} = \{\rho \mapsto \mathrm{Tr}(OU(\theta)\rho U^\dagger(\theta)) : \theta \in \Theta\}$.

Theorem 2.1 (Rademacher Complexity Bound)**.** *For any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size m drawn according to $\mathcal{D}$, following holds for any $h \in \mathcal{H}_{\mathrm{QNN}}$*

$$R(h) \leq \hat{R}_S(h) + \hat{\mathfrak{R}}_S(\mathrm{sgn} \circ \mathcal{H}_{\mathrm{QNN}}) + 3\sqrt{\frac{\log(2/\delta)}{2m}}. \quad (1)$$

Here, $\mathrm{sgn} \circ \mathcal{H}_{\mathrm{QNN}} = \{\rho \mapsto \mathrm{sgn}(h(\rho)) : h \in \mathcal{H}_{\mathrm{QNN}}\}$, and $\hat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma[\sup_{h\in\mathcal{H}} \frac{1}{m} \sum_i \sigma_i h(\rho_i)]$, where σ_i are i.i.d Rademacher random variables that takes value ± 1 with equal probability 1/2.

Although Theorem 2.1 provides rigorous theoretical guarantee for generalization, it can result in vacuous upper bound, especially when $\mathcal{H}_{\mathrm{QNN}}$ is extensive enough to overfit random labels. For example, consider a corrupted sample $\tilde{S} = \{\rho_i, \tilde{y}_i\}_{i=1}^m$, where each $\tilde{y}_i$ are independently assigned ± 1 with a probability 1/2, irrespective of the data ρ_i. Suppose $\mathcal{H}_{\mathrm{QNN}}$ can overfit the corrupted sample $\tilde{S}$, i.e. $\exists h \in \mathcal{H}_{\mathrm{QNN}}$ s.t. $\hat{R}_{\tilde{S}}(h) \approx 0$. Since the true error with respect to the corrupted distribution is 0.5 for all h, the analysis indicates that $0.5 \lesssim \hat{\mathfrak{R}}_S(\mathrm{sgn} \circ \mathcal{H}) + 3\sqrt{\log(2/\delta)/2m}$. Consequently, the Rademacher complexity bound $g(h) \lesssim 0.5$ is uninformative for binary classification.

Zhang et al. (2021) highlighted that modern (classical) machine learning models, due to their large size and extensive numbers of parameters, can overfit random labels, suggesting our understanding of generalization is incomplete. Similarly, Gil-Fuster et al. (2024) demonstrated that Quantum Convolutional Neural Networks (QCNNs) can also overfit random labels in the Quantum Phase Recognition problem, indicating this issue extends to quantum machine learning. It is important to note that this problem is not restricted to Rademacher Complexity bound, but any uniform generalization bounds, including the results from Caro et al. (2021; 2022; 2023); Bu et al. (2021; 2022; 2023).

3. Margin based Generalization in Quantum Machine Learning

The concept of margin has been extensively explored since the early days of machine learning, offering theoretical foundations for Support Vector Machines (Cortes & Vapnik, 1995). The margin quantifies the difference between the output for correct labels and incorrect labels. More specifically, in k-class classification, for a data point (x, y), where $x \in \mathcal{X}$ and $y \in [k]$, and a classifier $f : \mathcal{X} \mapsto \mathbb{R}^k$, margin is defined as $f(x)_y - \max_{j\neq y} f(x)_j$. Here, the k-dimensional vector output of the classifier corresponds to the probability of assigning x to each class. Recently, Bartlett et al. (2017) proposed a generalization bound based on margins, normalized by a spectral norm of the weights, in the context of deep neural networks. It illustrated that complexity

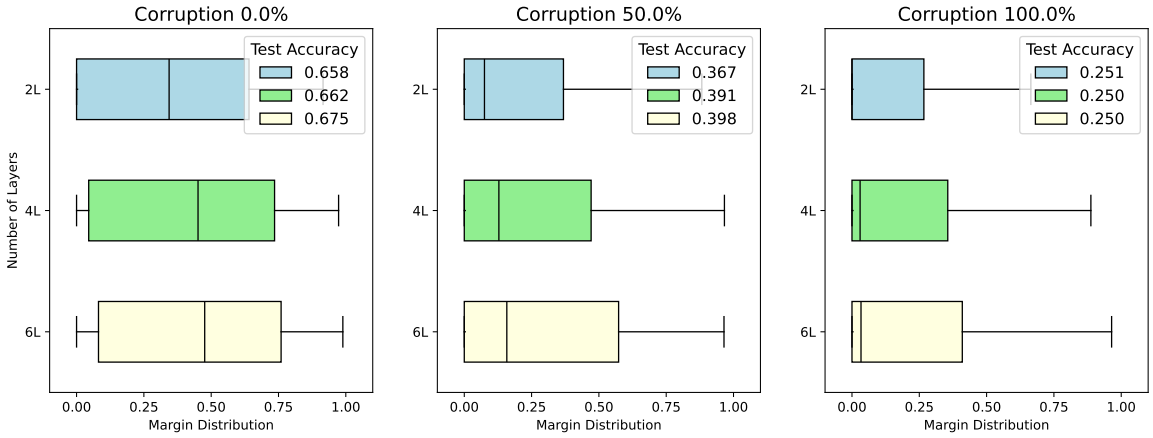

Figure 1. A (Tukey) box-and-whisker plot depicting the margin distributions of optimized four qubit Quantum Convolutional Neural Networks (QCNNs). The results for QCNNs with two, four, and six layers are displayed, along with their corresponding test accuracies. QCNNs were trained for 4-class classification task aimed at quantum phase recognition (QPR). The experiment was performed with varying degrees of label noise: QPR dataset with pure labels (left), half randomly labelled dataset (middle), and full randomly labelled datasets (right). As the noise (corruption) level increases, the margin distributions tend to exhibit a more pronounced skew towards the left, indicating that a greater proportion of samples are classified with smaller margins. Notably, the margin distribution exhibits a strong positive correlation with test accuracy across all scenarios.

measures based on margin can address the shortcomings of uniform generalization bounds, as will be explained in more detail later in this section. Furthermore, it empirically demonstrated a significant correlation between margin-based measures and generalization error. Since then, margin is extensively used as a tool to understand generalization in (classical) machine learning (Neyshabur et al., 2017; Jiang et al., 2018; Neyshabur et al., 2018; Farhang et al., 2022; Jiang et al., 2020).

The notion of margin can be extended to understand generalization performances of quantum machine learning models. Consider a k-class classification employing quantum neural networks, where the hypothesis class is defined as $\mathcal{H}_{\mathrm{QNN}} = \{\rho \mapsto [\mathrm{Tr}(\mathcal{M}_i U(\theta)\rho U^\dagger(\theta))]_{i=1}^k : \theta \in \Theta\}$. Here, measurement outcome of $\mathcal{M}_i$ represents the probability of assigning ρ to label i.

Theorem 3.1 (Margin Bound for Quantum Neural Networks)**.** *For any $\delta > 0$ and $\gamma > 0$, with probability at least $1-\delta$ over a sample S of size m drawn according to $\mathcal{D}$, following holds for any $h \in \mathcal{H}_{\mathrm{QNN}}$,*

$$R(h) \leq \hat{R}_\gamma(h) + \frac{2}{\gamma}\hat{\mathfrak{R}}_S(\mathcal{H}_{\mathrm{QNN}}) + 3\sqrt{\frac{\log(2/\delta)}{2m}}. \quad (2)$$

Here, $\hat{R}_\gamma(h)$ represents the empirical margin error, quantifying the number of samples whose classification margin falls below the threshold γ. Formally, it is defined as $\hat{R}_\gamma(h) = 1/|S| \sum_{\rho,y\in S} \mathbb{1}_{h(\rho)_y \leq \max_{j\neq y} h(\rho)_j + \gamma}$. The upper bound described in Equation 2 comprises of two competing terms: selecting a larger γ increases $\hat{R}_\gamma(h)$, while simultaneously decreasing $2\hat{\mathfrak{R}}_S(\mathcal{H}_{\mathrm{QNN}})/\gamma$. According to Theorem 3.1, a hypothesis that classifies S with large margins results in a tighter upper bound, as opting for a larger γ does not significantly increase $\hat{R}_\gamma(h)$. Thus the *margin distribution*, which is the distribution of margins of sample S, plays a crucial role in comprehending the generalization performance of QML models.

Unlike uniform generalization bounds, margin bound provides distinct results depending on the distribution of the data. For instance, if we corrupt the sample from S to $\tilde{S}$ (and the data distribution from $\mathcal{D}$ to $\tilde{\mathcal{D}}$) as outlined in Section 2, the margin distribution will also vary, leading to a different generalization upper bound. If the margin distribution skews toward left as the data are corrupted, the margin bounds correctly explains the increasing generalization gap, a subtlety that uniform generalization bounds fail to capture.

Remark 3.2. It is noteworthy that we can further upper bound the $\hat{\mathfrak{R}}_S(\mathcal{H}_{\mathrm{QNN}})$ and achieve more interpretable results. For instance, Ref (Bu et al., 2021; 2022; 2023) analyze the Rademacher complexity of QNN through the lens of quantum resource theory. Additionally, Ref (Caro et al., 2022) quantifies the covering number of QNN based on the numbers of parameters. This result, combined with Dudley's entropy integral, can be utilized to establish the upper bound of Rademacher complexity (Vershynin, 2018). However, in this study, our primary focus lies on exploring the margin distributions of quantum machine learning models and how margin-based complexity measures strongly correlate with generalization gap.

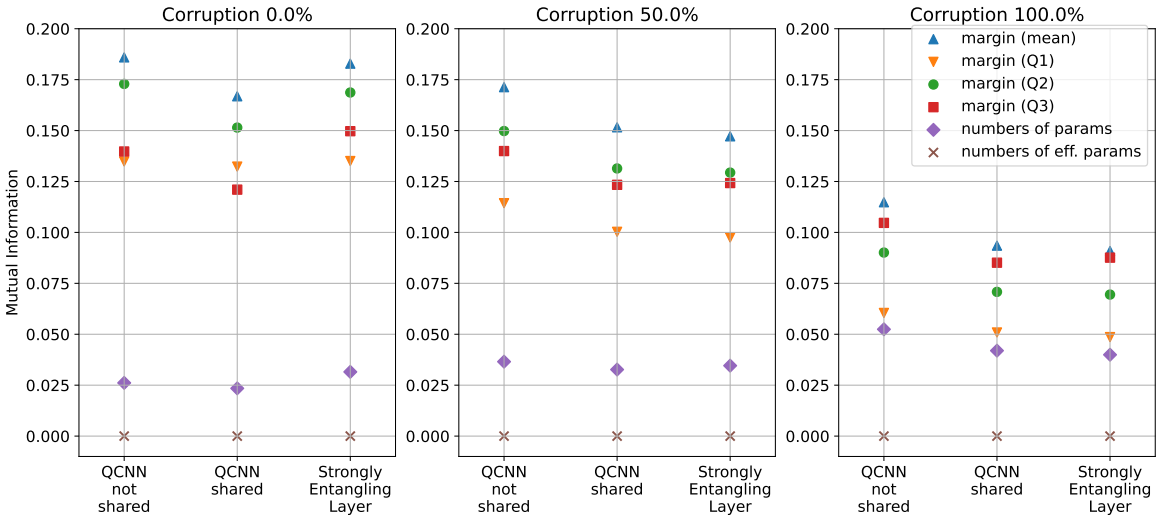

Figure 2. A comparative analysis demonstrating mutual information between generalization gap and various complexity measures. This includes four margin-based complexity measures: mean, lower quartile (Q1), median (Q2), and upper quartile (Q3), along with two parameter-based complexity measures: number of parameters and number of effective parameters. The experiments were conducted using three distinct variational ansatz: 1) QCNN without parameter sharing, 2) QCNN with parameter sharing 3) `StronglyEntanglingLayer` (see Bergholm et al. (2020) for details). Furthermore, the experiments were repeated under label corruption as outlined in Figure 1. In all scenarios, margin-based complexity measure exhibited more mutual information about the generalization gap compared to parameter-based complexity measures. Notably, the mutual information tends to decrease with higher levels of corruption.

4. Experimental Results

This section experimentally demonstrate strong correlation between margins and generalization performances of QML models. We conducted extensive tests on Quantum Neural Networks (QNNs) with various hyperparameters, including circuit architecture, variational ansatz, number of layers, number of training samples, training batch size, maximum training iteration. The models were trained to perform the Quantum Phase Recognition (QPR) task, which involves classifying the phases of the ground state of the generalized cluster Hamiltonian, defined as $H = \sum_{j=1}^{n}(Z_j - J_1 X_j X_{j+1} - J_2 X_{j-1} Z_j X_{j+1})$ (see Caro et al. (2022); Gil-Fuster et al. (2024) for details). Additionally, the experiments were conducted under different levels of label noise: pure labels (r=0.0), half random labels (r=0.5), and fully random labels (r=1.0).

Figure 1 illustrates margin distributions of the optimized QCNNs in a box-and-whisker plot, alongside their respective test accuracies, conducted with varying numbers of QCNN layers. Across all layer configurations, the test accuracy decreases (and consequently, the generalization gap increases) as the labels are randomly corrupted with increasing levels of noise. The margin distributions exhibit significant leftward skew as the labels are corrupted. Thus, the margin bounds (Equation 2) correctly captures the generalization behavior under label corruption. Moreover, QCNNs with a larger number of layers tend to have higher test accuracy and exhibit right-skewed margin distributions, which further validates that margin distribution effectively captures the generalization performance in QML.

In Figure 2, we compare four margin-based complexity measures—mean, lower quartile, median, and upper quartile of the margin distribution—against parameter-based complexity measures. The latter includes 1) the number of parameters and 2) the number of effective parameters, which underwent significant changes during the optimization process. We evaluated mutual information between generalization gap and various complexity measures, treating them as random variables depending on sample S and hyperparameters of the models. Intuitively, a larger mutual information value indicates that the complexity measure contains more information about the generalization gap, thereby reducing uncertainty about generalization given the complexity measure. The experiments were conducted with three distinct variational ansatz: 1) QCNN without parameter sharing (Grant et al., 2018), 2) QCNN with parameter sharing (Cong et al., 2019; Hur et al., 2022), and 3) `StronglyEntanglingLayers` (Bergholm et al., 2020). Across all models, the mutual information values with margin-based complexity measures are significantly larger than those with parameters-based counterparts, indicating that margin distribution is more effective tool for understanding the generalization performance of QML models.

References

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Bergholm, V., Izaac, J., Schuld, M., Gogolin, C., Alam, M. S., Ahmed, S., Arrazola, J. M., Blank, C., Delgado, A., Jahangiri, S., McKiernan, K., Meyer, J. J., Niu, Z., Száva, A., and Killoran, N. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, 2020.

Bu, K., Koh, D. E., Li, L., Luo, Q., and Zhang, Y. Rademacher complexity of noisy quantum circuits. *arXiv preprint arXiv:2103.03139*, 2021.

Bu, K., Koh, D. E., Li, L., Luo, Q., and Zhang, Y. Statistical complexity of quantum circuits. *Physical Review A*, 105 (6):062431, 2022.

Bu, K., Koh, D. E., Li, L., Luo, Q., and Zhang, Y. Effects of quantum resources and noise on the statistical complexity of quantum circuits. *Quantum Science and Technology*, 8 (2):025013, 2023.

Caro, M. C., Gil-Fuster, E., Meyer, J. J., Eisert, J., and Sweke, R. Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum*, 5:582, 2021.

Caro, M. C., Huang, H.-Y., Cerezo, M., Sharma, K., Sornborger, A., Cincio, L., and Coles, P. J. Generalization in quantum machine learning from few training data. *Nature communications*, 13(1):4919, 2022.

Caro, M. C., Huang, H.-Y., Ezzell, N., Gibbs, J., Sornborger, A. T., Cincio, L., Coles, P. J., and Holmes, Z. Out-of-distribution generalization for learning quantum dynamics. *Nature Communications*, 14(1):3751, 2023.

Cong, I., Choi, S., and Lukin, M. D. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, December 2019. ISSN 1745-2473, 1745-2481. doi: 10.1038/s41567-019-0648-8. URL http://www.nature.com/articles/s41567-019-0648-8.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Farhang, A. R., Bernstein, J. D., Tirumala, K., Liu, Y., and Yue, Y. Investigating generalization by controlling normalized margin. In *International Conference on Machine Learning*, pp. 6324–6336. PMLR, 2022.

Gil-Fuster, E., Eisert, J., and Bravo-Prieto, C. Understanding quantum machine learning also requires rethinking generalization. *Nature Communications*, 15(1):1–12, 2024.

Grant, E., Benedetti, M., Cao, S., Hallam, A., Lockhart, J., Stojevic, V., Green, A. G., and Severini, S. Hierarchical quantum classifiers. *npj Quantum Information*, 4(1): 65, December 2018. ISSN 2056-6387. doi: 10.1038/s41534-018-0116-9. URL http://www.nature.com/articles/s41534-018-0116-9.

Hur, T., Kim, L., and Park, D. K. Quantum convolutional neural network for classical data classification. *Quantum Machine Intelligence*, 4(1):3, 2022.

Jiang, Y., Krishnan, D., Mobahi, H., and Bengio, S. Predicting the generalization gap in deep networks with margin distributions. *arXiv preprint arXiv:1810.00113*, 2018.

Jiang, Y., Foret, P., Yak, S., Roy, D. M., Mobahi, H., Dziugaite, G. K., Bengio, S., Gunasekar, S., Guyon, I., and Neyshabur, B. Neurips 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*, 2020.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.