Study Notes on Statistical Learning Theory for Quantum Machine Learning

Kensuke Kamisoyama

(Dated: 2024/11)

Contents

I. References

This note is based on the following references: Overview [1–4], Covering number [5–7], Rademacher complexity [5, 8, 9].

II. Overview

A. Errors in Classical Machine Learning

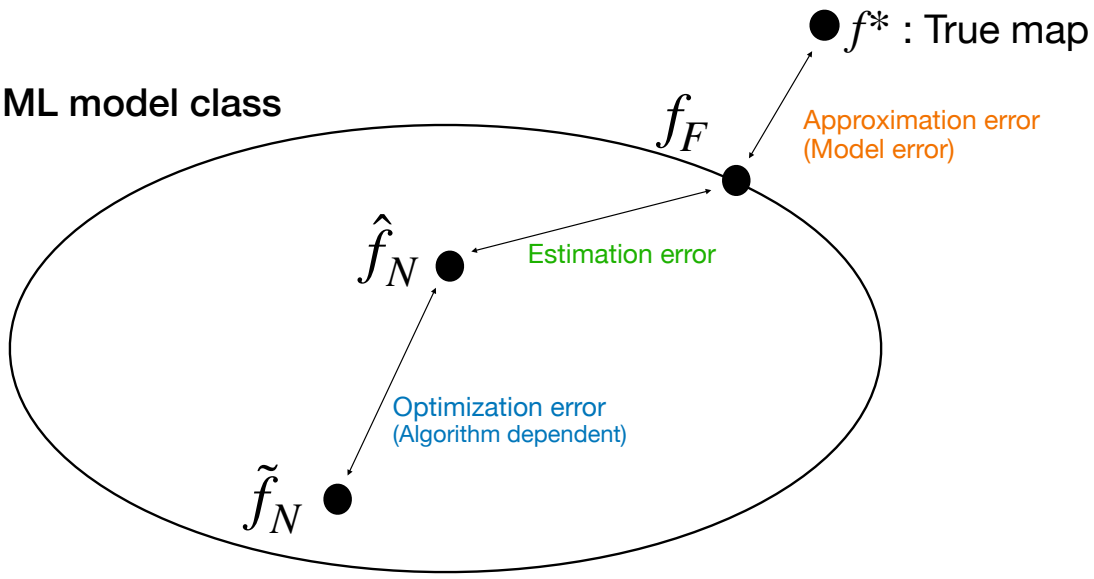

FIG. 1. Approximation error, Estimation error, Optimization error.

- $R(f) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathcal{L}(y, f_{\boldsymbol{\theta}}(x))$: Expected risk of a model f.
- $\hat{R}_N(f) := \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y_i, f_{\boldsymbol{\theta}}(x_i))$: Empirical risk of a model f.
- $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$: A hypothesis class, or a set of functions that a machine learning model can take.
- $\tilde{f}_N$: Model that we get after training on N samples using a learning algorithm.
- $\hat{f}_N := \arg\min_{f\in\mathcal{F}} \hat{R}_N(f)$: Model that minimizes the empirical risk over the hypothesis class $\mathcal{F}$.
- $f_{\mathcal{F}} := \arg\min_{f\in\mathcal{F}} R(f)$: Model that minimizes the expected risk over the hypothesis class $\mathcal{F}$.
- $f^* := \arg\min_{f\in\mathcal{Y}^{\mathcal{X}}} R(f)$: Model that minimizes the expected risk over all measurable functions.

$$\underbrace{R(\tilde{f}_N) - R(f^*)}_{\text{Exess risk}} = \underbrace{R(\tilde{f}_N) - R(\hat{f}_N)}_{\text{Optimization error}} + \underbrace{R(\hat{f}_N) - R(f_{\mathcal{F}})}_{\text{Estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{Approximation error}} \tag{1}$$

Remark 1.

- While Approximation and Optimization errors are important, they are relatively better managed through advances in model architecture, size, and training techniques in modern classical machine learning. Estimation error remains a significant challenge due to the inherent limitations of available training data.
- The relationship between the Approximation error and the Estimation error is similar to the Bias-Variance trade-off in conventional statistical learning theory. (Not exactly the same. Refer to Fig.2 in Ref. [10])

Lemma 1 (Estimation error and Generalization error)**.**

$$\underbrace{R(\hat{f}_N) - R(f_{\mathcal{F}})}_{\text{Estimation error}} \leq 2 \sup_{f \in \mathcal{F}} | \underbrace{R(f) - \hat{R}_N(f)}_{\text{Generalization error}} | \tag{2}$$

Proof.

$$\underbrace{R(\hat{f}_N) - R(f_{\mathcal{F}})}_{\text{Estimation error}} = R(\hat{f}_N) - \hat{R}_N(\hat{f}_N) + \underline{\hat{R}_N(\hat{f}_N) - \hat{R}_N(f_{\mathcal{F}})}^{\leq 0} + \hat{R}_N(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \tag{3}$$

$$\leq R(\hat{f}_N) - \hat{R}_N(\hat{f}_N) + \hat{R}_N(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \tag{4}$$

$$\leq 2 \sup_{f \in \mathcal{F}} | \underbrace{R(f) - \hat{R}_N(f)}_{\text{Generalization error}} | \tag{5}$$

□

B. Errors in Quantum Machine Learning

In quantum machine learning, we have another source of error: "measurement error". Let $\tilde{f}_{N,M}$ be the hypothesis that we get after training on N samples with M measurements and define

$$\hat{f}_{N,M} \coloneqq \arg\min_{f \in \mathcal{F}} \hat{R}_{N,M}(f) \tag{6}$$

, where $\hat{R}_{N,M}(f)$ is the empirical risk of f estimated from M measurements for each training example out of N samples. Then, the total error in quantum machine learning is given by

$$R(\tilde{f}_{N,M}) - R(h^*) = \underbrace{R(\tilde{f}_{N,M}) - R(\hat{f}_{N,M})}_{\text{Optimization error}} + \underbrace{R(\hat{f}_{N,M}) - R(f_{\mathcal{F}})}_{\text{Estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{Approximation error}} \tag{7}$$

Lemma 2 (Estimation error and Generalization error in quantum machine learning)**.**
Estimation error in quantum machine learning can be bounded by the Generalization error and the Measurement error as follows:

$$\underbrace{R(\hat{f}_{N,M}) - R(f_{\mathcal{F}})}_{\text{Estimation error}} \leq 2 \sup_{f \in \mathcal{F}} | \underbrace{R(f) - \hat{R}_N(f)}_{\text{Generalization error}} | + 2 \sup_{f \in \mathcal{F}} | \underbrace{\hat{R}_N(f) - \hat{R}_{N,M}(f)}_{\text{Measurement error}} | \tag{8}$$

Proof.

$$\underbrace{R(\hat{f}_{N,M}) - R(f_{\mathcal{F}})}_{\text{Estimation error}} = R(\hat{f}_{N,M}) - \hat{R}_{N,M}(\hat{f}_{N,M}) + \underline{\hat{R}_{N,M}(\hat{f}_{N,M}) - \hat{R}_{N,M}(f_{\mathcal{F}})}^{\leq 0} + \hat{R}_{N,M}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \tag{9}$$

$$\leq R(\hat{f}_{N,M}) - \hat{R}_{N,M}(\hat{f}_{N,M}) + \hat{R}_{N,M}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \tag{10}$$

$$\leq 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_{N,M}(f)| \tag{11}$$

$$\leq 2 \sup_{f \in \mathcal{F}} | \underbrace{R(f) - \hat{R}_N(f)}_{\text{Generalization error}} | + 2 \sup_{f \in \mathcal{F}} | \underbrace{\hat{R}_N(f) - \hat{R}_{N,M}(f)}_{\text{Measurement error}} | \tag{12}$$

□

Remark 2. If we measure the output of the quantum model M times for each training example, then the Measurement error is $LC\sqrt{\frac{2\log(2/\delta)}{NM}}$ as shown in "6. Extension to unbiased estimates of measurement statistics" in Ref. [3].

Since the Generalization error bound is $\mathcal{O}\left(\sqrt{\frac{T}{N}}\right)$, the generalization error is still the dominant source of error in quantum machine learning.

$$R(f_{\boldsymbol{\theta}}) - \hat{R}_N(f_{\boldsymbol{\theta}}) \leq 2L\,\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) + LC\sqrt{\frac{2\log(2/\delta)}{NM}} + 3LC\sqrt{\frac{2\log(2/\delta)}{N}} \tag{13}$$

III. Notation and Setup

This section is maily based on the Ref. [4].

A. Supervised Learning Framework

We consider a standard supervised learning setting defined by an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. A learning algorithm is provided with a training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^N$, where samples are drawn i.i.d. from an unknown distribution $\mathcal{D}$. The objective is to learn a hypothesis function $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$, parameterized by $\boldsymbol{\theta}$, that minimizes the discrepancy between predictions and true labels.

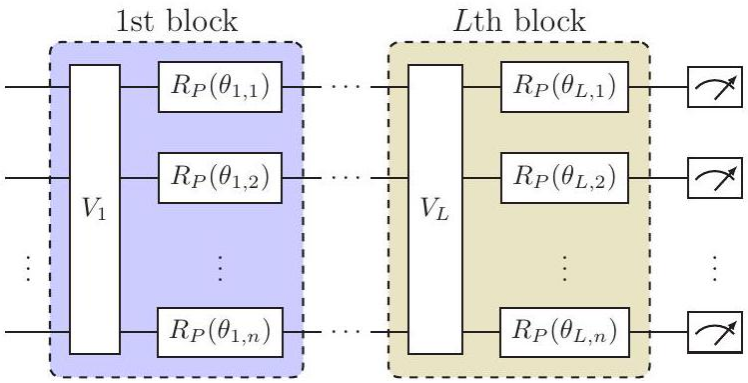

FIG. 2. Setup of PQC. It consists of L blocks, each of which begins with a fixed unitary operator followed by Pauli rotations applied to each qubit. There are nL trainable gates in total. [4]

B. Quantum Circuit Ansatz

Our hypothesis class consists of variational quantum machine learning (VQML) models. We specifically adopt a hardware-efficient ansatz for an N_q-qubit system composed of L layers (blocks). As shown in the circuit diagram (Fig. 2), the total unitary operation $U(\boldsymbol{\theta})$ is a product of layer-wise operations:

$$U(\boldsymbol{\theta}) = \prod_{l=L}^{1} R_l\left(\boldsymbol{\theta}_l\right) V_l, \tag{14}$$

where V_l represents a fixed unitary and $R_l(\boldsymbol{\theta}_l)$ represents a trainable rotation layer. The parameters are organized as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L)$, where each layer's parameters $\boldsymbol{\theta}_l$ contain N_q rotation angles. The rotation operator for the l-th layer is defined as a tensor product of single-qubit rotations:

$$R_l\left(\boldsymbol{\theta}_l\right) = \bigotimes_{m=1}^{N_q} e^{-\frac{i}{2}\theta_{l,m} P_{l,m}}. \tag{15}$$

Here, $P_{l,m} \in \{\mathbb{1}, X, Y, Z\}$ represents a standard Pauli operator chosen for the m-th qubit in layer l.

C. Quantum Embedding and Measurement

The interaction between the classical data and the quantum model proceeds as follows: 1. Encoding: An input x is mapped to a quantum state $\rho(x)$. 2. Evolution: The parameterized channel $\mathcal{E}_{\boldsymbol{\theta}}$ acts on the state via $\mathcal{E}_{\boldsymbol{\theta}}(x) = U(\boldsymbol{\theta})\rho(x)U^\dagger(\boldsymbol{\theta})$. 3. Measurement: An observable O is measured to produce the scalar output:

$$f_{\boldsymbol{\theta}}(x) \coloneqq \mathrm{Tr}[O\,\mathcal{E}_{\boldsymbol{\theta}}(x)]. \tag{16}$$

Note: While the standard setup assumes fixed V_l, this formalism accommodates "data re-uploading" schemes where V_l depends on input x, as well as noisy channels.

D. Risk and Generalization

We evaluate performance using a loss function $\mathcal{L}(y, \hat{y})$. We adopt two standard assumptions for the theoretical analysis: 1. Boundedness: The model output is bounded, $|f_{\boldsymbol{\theta}}(x)| \leq C$. 2. Lipschitz Continuity: The loss is L-Lipschitz with respect to the prediction: $|\mathcal{L}(y, z) - \mathcal{L}(y, z')| \leq L|z - z'|$. For example, $\mathcal{L}$ can be the square loss, the cross-entropy loss, or the hinge loss.

We define the Empirical Risk (training error) and Expected Risk (test error) respectively as:

$$\hat{R}_N(f_{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y_i, f_{\boldsymbol{\theta}}(x_i)), \tag{17}$$

$$R(f_{\boldsymbol{\theta}}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(y, f_{\boldsymbol{\theta}}(x))]. \tag{18}$$

The central quantity of interest in these notes is the **generalization error**, defined as $\mathrm{gen}(f_{\boldsymbol{\theta}}) \coloneqq R(f_{\boldsymbol{\theta}}) - \hat{R}_N(f_{\boldsymbol{\theta}})$. Subsequent sections will bound this quantity using covering numbers and Rademacher complexity. Finally, we apply the generalization error bound to the VQML model.

IV. Norm

In this section, we introduce the concept of norm and its properties. The p-norm extends the concept of the Euclidean norm, while the Schatten p-norm generalizes it to matrices.

Definition 1 (Norm)**.** Let $\mathcal{M}$ be a linear space (vector space) over $\mathbb{R}$ or $\mathbb{C}$. A norm on $\mathcal{M}$ is a function $\|\cdot\| : \mathcal{M} \to \mathbb{R}$ satisfying the following properties for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$ and $c \in \mathbb{R}$ or $\mathbb{C}$:

1. (Positivity) $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \boldsymbol{0}$.
2. (Homogeneity) $\|c\boldsymbol{x}\| = |c|\|\boldsymbol{x}\|$.
3. (Triangle inequality) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

Definition 2 (p-norm or L^p-norm)**.** Let $\boldsymbol{x} \coloneqq (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and $1 \leq p$. Then p-norm $\|\cdot\|_p$ of $\boldsymbol{x}$ is defined as

$$\|\boldsymbol{x}\|_p \coloneqq \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} & (1 \leq p < \infty) \\ \max_i |x_i| & (p = \infty) \end{cases} \tag{19}$$

and normalized p-norm $\|\boldsymbol{x}\|_{p,n}$ is defined as

$$\|\boldsymbol{x}\|_{p,n} \coloneqq \left(\frac{1}{n}\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} = \frac{1}{n^{\frac{1}{p}}}\|\boldsymbol{x}\|_p \tag{20}$$

Lemma 3 (Hölder's inequality)**.** Let $1 \le p, q \le \infty$ and $1/p + 1/q = 1$. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\left(|\langle \boldsymbol{x} | \boldsymbol{y} \rangle| := \left| \sum_{i=1}^{n} x_i y_i \right| \le \right) \sum_{i=1}^{n} |x_i y_i| \le \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q \tag{21}$$

Lemma 4 (Inequality of p-norms)**.** Let $1 \le p \le q$ and $\boldsymbol{x} \in \mathbb{R}^n$. The p-norm $\|\boldsymbol{x}\|_p$ is decreasing in p, that is,

$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_q \le \|\boldsymbol{x}\|_p \le \|\boldsymbol{x}\|_1 \tag{22}$$

And any two p-norms are related as

$$\|\boldsymbol{x}\|_q \le \|\boldsymbol{x}\|_p \le n^{(1/p - 1/q)} \|\boldsymbol{x}\|_q \tag{23}$$

In particular,

$$\|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1 \le \sqrt{n} \|\boldsymbol{x}\|_2 \tag{24}$$
$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2 \le \sqrt{n} \|\boldsymbol{x}\|_\infty \tag{25}$$
$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_1 \le n \|\boldsymbol{x}\|_\infty \tag{26}$$

Proof.

(First inequality) For any $\boldsymbol{x} \in \mathbb{R}^n$, there exists $\boldsymbol{y} \in \mathbb{R}^n$ such that $\boldsymbol{x} = \|\boldsymbol{x}\|_r \boldsymbol{y}$. Then, since $\|\boldsymbol{y}\|_r = 1$, for any $1 \le r < p$,

$$\|\boldsymbol{y}\|_p^p = \sum_{i=1}^{n} |y_i|^p \le \sum_{i=1}^{n} |y_i|^r = \|\boldsymbol{y}\|_r^r = 1 \quad (\because t^p \le t^r \text{ for } 0 \le t \le 1) \implies \|\boldsymbol{y}\|_p \le 1 \tag{27}$$

Therefore, $\|\boldsymbol{x}\|_p = \|\boldsymbol{x}\|_r \|\boldsymbol{y}\|_p \le \|\boldsymbol{x}\|_r \|\boldsymbol{y}\|_r = \|\boldsymbol{x}\|_r$.

(Second inequality) By Hölder's inequality,

$$\sum_{i=1}^{n} |a_i b_i| \le \left(\sum_{i=1}^{n} |a_i|^r \right)^{\frac{1}{r}} \left(\sum_{i=1}^{n} |b_i|^{\frac{r}{r-1}} \right)^{1 - \frac{1}{r}} \tag{28}$$

Let $|a_i| = |x_i|^p, |b_i| = 1$ and $r = q/p > 1$. Then

$$\|x\|_p^p = \sum_{i=1}^{n} |x_i|^p = \sum_{i=1}^{n} |x_i|^p \cdot 1 \le \left(\sum_{i=1}^{n} (|x_i|^p)^{\frac{q}{p}} \right)^{\frac{p}{q}} \left(\sum_{i=1}^{n} 1^{\frac{q}{q-p}} \right)^{1 - \frac{p}{q}} = \left(\sum_{i=1}^{n} |x_i|^q \right)^{\frac{p}{q}} n^{1 - \frac{p}{q}} \tag{29}$$

Therefore,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \le \left(\left(\sum_{i=1}^{n} |x_i|^q \right)^{\frac{p}{q}} n^{1 - \frac{p}{q}} \right)^{1/p} = \left(\sum_{i=1}^{n} |x_i|^q \right)^{\frac{1}{q}} n^{\frac{1}{p} - \frac{1}{q}} = n^{1/p - 1/q} \|x\|_q \tag{30}$$

□

Remark 3. Let $1 \le r$. If $\|\boldsymbol{y}\|_r = 1 (\iff \sum_{i=1}^{n} |y_i|^r = 1)$, then $|y_i| \le 1$ for any i. Because if there exists y_i such that $1 < |y_i|$, then $\sum_{i=1}^{n} |y_i|^r > 1$.

Remark 4. From the last inequality in the proof, we have $\|x\|_p \le n^{1/p - 1/q} \|x\|_q \iff \|x\|_{p,n} \le \|x\|_{q,n}$. Therefore, the normalized p-norm is increasing in p.

Definition 3 (Schatten p-norm)**.** Let $A \in \mathbb{C}^{m\times n}$ and $1 \le p$. Then Schatten p-norm of A is defined as

$$\|A\|_p = \begin{cases} (\mathrm{Tr}[|A|^p])^{\frac{1}{p}} = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p\right)^{\frac{1}{p}} & (1 \le p < \infty) \\ \sigma_1 & (p = \infty) \end{cases} \tag{31}$$

, where $|A| = \sqrt{A^\dagger A}$ and σ_i is the i-th largest singular value of A (i.e., the i-th largest eigenvalue of $|A|$).

Remark 5.

- When $p = 1$, the Schatten 1-norm is called the trace norm or nuclear norm.
- When $p = 2$, the Schatten 2-norm is called the Frobenius norm or Hilbert-Schmidt norm.
- When $p = \infty$, the Schatten ∞-norm is called the operator norm or spectral norm.

Lemma 5 (Properties of Schatten p-norm)**.** Let $A, B \in \mathcal{L}(\mathcal{H})$ and $1 \le p \le q \le \infty$. The Schatten p-norm has the following properties.

- For any $p \le q$, $\|A\|_q \le \|A\|_p$.
- For any $p \in [1, \infty]$ and $U, V \in \mathcal{U}(d)$, $\left\|UAV^\dagger\right\|_p = \|A\|_p$.
- $\|A\|_p = \left\|A^\mathsf{T}\right\|_p = \|A^*\|_p = \left\|A^\dagger\right\|_p$.
- For any p, q, r and $1/p + 1/q \le 1/r$, $\|AB\|_r \le \|A\|_p \|B\|_q$ (Hölder's inequality).
  - $\|AB\|_1 \le \|A\|_1 \|B\|_\infty$.
  - $\|AB\|_1 \le \|A\|_2 \|B\|_2$.

Definition 4 (norm ball)**.** Let $(\mathcal{M}, \|\cdot\|)$ be a norm space. A norm ball centered at $\boldsymbol{x} \in \mathcal{M}$ with radius $0 < \varepsilon$ is defined as

$$B(\boldsymbol{x}, \varepsilon, \|\cdot\|) \coloneqq \{\ \boldsymbol{y} \in \mathcal{M} \mid \|\boldsymbol{y} - \boldsymbol{x}\| \le \varepsilon\ \} \tag{32}$$

Remark 6. When p-norm is used, we call it p-norm ball and denote it as $B_p(\boldsymbol{x}, \varepsilon)$. From the Fig. 5, we can see that $B_1(\boldsymbol{x}, \varepsilon) \subset B_p(\boldsymbol{x}, \varepsilon) \subset B_q(\boldsymbol{x}, \varepsilon) \subset B_\infty(\boldsymbol{x}, \varepsilon)$ for $1 \le p \le q \le \infty$.

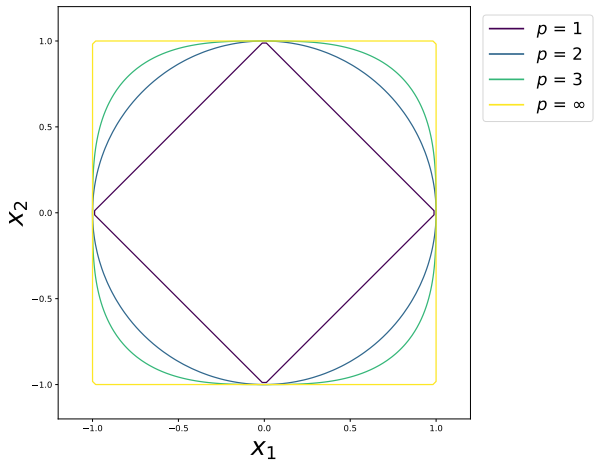

FIG. 3. The contours of p-norm balls with $\varepsilon = 1$ in $\mathbb{R}^2$ for $p = 1, 2, 3, \infty$.

Definition 5 (Diamond norm)**.** Let $\mathcal{E} : \mathcal{L}(\mathcal{H}_1) \to \mathcal{L}(\mathcal{H}_2)$ be a linear map. The diamond norm of $\mathcal{E}$ is defined as

$$\|\mathcal{E}\|_\diamond \coloneqq \sup_{n\in\mathbb{N}} \sup_{\|A\|_1\leq 1} \|(\mathcal{E}\otimes \mathbb{1}_{\mathbb{C}^n})(A)\|_1 \tag{33}$$

$$= \sup\{\ \|(\mathcal{E}\otimes \mathbb{1}_{\mathcal{H}_1})(|\psi\rangle\langle\phi|)\|_1 \mid |\psi\rangle, |\phi\rangle \in \mathcal{H}_1\otimes\mathcal{H}_1\ \} \tag{34}$$

Remark 7.

- When $\mathcal{E}$ is a quantum channel, or a completely positive and trace-preserving (CPTP) map, the diamond norm is $\|\mathcal{E}\|_\diamond = 1$.
- If the map $\mathcal{E}$ is Hermiticity-preserving (e.g. $\mathcal{E}$ is the difference of two quantum channels), one can optimize over $|\psi\rangle = |\phi\rangle$ in the formula above.
- Refer to Ref. [11] or Ref. [12] for further details on the diamond norm.

Lemma 6 (Subadditivity of diamond distance [4])**.** For any quantum channels $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4$, where $\mathcal{E}_2$ and $\mathcal{E}_4$ map from n-qubit to m-qubit systems and $\mathcal{E}_1$ and $\mathcal{E}_3$ map from m-qubit to k-qubit systems, we have

$$\|\mathcal{E}_1\circ\mathcal{E}_2 - \mathcal{E}_3\circ\mathcal{E}_4\|_\diamond \leq \|\mathcal{E}_1 - \mathcal{E}_3\|_\diamond + \|\mathcal{E}_2 - \mathcal{E}_4\|_\diamond \tag{35}$$

, where $\|\cdot\|_\diamond$ denotes the diamond norm of a quantum channel.

Proof.

$$\|\mathcal{E}_1\circ\mathcal{E}_2 - \mathcal{E}_3\circ\mathcal{E}_4\|_\diamond = \|\mathcal{E}_1\circ\mathcal{E}_2 - \mathcal{E}_3\circ\mathcal{E}_2 + \mathcal{E}_3\circ\mathcal{E}_2 - \mathcal{E}_3\circ\mathcal{E}_4\|_\diamond \tag{36}$$

$$= \|(\mathcal{E}_1 - \mathcal{E}_3)\circ\mathcal{E}_2\| + \|\mathcal{E}_3\circ(\mathcal{E}_2 - \mathcal{E}_4)\|_\diamond \tag{37}$$

$$\leq \|\mathcal{E}_1 - \mathcal{E}_3\|_\diamond\|\mathcal{E}_2\|_\diamond + \|\mathcal{E}_3\|_\diamond\|\mathcal{E}_2 - \mathcal{E}_4\|_\diamond \tag{38}$$

$$\leq \|\mathcal{E}_1 - \mathcal{E}_3\|_\diamond + \|\mathcal{E}_2 - \mathcal{E}_4\|_\diamond \tag{39}$$

□

Lemma 7 (Spectral norm and diamond norm of unitary channels [4])**.** Let $\mathcal{E}_U(\rho) = U\rho U^\dagger$ and $\mathcal{E}_V(\rho) = V\rho V^\dagger$ be

unitary channels. Then,

$$\|\mathcal{E}_U - \mathcal{E}_V\|_\diamond \leq 2\|U - V\|_\infty$$

where $\|\cdot\|_\infty$ is the Schatten ∞-norm.

Proof. For any $|u\rangle, |v\rangle \in \mathcal{H}$, we have $|v\rangle = a|u\rangle + b|u^\perp\rangle$, where $|u^\perp\rangle$ is orthogonal to $|u\rangle$ and $|a|^2 + |b|^2 = 1$. Then,

$$\begin{aligned}
|u\rangle\langle u| - |v\rangle\langle v| &= |u\rangle\langle u| - (a|u\rangle + b|u^\perp\rangle)(a^*\langle u| + b^*\langle u^\perp|) && (40)\\
&= |u\rangle\langle u| - |a|^2|u\rangle\langle u| - |b|^2|u^\perp\rangle\langle u^\perp| - ab^*|u\rangle\langle u^\perp| - a^*b|u^\perp\rangle\langle u| && (41)\\
&= \begin{pmatrix} 1-|a|^2 & -ab^* \\ -a^*b & -|b|^2 \end{pmatrix} && (42)\\
&= \begin{pmatrix} |b|^2 & -ab^* \\ -a^*b & -|b|^2 \end{pmatrix} && (43)
\end{aligned}$$

Since

$$\begin{aligned}
&\det(\lambda I - (|u\rangle\langle u| - |v\rangle\langle v|)) = 0 && (44)\\
\iff &\lambda^2 - (|a|^2 + |b|^2)|b|^2 = 0 && (45)\\
\iff &\lambda^2 - |b|^2 = 0 && (46)\\
\iff &\lambda = \pm|b| = \pm\sqrt{1-|a|^2} = \pm\sqrt{1 - |\langle u|v\rangle|^2} && (47)
\end{aligned}$$

, and trace norm is the sum of singular values, we have

$$\begin{aligned}
\frac{1}{2}\||u\rangle\langle u| - |v\rangle\langle v|\|_1 &= \sqrt{1 - |\langle u|v\rangle|^2} && (48)\\
&= \sqrt{(1 + |\langle u|v\rangle|)(1 - |\langle u|v\rangle|)} && (49)\\
&\leq \sqrt{2(1 - \mathrm{Re}(\langle u|v\rangle))} && (50)\\
&= \||u\rangle - |v\rangle\|_2 && (51)
\end{aligned}$$

The last equality is due to the fact that

$$\||u\rangle - |v\rangle\|_2 = \sqrt{(\langle u| - \langle v|)(|u\rangle - |v\rangle)} = \sqrt{2 - \langle u|v\rangle - \langle v|u\rangle} = \sqrt{2(1 - \mathrm{Re}(\langle u|v\rangle))}.$$

The diamond distance bound then is a direct consequence of this relation.

$$\begin{aligned}
\frac{1}{2}\|\mathcal{E}_U - \mathcal{E}_V\|_\diamond &= \sup_{|\psi\rangle\langle\psi|} \frac{1}{2}\|\mathcal{E}_U(|\psi\rangle\langle\psi|) - \mathcal{E}_V(|\psi\rangle\langle\psi|)\|_1 && (52)\\
&\leq \sup_{|\psi\rangle} \|(U - V)|\psi\rangle\|_2 && (53)\\
&= \text{largest singular value of } U - V && (54)\\
&= \|U - V\|_\infty && (55)
\end{aligned}$$

□

The following lemma translates the distance of rotation operators to the distance of their corresponding angles.

Lemma 8 ([4])**.** Given an arbitrary Pauli matrix $P \in \{I, X, Y, Z\}$ and two arbitrary angles θ and $\tilde{\theta}$, the corresponding 1-qubit rotation operators are $R(\theta) = e^{-\frac{i}{2}\theta P}$ and $R(\tilde{\theta}) = e^{-\frac{i}{2}\tilde{\theta} P}$, respectively. Then, the distance between the two operators measured by the Schatten ∞-norm can be upper bounded as

$$\|R(\theta) - R(\tilde{\theta})\|_\infty \leq \frac{1}{2}|\theta - \tilde{\theta}| \quad (56)$$

Proof. According to the definition of rotation operators, we have $R(\theta)-R(\tilde{\theta}) = \left(\cos\frac{\theta}{2} - \cos\frac{\tilde{\theta}}{2}\right) I - i\left(\sin\frac{\theta}{2} - \sin\frac{\tilde{\theta}}{2}\right) P$, whose singular value is $2\sin\frac{\theta-\tilde{\theta}}{4}$ with the multiplicity 2. Thus,

$$\|R(\theta) - R(\tilde{\theta})\|_\infty = 2\left|\sin\frac{\theta-\tilde{\theta}}{4}\right| \leq \frac{1}{2}|\theta - \tilde{\theta}| \tag{57}$$

□

V. Covering Number

We first introduce the covering number, which is a geometrically intuitive measure of complexity and has been widely used in machine learning theory to analyze generalization error."

Definition 6 (Covering net and covering number [4])**.** Let $(\mathcal{M}, d)$ be a metric space. Consider a bounded subset $\mathcal{A} \subset \mathcal{M}$ and let $\varepsilon > 0$. A subset $\mathcal{N} \subseteq \mathcal{M}$ is called an ε**-covering net** of $\mathcal{A}$ if every point in $\mathcal{A}$ is within a distance ε of some point of $\mathcal{N}$, i.e.,

$$\forall x \in \mathcal{A},\ \exists y \in \mathcal{N}\ \text{ s.t. } d(x, y) \leq \varepsilon \tag{58}$$

The smallest cardinality of an ε-covering net of $\mathcal{A}$ w.r.t. d is called the ε**-covering number** of $\mathcal{A}$ w.r.t. d, denoted by $N(\mathcal{A}, \varepsilon, d)$, i.e.,

$$N(\mathcal{A}, \varepsilon, d) \coloneqq \min\{\ n \in \mathbb{N} \mid \exists x_1, x_2, \ldots, x_n \in \mathcal{M}\ \text{ s.t. } \mathcal{A} \subseteq \textstyle\bigcup_{i=1}^n B_d(x_i, \varepsilon)\ \},$$

If $\mathcal{N} \subseteq \mathcal{A}$, the covering number is denoted by $N_{\text{in}}(\mathcal{A}, \varepsilon, d)$ and called **internal covering number**.

$$N_{\text{in}}(\mathcal{A}, \varepsilon, d) \coloneqq \min\{\ n \in \mathbb{N} \mid \exists x_1, x_2, \ldots, x_n \in \mathcal{A}\ \text{ s.t. } \mathcal{A} \subseteq \textstyle\bigcup_{i=1}^n B_d(x_i, \varepsilon)\ \},$$

You can intuitively understand the covering number as the minimum number of balls with radius ε needed to cover the set $\mathcal{A}$ with the metric d. The covering number is a measure of the complexity of the set $\mathcal{A}$. By definition, $N(\mathcal{A}, \varepsilon, d) \leq N_{\text{in}}(\mathcal{A}, \varepsilon, d)$ always holds. For example, when $\mathcal{A}$ is a doughnut-shaped set, the covering number can be smaller than the internal covering number.

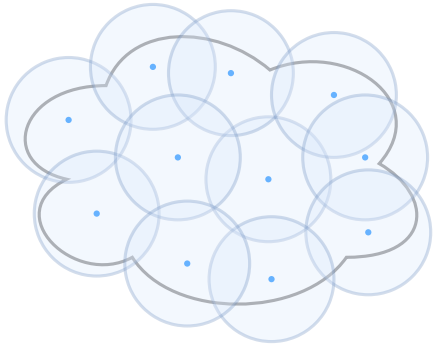

FIG. 4. Illustration of a covering net

Lemma 9 (Properties of covering number)**.** Let $(\mathcal{M}, \|\cdot\|)$ be a normed vector space and $\alpha > 0$. Then, for any bounded subset $\mathcal{A} \subset \mathcal{M}$, $\varepsilon > 0$ and $\|\cdot\|$, the following equality holds:

$$N(\alpha\mathcal{A}, \varepsilon, \|\cdot\|) = N(\mathcal{A}, \varepsilon/\alpha, \|\cdot\|) = N(\mathcal{A}, \varepsilon, \alpha\|\cdot\|) \tag{A}$$

$$\begin{aligned} N(\mathcal{A}, \varepsilon, \|\cdot\|) &= N(\alpha\mathcal{A}, \alpha\varepsilon, \|\cdot\|) \\ &= N(\mathcal{A}, \alpha\varepsilon, \alpha\|\cdot\|) \end{aligned} \tag{B}$$

$$= N(\mathcal{A}/\alpha, \varepsilon, \alpha\|\cdot\|)$$

Let $\mathcal{A} \subset \mathcal{B} \subset \mathcal{M}$, $0 < \varepsilon_1 < \varepsilon_2$, and $\|\cdot\|_{\spadesuit} \leq \|\cdot\|_{\clubsuit}$. Then the following inequality holds:

$$\begin{aligned} N(\mathcal{A}, \varepsilon, \|\cdot\|) &\leq N(\mathcal{B}, \varepsilon, \|\cdot\|) \\ N(\mathcal{A}, \varepsilon_2, \|\cdot\|) &\leq N(\mathcal{A}, \varepsilon_1, \|\cdot\|) \\ N(\mathcal{A}, \varepsilon, \|\cdot\|_{\spadesuit}) &\leq N(\mathcal{A}, \varepsilon, \|\cdot\|_{\clubsuit}) \end{aligned} \tag{C}$$

Proof.
(Proof of (A))

- The first equality holds because expanding (or shrinking) the set $\mathcal{A}$ by a factor of α is equivalent to shrinking (or expanding) the radius of the balls by a factor of $1/\alpha$ since the covering number only depends on the relative size of the radius and the set.
- The second equality holds because $d(x, y) \leq \varepsilon/\alpha$ is equivalent to $\alpha d(x, y) \leq \varepsilon$.

(Proof of (B))

- Since the covering number only depends on the relative size of the radius and the set, expanding (or shrinking) the set $\mathcal{A}$ and the radius of the balls by a factor of α at the same time does not change the covering number.
- The second equality holds because $d(x, y) \leq \varepsilon$ is equivalent to $\alpha d(x, y) \leq \alpha\varepsilon$.
- Use the first and second equalities.

(Proof of (C))

- The first inequality holds because the covering number of a larger set is always larger than that of a smaller set.
- The second inequality holds because the covering number of a set with a ball of larger radius is always smaller than that of a set with a ball of smaller radius.
- If $d_2(x, y) \leq \varepsilon$, then $d_1(x, y) \leq d_2(x, y) \leq \varepsilon$. This implies that if Q is an ε-covering net of $\mathcal{A}$ with d_2, then Q is also an ε-covering net of $\mathcal{A}$ with d_1 and thus

$$W(\mathcal{A}, \varepsilon, d_1) \supseteq W(\mathcal{A}, \varepsilon, d_2) \tag{59}$$

, where $W(\mathcal{A}, \varepsilon, d)$ represents the set of all ε-covering nets of $\mathcal{A}$ with d. Since the covering number is the minimum cardinality of the covering net, it follows that

$$\begin{aligned} N(\mathcal{A}, d_1, \varepsilon) &= \min_{Q \in W(\mathcal{A}, \varepsilon, d_1)} |Q| && (60) \\ &\leq \min_{Q \in W(\mathcal{A}, \varepsilon, d_2)} |Q| && (61) \\ &= N(\mathcal{A}, d_2, \varepsilon) && (62) \end{aligned}$$

□

Lemma 10 (Covering number of normalized norm)**.**

$$N(\mathcal{A}, \varepsilon, \|\cdot\|_p) = N(\mathcal{A}, \varepsilon/m^{\frac{1}{p}}, \|\cdot\|_{p,m}) \tag{63}$$

Proof. From the definition of normalized p-norm, $\|a - a'\|_p \leq \varepsilon$ and $\|a - a'\|_{p,m} \leq \varepsilon/m^{\frac{1}{p}}$ is equivalent, thus the equation holds. □

Lemma 11 (Inequality of covering numbers with different norms)**.** The following inequality holds for any $\varepsilon > 0$

and any bounded subset $\mathcal{A} \subset \mathbb{R}^m$:

$$N(\mathcal{A}, \varepsilon, \|\cdot\|_\infty) \leq N(\mathcal{A}, \varepsilon, \|\cdot\|_q) \quad \leq N(\mathcal{A}, \varepsilon, \|\cdot\|_p) \quad \leq N(\mathcal{A}, \varepsilon, \|\cdot\|_1) \tag{64}$$

$$N(\mathcal{A}, \varepsilon, \|\cdot\|_{1,n}) \leq N(\mathcal{A}, \varepsilon, \|\cdot\|_{p,n}) \leq N(\mathcal{A}, \varepsilon, \|\cdot\|_{q,n}) \leq N(\mathcal{A}, \varepsilon, \|\cdot\|_{\infty,n}) \tag{65}$$

where $1 \leq p \leq q < \infty$.

Proof. Use the Lemma 4 and the inequality (C) in Lemma 9. □

Remark 8. From the properties of p-norm, for any two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ and $1 \leq p \leq q < \infty$, you have $\|\boldsymbol{x} - \boldsymbol{y}\|_\infty \leq \|\boldsymbol{x} - \boldsymbol{y}\|_q \leq \|\boldsymbol{x} - \boldsymbol{y}\|_p \leq \|\boldsymbol{x} - \boldsymbol{y}\|_1$. This implies that $B_1(\boldsymbol{x}, \varepsilon) \subset B_p(\boldsymbol{x}, \varepsilon) \subset B_q(\boldsymbol{x}, \varepsilon) \subset B_\infty(\boldsymbol{x}, \varepsilon)$. This is also obvious from Fig. 5. From this fact, you can see that you need more balls to cover $\mathcal{A}$ with the q-norm ball than with the p-norm ball.

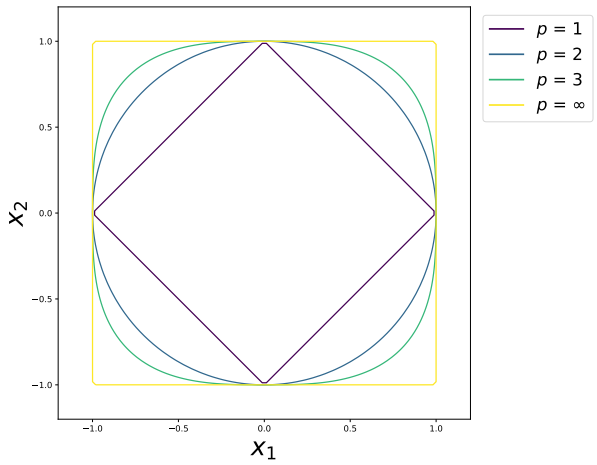

FIG. 5. The contours of p-norm balls with $\varepsilon = 1$ in $\mathbb{R}^2$ for $p = 1, 2, 3, \infty$.

Lemma 12. The following inequality holds for any $\varepsilon > 0$, some subset $\mathcal{A} \subset \mathbb{R}^m$, and $1 \leq r < p$:

$$N(\mathcal{A}, \varepsilon, \|\cdot\|_p) \leq N(\mathcal{A}, \varepsilon, \|\cdot\|_r) \leq N(\mathcal{A}, \varepsilon, m^{(1/r-1/p)}\|\cdot\|_p) \tag{66}$$

Proof. From Lemma 4, it holds that $\|\boldsymbol{x}\|_p \leq \|\boldsymbol{x}\|_r \leq m^{(1/r-1/p)}\|\boldsymbol{x}\|_p$ for $1 \leq r < p$ and $\boldsymbol{x} \in \mathbb{R}^n$. Thus, you can use the last inequality in (C) in Lemma 9. □

Example 1. When $p = \infty$ and $r = 1$, you have $N(\mathcal{A}, \varepsilon, \|\cdot\|_\infty) \leq N(\mathcal{A}, \varepsilon, \|\cdot\|_1) \leq N(\mathcal{A}, \varepsilon, m\|\cdot\|_\infty) = N(\mathcal{A}, \varepsilon/m, \|\cdot\|_\infty)$. The Fig. 6 illustrates the inequality between covering numbers visually. The Fig. 6 shows the contours of 1-norm ball with $\varepsilon = 1$ and ∞-norm balls with $\varepsilon = 1/2, 1$ in $\mathbb{R}^2$. You can see that $B_\infty(\boldsymbol{x}, \varepsilon/2) \subset B_1(\boldsymbol{x}, \varepsilon)$. In $\mathbb{R}^m$, you have $B_\infty(\boldsymbol{x}, \varepsilon/m) \subset B_1(\boldsymbol{x}, \varepsilon)$. Thus, you need more balls to cover $\mathcal{A}$ with the ∞-norm ball with radius ε/m than with the 1-norm ball with radius ε.

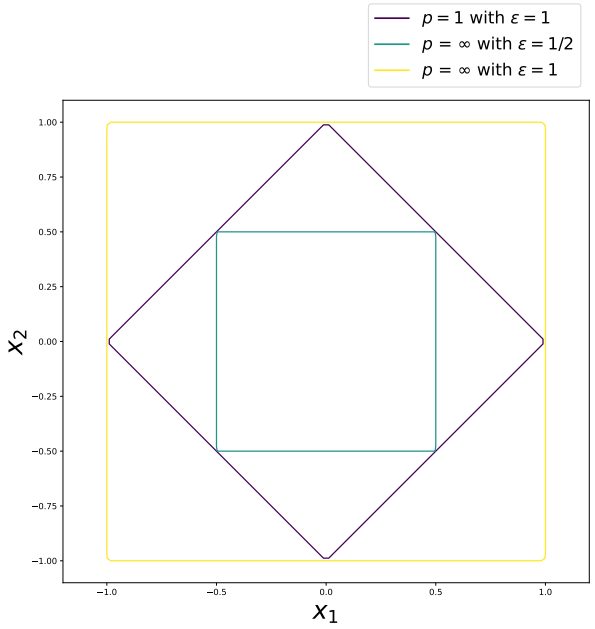

FIG. 6. The contours of 1-norm ball with $\varepsilon = 1$ in $\mathbb{R}^2$ and ∞-norm balls with $\varepsilon = 1/2, 1$ in $\mathbb{R}^2$

The following lemma enables us to employ the covering number of one metric space to bound the covering number of another metric space. (Lemma 5 in Ref. [7].)

Lemma 13 (Covering numbers of two metric spaces [4])**.** Let $(\mathcal{M}_1, d_1)$ and $(\mathcal{M}_2, d_2)$ be two metric spaces and $f : \mathcal{M}_1 \to \mathcal{M}_2$ be K-Lipschitz such that

$$d_2(f(x), f(y)) \leq K\, d_1(x, y), \forall x, y \in \mathcal{M}_1 \tag{67}$$

where K is a constant. Then, you can use the covering number of a bounded subset $\mathcal{A}_1 \subset \mathcal{M}_1$ with d_1 to bound the covering number of a bounded subset $f(\mathcal{A}_1) \subset \mathcal{M}_2$ with d_2 as

$$N\left(f(\mathcal{A}_1), \varepsilon, d_2\right) \leq N\left(\mathcal{A}_1, \varepsilon/K, d_1\right) \tag{68}$$

Proof. Let $\mathcal{Q}_1$ be an ε/K-covering net for $\mathcal{A}_1$ with d_1, that is, $\forall x \in \mathcal{A}_1$, $\exists y \in \mathcal{Q}_1$ such that $d_1(x, y) \leq \varepsilon/K$. Since, for any $X \in f(\mathcal{A}_1)$, there exists $x \in \mathcal{A}_1$ such that $X = f(x)$, and there exists $y \in \mathcal{Q}_1$ such that $d_1(x, y) \leq \varepsilon/K$, it follows $\forall X \in f(\mathcal{A}_1)$, $\exists Y = f(y) \in f(\mathcal{Q}_1)$ such that $d_2(X, Y) \leq K d_1(x, y) \leq \varepsilon$ (the first inequality is due to assumption of K-Lipschitzness of f). Therefore, $f(\mathcal{Q}_1)$ is an ε-covering net of $f(\mathcal{A}_1)$ with d_2. It implies that if Q_1 is an ε/K-covering net of $\mathcal{A}_1$ with d_1, then $f(Q_1)$ is an ε-covering net of $f(\mathcal{A}_1)$ with d_2, and thus

$$W(f(\mathcal{A}_1), \varepsilon, d_2) \supseteq \{\ f(\mathcal{Q}_1) \mid \mathcal{Q}_1 \in W(\mathcal{A}_1, \varepsilon/K, d_1)\ \} \tag{69}$$

, where $W(f(\mathcal{A}_1), \varepsilon, d_2)$ is the set of all ε-covering nets of $f(\mathcal{A}_1)$ with d_2 and $W(\mathcal{A}_1, \varepsilon/K, d_1)$ is the set of all ε/K-covering nets of $\mathcal{A}_1$ with d_1. Since the covering number is the minimum cardinality of the covering net, it follows that

$$N(f(\mathcal{A}_1), d_2, \varepsilon) = \min_{\mathcal{Q}_2 \in W(f(\mathcal{A}_1), \varepsilon, d_2)} |\mathcal{Q}_2| \tag{70}$$

$$\overset{(a)}{\leq} \min_{\mathcal{Q}_1 \in W(\mathcal{A}_1, \varepsilon/K, d_1)} |f(\mathcal{Q}_1)| \tag{71}$$

$$\overset{(b)}{\leq} \min_{\mathcal{Q}_1 \in W(\mathcal{A}_1, \varepsilon/K, d_1)} |\mathcal{Q}_1| \tag{72}$$

$$= N(\mathcal{A}_1, d_1, \varepsilon/K) \tag{73}$$

Equality in (a) holds if the equality in (69) holds, and equality in (b) holds if f is injective. □

Example 2. Let $\mathcal{A}_1 := [0,1]$ and $f : \mathcal{A}_1 \ni x \mapsto \pi x \subset [0,\pi] = f(\mathcal{A}_1)$, then f is K-Lipschitz with $K = \pi$. Then, $N(\mathcal{A}_1, \varepsilon, \|\cdot\|) = \lceil \frac{1}{2\varepsilon} \rceil$ and $N(f(\mathcal{A}_1), \varepsilon, \|\cdot\|) = \lceil \frac{\pi}{2\varepsilon} \rceil$. Thus, $N(f(\mathcal{A}_1), \varepsilon, \|\cdot\|) = \lceil \frac{\pi}{2\varepsilon} \rceil \leq \lceil \frac{\pi}{2\varepsilon} \rceil = N(\mathcal{A}_1, \varepsilon/\pi, \|\cdot\|)$.

Example 3. Let $\mathcal{A}_1 := [0,1]$ and $f : \mathcal{A}_1 \ni x \mapsto \sin(\pi x) \subset [0,1] = f(\mathcal{A}_1)$, then f is K-Lipschitz with $K = \pi$. Then, $N(\mathcal{A}_1, \varepsilon, \|\cdot\|) = \lceil \frac{1}{2\varepsilon} \rceil$ and $N(f(\mathcal{A}_1), \varepsilon, \|\cdot\|) = \lceil \frac{1}{2\varepsilon} \rceil$. Thus, $N(f(\mathcal{A}_1), \varepsilon, \|\cdot\|) = \lceil \frac{1}{2\varepsilon} \rceil \leq \lceil \frac{\pi}{2\varepsilon} \rceil = N(\mathcal{A}_1, \varepsilon/\pi, \|\cdot\|)$.

Definition 7 (Packing)**.** Let $(\mathcal{M}, d)$ be a metric space. Consider a bounded subset $\mathcal{A} \subset \mathcal{M}$ and let $\varepsilon > 0$. A subset $\mathcal{P} \subseteq \mathcal{A}$ is called an ε**-packing** of $\mathcal{A}$ if every pair of distinct points in $\mathcal{P}$ is separated by a distance greater than ε, i.e.,

$$\forall x, y \in \mathcal{P},\ d(x,y) > \varepsilon. \tag{74}$$

The largest possible cardinality of an ε-packing of $\mathcal{A}$ w.r.t. d is called the ε**-packing number** of $\mathcal{A}$ w.r.t. d, denoted by $M(\mathcal{A}, \varepsilon, d)$, i.e.,

$$M(\mathcal{A}, \varepsilon, d) \coloneqq \max\{\ |\mathcal{P}| : \mathcal{P} \text{ is an } \varepsilon\text{-packing of } \mathcal{A}\ \} \tag{75}$$

$$\coloneqq \max\{\ n \in \mathbb{N} \mid \exists x_1, x_2, \ldots, x_n \in \mathcal{A} \text{ s.t. } \forall i, j \in [n],\ i \neq j \Rightarrow d(x_i, x_j) > \varepsilon\ \} \tag{76}$$

, where $[n] \coloneqq \{\ 1, 2, \ldots, n\ \}$.

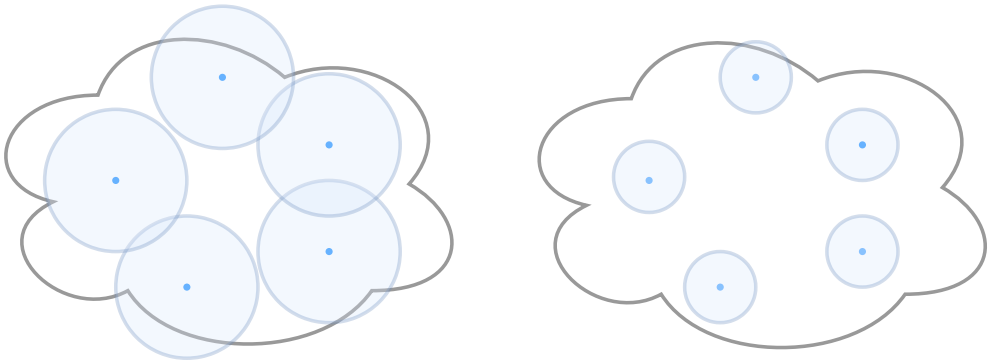

FIG. 7. (Left) Illustration of a packing. All points in the set $\mathcal{P}$ are separated by a distance greater than ε. You can add more points to $\mathcal{P}$ without violating the condition because there is still some space not covered by the balls. (Right) If we consider the balls with radius $\varepsilon/2$ centered at the points in $\mathcal{P}$, then the balls are disjoint.

Lemma 14 (Inequality between covering number and packing number)**.** Let $\mathcal{A} \subset \mathbb{R}^m$ be a bounded subset and $\varepsilon > 0$. Then, the following inequality holds:

$$N(\mathcal{A}, \varepsilon, d) \leq N_{\text{in}}(\mathcal{A}, \varepsilon, d) \leq M(\mathcal{A}, \varepsilon, d) \leq N(\mathcal{A}, \varepsilon/2, d) \tag{77}$$

Proof.

The first inequality is trivial from the definition of the covering number.

(Proof of the second inequality)

If $\mathcal{P}$ is an ε-packing of $\mathcal{A}$ such that $|\mathcal{P}| = M(\mathcal{A}, \varepsilon, d)$, for any point $a \in \mathcal{A} \backslash \mathcal{P}$, the set $\mathcal{P} \cup \{a\}$ is not an ε-packing of $\mathcal{A}$. Namely, for any $a \in \mathcal{A}$ there exists $p \in \mathcal{P}$ such that $d(a,p) \leq \varepsilon$. Hence, $\mathcal{P}$ is an internal ε-covering net of $\mathcal{A}$, and so $N_{\text{in}}(\mathcal{A}, \varepsilon, d) \leq |\mathcal{P}|$.

(Proof of the third inequality)

Let $\mathcal{B}$ be an $\varepsilon/2$-covering net of $\mathcal{A}$ such that $|\mathcal{B}| = N(\mathcal{A}, \varepsilon/2, d)$. Let $\mathcal{P}$ be any ε-packing of $\mathcal{A}$. Since any two points in $\mathcal{P}$ are separated by a distance greater than ε (i.e., $d(p,p') > \varepsilon$ for any $p, p' \in \mathcal{P}$), for any $b \in \mathcal{B}$, there exists at most one $p \in \mathcal{P}$ such that $d(b,p) \leq \varepsilon/2$. Hence, $|\mathcal{P}| \leq |\mathcal{B}|$, and so $M(\mathcal{A}, \varepsilon, d) \leq N(\mathcal{A}, \varepsilon/2, d)$. □

Lemma 15 (covering number of a norm ball)**.** Let $\boldsymbol{x} \in \mathbb{R}^m$, $0 < \varepsilon \leq R$, and $\|\cdot\|$ be any norm on $\mathbb{R}^m$. Then, the

covering number of the norm ball $B(\boldsymbol{x}, R, \|\cdot\|)$ is bounded as

$$\left(\frac{R}{\varepsilon}\right)^m \leq N(B(\boldsymbol{x}, R, \|\cdot\|), \varepsilon, \|\cdot\|) \leq \left(1 + \frac{2R}{\varepsilon}\right)^m \leq \left(\frac{3R}{\varepsilon}\right)^m. \tag{78}$$

Proof. We denote the volume of the unit ball in $\mathbb{R}^m$ as V. Then, the volume of $B(\boldsymbol{x}, r, \|\cdot\|)$ is $r^m V$.
(Proof of the first inequality)
Let $\mathcal{N}$ be an ε-covering net of $B(\boldsymbol{x}, R, \|\cdot\|)$ such that $|\mathcal{N}| = N(B(\boldsymbol{x}, R, \|\cdot\|), \varepsilon, \|\cdot\|)$. Then, since $B(\boldsymbol{x}, R, \|\cdot\|) \subset \cup_{n\in\mathcal{N}} B(n, \varepsilon, \|\cdot\|)$, we have

$$\text{Volume}(B(\boldsymbol{x}, R, \|\cdot\|)) \leq \text{Volume}(\cup_{n\in\mathcal{N}} B(n, \varepsilon, \|\cdot\|)) \leq \sum_{n\in\mathcal{N}} \text{Volume}(B(n, \varepsilon, \|\cdot\|)). \tag{79}$$

Thus, $R^m V \leq |\mathcal{N}| \times \varepsilon^m V$, and so $(R/\varepsilon)^m \leq |\mathcal{N}|$.

(Proof of the second inequality)
Let $\mathcal{P}$ be an ε-packing of $B(\boldsymbol{x}, R, \|\cdot\|)$ such that $|\mathcal{P}| = M(B(\boldsymbol{x}, R, \|\cdot\|), \varepsilon, \|\cdot\|)$. Then, since the balls $B(p, \varepsilon/2, \|\cdot\|)$ are disjoint for $p \in \mathcal{P}$ and all these balls are contained in $B(\boldsymbol{x}, R + \varepsilon/2, \|\cdot\|)$, we have

$$\text{Volume}(\cup_{p\in\mathcal{P}} B(p, \varepsilon/2, \|\cdot\|)) = \sum_{p\in\mathcal{P}} \text{Volume}(B(p, \varepsilon/2, \|\cdot\|)) \leq \text{Volume}(B(\boldsymbol{x}, R + \varepsilon/2, \|\cdot\|)). \tag{80}$$

Thus, $|\mathcal{P}| \times (\varepsilon/2)^m V \leq (\varepsilon/2 + R)^m V$, and so $|\mathcal{P}| \leq (1 + 2R/\varepsilon)^m$. The last inequality holds because $1 + 2R/\varepsilon \leq 3R/\varepsilon$. Finally, we use the fact that $|\mathcal{N}| \leq |\mathcal{P}|$. □

Lemma 16 (covering number of a hypercube)**.** The covering number of a hypercube $\Theta := \{\ (\theta_1, \theta_2, \ldots, \theta_m)\ |\ \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R}\ \} \subset \mathbb{R}^m$ w.r.t. the ∞-norm is equal to

$$N(\Theta,\ \varepsilon, \|\cdot\|_\infty) = \prod_{i=1}^m \left\lceil \frac{c_i}{2\varepsilon} \right\rceil \tag{81}$$

The ∞-norm ball is a hypercube with side length 2ε. For each side of the hypercube Θ, you need $\lceil \frac{c_i}{2\varepsilon} \rceil$ hypercubes to cover the side. Thus, you need $\prod_{i=1}^m \lceil \frac{c_i}{2\varepsilon} \rceil$ hypercubes to cover the hypercube Θ.

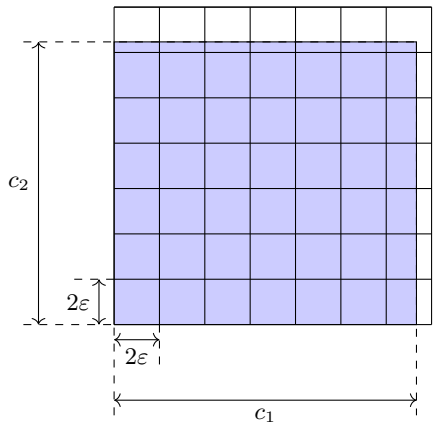

FIG. 8. ε-covering net of a hypercube $\Theta \subset \mathbb{R}^2$ w.r.t. $\|\cdot\|_\infty$

In the following, we consider $\Theta := \{\ (\theta_1, \theta_2, \ldots, \theta_m)\ |\ \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R}\ \} \subset \mathbb{R}^m$ in which some parameters have correlations, i.e., $\theta_i = \theta_j$ for some $i \neq j$. We can derive the covering number of Θ as follows.

Lemma 17 (covering number of a hypercube with some correlations)**.** Let $\Theta := \{\ (\theta_1, \theta_2, \ldots, \theta_m)\ |\ \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R}\ \} \subset \mathbb{R}^m$ in which some parameters have correlations, i.e., $\theta_1 = k\,\theta_2$

with $k \in \mathbb{R}$. Then, the covering number of Θ w.r.t. the ∞-norm is upper bounded as

$$N(\Theta, \varepsilon, \|\cdot\|_\infty) \leq \prod_{j=1}^{T} \left\lceil \frac{c_{i_j}}{2\varepsilon} \right\rceil \tag{82}$$

, where T is the number of independent parameters. c_{i_j} is the largest one in those of dependent parameters.

Proof. You can understand Lemma 17 from the following example in $\mathbb{R}^3$. When all parameters are independent, then the covering number of Θ is equal to $\lceil \frac{c_1}{2\varepsilon} \rceil \times \lceil \frac{c_2}{2\varepsilon} \rceil \times \lceil \frac{c_3}{2\varepsilon} \rceil$. However, when $\theta_2 = \theta_3$, Θ is a plane in $\mathbb{R}^3$, and you only need to consider the covering number of the plane. As shown in Fig. 9, you can align the ∞-norm ball with the plane, and see that you can cover the plane with the same covering number of the hypercube $[a_1, a_1+c_1]\times[a_2, a_2+c_2]$ in $\mathbb{R}^2$, which is $\lceil \frac{c_1}{2\varepsilon} \rceil \times \lceil \frac{c_2}{2\varepsilon} \rceil$.

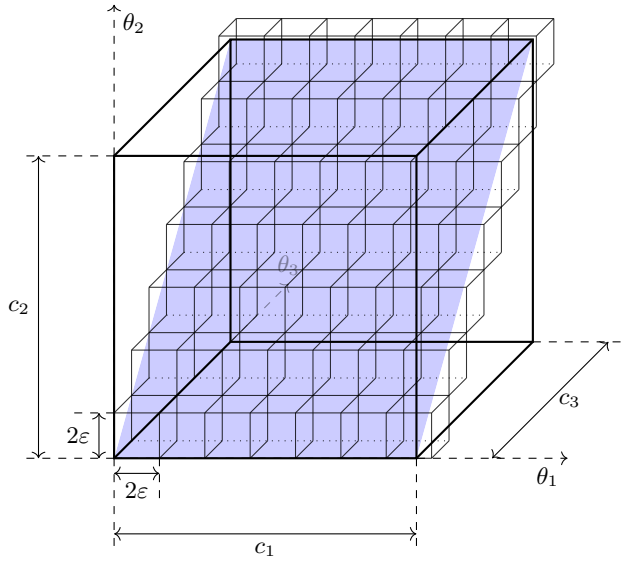

FIG. 9. The blue area represents $\Theta \coloneqq \{ (\theta_1, \theta_2, \theta_3) \mid \theta_2 = \theta_3, \theta_i \in [a_i, a_i + c_i] \} \subset \mathbb{R}^3$, and the small cubes represent the ε-covering net of Θ w.r.t. $\|\cdot\|_\infty$

□

Theorem 1 (Covering number of general VQML model, generalized from Theorem 3.6 in Ref. [4])**.** Let $\Theta \coloneqq \{ (\theta_1, \theta_2, \ldots, \theta_m) \mid \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R} \} \subset \mathbb{R}^m$, and $\mathcal{E}_\Theta$ be the set of all possible VQML channels $\mathcal{E}_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \Theta$. And some parameters may have correlations, i.e., $\theta_1 = k\,\theta_2$ with $k \in \mathbb{R}$. Then, the covering number of $\mathcal{E}_\Theta$ w.r.t. the diamond norm is upper bounded as

$$N\left(\mathcal{E}_\Theta, \varepsilon, \|\cdot\|_\diamond\right) \leq \prod_{j=1}^{T} \left\lceil \frac{m c_{i_j}}{2\varepsilon} \right\rceil \tag{83}$$

, where T is the number of independent parameters. c_{i_j} is the largest one in those of dependent parameters.

Proof. The map $\mathcal{E}_{\boldsymbol{\theta}}$ can be expressed as the product of sequential trainable maps and some fixed maps, so we have

$$\left\|\mathcal{E}_{\boldsymbol{\theta}} - \mathcal{E}_{\tilde{\boldsymbol{\theta}}}\right\|_\diamond \coloneqq \left\| \bigcirc_{l=1}^{L} \left(\bigcirc_{m=1}^{N_q} \mathcal{R}_{l,m}(\theta_{l,m}) \right) \mathcal{V}_l - \bigcirc_{l=1}^{L} \left(\bigcirc_{m=1}^{N_q} \mathcal{R}_{l,m}(\tilde{\theta}_{l,m}) \right) \mathcal{V}_l \right\|_\diamond \tag{84}$$

$$\overset{(1)}{\leq} \sum_{l=1}^{L} \left\| \bigcirc_{m=1}^{N_q} \mathcal{R}_{l,m}(\theta_{l,m}) - \bigcirc_{m=1}^{N_q} \mathcal{R}_{l,m}(\tilde{\theta}_{l,m}) \right\|_\diamond \tag{85}$$

$$\overset{(2)}{\leq} \sum_{l=1}^{L} \sum_{m=1}^{N_q} \left\| \mathcal{R}_{l,m}(\theta_{l,m}) - \mathcal{R}_{l,m}(\tilde{\theta}_{l,m}) \right\|_\diamond \tag{86}$$

$$\overset{(3)}{\leq} 2\sum_{l=1}^{L}\sum_{m=1}^{N_q} \|R_{l,m}(\theta_{l,m}) - R_{l,m}(\tilde{\theta}_{l,m})\|_{\infty} \tag{87}$$

$$\overset{(4)}{\leq} \sum_{l=1}^{L}\sum_{m=1}^{N_q} |\theta_{l,m} - \tilde{\theta}_{l,m}| \tag{88}$$

$$=: \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1 \tag{89}$$

, where $\mathcal{R}_{l,m}(\theta_{l,m}), \mathcal{V}_l$ are the channels corresponding to the rotation gate and fixed gate $R_{l,m}(\theta_{l,m}), V_l$, respectively. Here, inequality (1),(2) is from Lemma 6, and inequality (3) is from Lemma 7. Inequality (4) is from Lemma 8. Thus, by Lemma 13, 12, and 17, we have

$$N(\mathcal{E}_\Theta, \varepsilon, \|\cdot\|_\diamond) \leq N(\Theta, \varepsilon, \|\cdot\|_1) \leq N(\Theta, \varepsilon/m, \|\cdot\|_\infty) = \prod_{j=1}^{T} \left\lceil \frac{mc_{i_j}}{2\varepsilon} \right\rceil \tag{90}$$

□

Remark 9.

- In Lemma 1, we lose the Information about the structure of the ansatz. If we know some properties on the eigenvalues of $U(\boldsymbol{\theta}) - U(\tilde{\boldsymbol{\theta}})$, we may be able to derive an tighter upper bound, using $\left\|\mathcal{E}_{\boldsymbol{\theta}} - \mathcal{E}_{\tilde{\boldsymbol{\theta}}}\right\|_\diamond \leq 2\|U(\boldsymbol{\theta}) - U(\tilde{\boldsymbol{\theta}})\|_\infty$
- More generally, we can consider the covering number of unitary group $U(N)$ with respect to the infinity norm as in Lemma 1 of Ref. [7]

VI. Probability

Lemma 18 (Union bound)**.** Let $A_1, \ldots, A_N$ be events in a probability space. Then,

$$\mathbb{P}\left[\bigcup_{i=1}^{N} A_i\right] \leq \sum_{i=1}^{N} \mathbb{P}[A_i]. \tag{91}$$

Lemma 19 (McDiarmid's inequality)**.** Let $Z = (X_1, \ldots, X_N)$ be a finite set of independent $\mathbb{R}$-valued random variables and $\phi : \mathbb{R}^N \to \mathbb{R}$ be a measurable function such that, for every $1 \leq i \leq N$, there exists a constant $c_i > 0$ such that $|\phi(z) - \phi(z')| \leq c_i$ for all $z, z' \in \mathbb{R}^N$ that differ only in the i-th coordinate. Then, for every $\varepsilon > 0$,

$$\mathbb{P}\left[\phi(Z) - \mathbb{E}[\phi(Z)] \geq \varepsilon\right] \leq \exp\left(-2\varepsilon^2 / \sum_{i=1}^{N} c_i^2\right), \tag{92}$$

$$\mathbb{P}\left[\mathbb{E}[\phi(Z)] - \phi(Z) \geq \varepsilon\right] \leq \exp\left(-2\varepsilon^2 / \sum_{i=1}^{N} c_i^2\right), \tag{93}$$

$$\mathbb{P}\left[|\phi(Z) - \mathbb{E}[\phi(Z)]| \geq \varepsilon\right] \leq 2\exp\left(-2\varepsilon^2 / \sum_{i=1}^{N} c_i^2\right). \tag{94}$$

VII. Rademacher complexity

A. Notation

- $\mathcal{X}, \mathcal{Y}$: input and output spaces
- $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$
- $\mathcal{S} \coloneqq \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\} \subset \mathcal{Z}$: : a set of N training samples
- $\mathcal{F} = \{\ f : \mathcal{X} \to [0, C]\ \} \subset \mathcal{Y}^{\mathcal{X}}$: hypotheses class. $\mathcal{Y}^{\mathcal{X}}$ denotes the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$

- $\mathcal{L}(y, f(\boldsymbol{x})) : \mathcal{Y} \times \mathcal{Y} \to [0, LC]$: loss function. L is the Lipschitz constant of the loss function
- $\hat{R}_N(f) \coloneqq \frac{1}{N}\sum_{i=1}^N \mathcal{L}(y_i, f(\boldsymbol{x}_i))$: empirical risk of a hypothesis $f \in \mathcal{F}$ over a sample $\mathcal{S}$
- $R(f) \coloneqq \mathbb{E}_{(\boldsymbol{x},y)\sim P}[\mathcal{L}(y, f(\boldsymbol{x}))]$: expected risk of a hypothesis $f \in \mathcal{F}$
- $\mathcal{G} \coloneqq \{(\boldsymbol{x}, y) \mapsto \mathcal{L}(y, f(\boldsymbol{x})) \mid f \in \mathcal{F}\} \subseteq [0, LC]^{\mathcal{Z}}$: loss function class
- $\mathcal{F} \bullet \mathcal{S} \coloneqq \{\ (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)) \mid f \in \mathcal{F}, (\boldsymbol{x}_i, y_i) \in \mathcal{S}\ \} \subset [0, C]^N$: prediction evaluated on a sample $\mathcal{S}$
- $\mathcal{G} \bullet \mathcal{S} \coloneqq \{\ (\mathcal{L}(y_1, f(\boldsymbol{x}_1)), \ldots, \mathcal{L}(y_N, f(\boldsymbol{x}_N))) \mid f \in \mathcal{F}, (\boldsymbol{x}_i, y_i) \in \mathcal{S}\ \} \subset [0, LC]^N$: loss function evaluated on a sample $\mathcal{S}$

B. Definition of Rademacher complexity

Definition 8 (Rademacher complexity of a class of functions)**.** Given a set $\mathcal{S} = \{z_1, z_2, \ldots, z_N\} \subset \mathcal{Z}^N$, the empirical Rademacher complexity of $\mathcal{G}$ w.r.t. $\mathcal{S}$ is defined as

$$\mathfrak{R}(\mathcal{G} \bullet \mathcal{S}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{i=1}^N \sigma_i g(z_i)\right] \tag{95}$$

, where $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_N)$ is a vector of independent uniform random variables taking values in $\{-1, +1\}$.

Remark 10. Rademacher complexity quantifies the capacity of the class to fit random labels. $\mathfrak{R}(\mathcal{G} \bullet \mathcal{S})$ is small if the class $\mathcal{G}$ is simple and large if the class is complex. This is exlpained as follows: to make the supremum large, the class $\mathcal{G}$ should have a function g that takes a large value at z_i for $\sigma_i = 1$ and a small value for $\sigma_i = -1$ for as many i and $\boldsymbol{\sigma}$ as possible.

If the class $\mathcal{G}\ (\subset [0, b]^{\mathcal{Z}})$ is complex enough, there exists a function g that takes b for $\sigma_i = 1$ and 0 for $\sigma_i = -1$ for all i and for all $\boldsymbol{\sigma}$. In this situation, when $\boldsymbol{\sigma} = (+1, +1, \ldots, +1)$, the supremum becomes b, and when $\boldsymbol{\sigma} = (-1, -1, \ldots, -1)$, the supremum becomes 0. And for any $\boldsymbol{\sigma}$, $\sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{i=1}^N \sigma_i g(z_i) + \sup_{g'\in\mathcal{G}} \frac{1}{N}\sum_{i=1}^N (-\sigma_i) g'(z_i) = b$. Therefore, the Rademacher complexity is $b/2$ in this case.

On the other hand, if the class $\mathcal{G}$ is the simplest – i.e., $\mathcal{G}$ is a singleton set –,

$$\mathfrak{R}(\mathcal{G} \bullet \mathcal{S}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\frac{1}{N}\sum_{i=1}^N \sigma_i g(z_i)\right] = \frac{1}{N}\sum_{i=1}^N g(z_i) \mathop{\mathbb{E}}_{\sigma_i\in\{\pm1\}}[\sigma_i] = 0. \tag{96}$$

Theorem 2 (Properties of Rademacher complexity, from Theorem 1.16 in [5])**.** Let $\mathcal{G}, \mathcal{G}_1, \mathcal{G}_2 \subseteq [0, c]^{\mathcal{Z}}$ be classes of real-valued functions on $\mathcal{Z}$, and $\mathcal{S} = \{z_1, z_2, \ldots, z_N\} \subset \mathcal{Z}^N$. Then, the Rademacher complexity of $\mathcal{G}, \mathcal{G}_1, \mathcal{G}_2$ w.r.t. $\mathcal{S}$ satisfies the following properties:

1. If $c \in \mathbb{R}$ is a constant, then $\mathfrak{R}((c\,\mathcal{G}) \bullet \mathcal{S}) = |c|\mathfrak{R}(\mathcal{G} \bullet \mathcal{S})$, where $c\,\mathcal{G} \coloneqq \{\ cg \mid g \in \mathcal{G}\ \}$
2. $\mathfrak{R}(\mathcal{G} \bullet \mathcal{S}) \le \mathfrak{R}(\mathcal{G}_1 \bullet \mathcal{S})$ for any $\mathcal{G} \subset \mathcal{G}_1$
3. $\mathfrak{R}((\mathcal{G}_1 + \mathcal{G}_2) \bullet \mathcal{S}) = \mathfrak{R}(\mathcal{G}_1 \bullet \mathcal{S}) + \mathfrak{R}(\mathcal{G}_2 \bullet \mathcal{S})$, where $\mathcal{G}_1 + \mathcal{G}_2 \coloneqq \{\ g_1(z) + g_2(z) \mid g_1 \in \mathcal{G}_1,\ g_2 \in \mathcal{G}_2\ \}$

Lemma 20. Let $\delta > 0$ and suppose P is a probability measure on the space $\mathcal{Z}$. With probability at least $1 - \delta$, when drawing training data $\mathcal{S} \in \mathcal{Z}^N$ from the distribution P^N, the following inequality holds for any function class $\mathcal{G}$:

$$\mathbb{E}_{\mathcal{S}\sim P^N}[\mathfrak{R}(\mathcal{G} \bullet \mathcal{S})] \le \mathfrak{R}(\mathcal{G} \bullet \mathcal{S}) + LC\sqrt{\frac{\log\frac{1}{\delta}}{2N}} \tag{97}$$

Proof. Let $\phi(\mathcal{S}) = \mathfrak{R}(\mathcal{G} \bullet \mathcal{S})$. The following inequality holds for any $\mathcal{S}, \mathcal{S}' \in \mathcal{Z}^N$ that differ in only j-th element:

$$|\phi(\mathcal{S}) - \phi(\mathcal{S}')| \le \left|\sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{i=1}^N \sigma_i g(z_i) - \sup_{g\in\mathcal{G}} \frac{1}{N}\sum_{i=1}^N \sigma_i g(z'_i)\right| \tag{98}$$

$$\leq \left| \sup_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \{g(z_i) - g(z_i')\} \right| \tag{99}$$

$$= \left| \sup_{g \in \mathcal{G}} \frac{1}{N} \sigma_j \{g(z_j) - g(z_j')\} \right| \tag{100}$$

$$\leq \frac{LC}{N} \quad (\text{since } g \in \mathcal{G} \subset [0, LC]^{\mathcal{Z}}) \tag{101}$$

From this, we can apply McDiarmid's inequality 19 to $\phi(\mathcal{S})$ to obtain

$$\mathbb{P}[\mathbb{E}_{\mathcal{S} \sim P^N}[\phi(\mathcal{S})] - \phi(\mathcal{S}) \geq \varepsilon] \leq \exp\left(-\frac{2N\varepsilon^2}{L^2 C^2}\right) \tag{102}$$

Setting the r.h.s. equal to δ and solving for ε then gives that with probability at least $1 - \delta$ we have

$$\mathbb{E}_{\mathcal{S} \sim P^N}[\mathfrak{R}(\mathcal{G} \bullet \mathcal{S})] \leq \mathfrak{R}(\mathcal{G} \bullet \mathcal{S}) + LC\sqrt{\frac{\log \frac{1}{\delta}}{2N}} \tag{103}$$

□

The following lemma gives an important upper bound on the generalization error of a hypothesis f in terms of the Rademacher complexity of $\mathcal{G} \bullet \mathcal{S}$.

Lemma 21 (Connection between Generalization Error and Rademacher Complexity)**.** Let $\delta > 0$ and suppose P is a probability measure on the space $\mathcal{Z}$. With probability at least $1 - \delta$, when drawing training data $\mathcal{S} \in \mathcal{Z}^N$ from the distribution P^N, the following inequality holds for any function $f \in \mathcal{F}$:

$$R(f) - \hat{R}_N(f) \leq 2\mathfrak{R}(\mathcal{G} \bullet \mathcal{S}) + 3LC\sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \tag{104}$$

$$R(f) - \hat{R}_N(f) \leq 2\mathbb{E}_{\mathcal{S} \sim P^N}[\mathfrak{R}(\mathcal{G} \bullet \mathcal{S})] + LC\sqrt{\frac{\log \frac{1}{\delta}}{2N}} \tag{105}$$

Proof. Let $\phi(\mathcal{S}) = \sup_{f \in \mathcal{F}} R(f) - \hat{R}_N(f)$. The following inequality holds for any $\mathcal{S}, \mathcal{S}' \in \mathcal{Z}^N$ that differ in only j-th element:

$$|\phi(\mathcal{S}) - \phi(\mathcal{S}')| \leq |\sup_{f \in \mathcal{F}} \{R(f) - \hat{R}_N(f)\} - \sup_{f \in \mathcal{F}} \{R(f) - \hat{R}_{\mathcal{S}'}(f)\}| \tag{106}$$

$$\overset{(a)}{\leq} |\sup_{f \in \mathcal{F}} \{R(f) - \hat{R}_N(f) - R(f) + \hat{R}_{\mathcal{S}'}(f)\}| \tag{107}$$

$$\leq |\sup_{f \in \mathcal{F}} \{\hat{R}_{\mathcal{S}'}(f) - \hat{R}_N(f)\}| \tag{108}$$

$$\leq |\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \{\mathcal{L}(y_i', f(\boldsymbol{x}_i')) - \mathcal{L}(y_i, f(\boldsymbol{x}_i))\}| \tag{109}$$

$$= |\sup_{f \in \mathcal{F}} \frac{1}{N} \{\mathcal{L}(y_j', f(\boldsymbol{x}_j')) - \mathcal{L}(y_j, f(\boldsymbol{x}_j))\}| \tag{110}$$

$$\leq \frac{LC}{N} \tag{111}$$

, where we used the fact that $\sup_x\{f(x)\} - \sup_x\{g(x)\} \leq \sup_x\{f(x) - g(x)\}$ in (a). From this, we can apply McDiarmid's inequality 19 to $\phi(\mathcal{S})$ to obtain

$$\mathbb{P}[\phi(\mathcal{S}) - \mathbb{E}_{\mathcal{S} \sim P^N}[\phi(\mathcal{S})] \geq \varepsilon] \leq \exp\left(-\frac{2N\varepsilon^2}{L^2 C^2}\right) \tag{112}$$

Setting the r.h.s. equal to δ and solving for ε then gives that with probability at least $1-\delta$ we have

$$\sup_{f\in\mathcal{F}} R(f) - \hat{R}_N(f) \leq \mathbb{E}_{\mathcal{S}\sim P^N}\left[\sup_{f\in\mathcal{F}} R(f) - \hat{R}_N(f)\right] + LC\sqrt{\frac{\log\frac{1}{\delta}}{2N}} \tag{113}$$

In order to bound the expectation term, we define $\mathcal{S}' = \{z'_1, z'_2, \ldots, z'_N\} \in \mathcal{Z}^N$ as a new sample drawn i.i.d. from P^N. Then, we have

$$\mathbb{E}_{\mathcal{S}\sim P^N}\left[\sup_{f\in\mathcal{F}}\{R(f) - \hat{R}_N(f)\}\right] \tag{114}$$

$$= \mathbb{E}_{\mathcal{S}\sim P^N}\left[\sup_{f\in\mathcal{F}}\{\mathbb{E}_{\mathcal{S}'\sim P^N}[\hat{R}_{\mathcal{S}'}(f)] - \hat{R}_N(f)\}\right] \tag{115}$$

$$\overset{(a)}{\leq} \mathbb{E}_{\mathcal{S},\mathcal{S}'\sim P^N}[\sup_{f\in\mathcal{F}}\{\hat{R}_{\mathcal{S}'}(f) - \hat{R}_N(f)\}] \tag{116}$$

$$= \mathbb{E}_{\mathcal{S},\mathcal{S}'\sim P^N}\left[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\{\mathcal{L}(y'_i, f(\boldsymbol{x}'_i)) - \mathcal{L}(y_i, f(\boldsymbol{x}_i))\}\right] \tag{117}$$

$$\overset{(b)}{=} \mathbb{E}_{\mathcal{S},\mathcal{S}'\sim P^N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\{\mathcal{L}(y'_i, f(\boldsymbol{x}'_i)) - \mathcal{L}(y_i, f(\boldsymbol{x}_i))\}\right] \tag{118}$$

$$\overset{(c)}{\leq} \mathbb{E}_{\mathcal{S},\mathcal{S}'\sim P^N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}(y'_i, f(\boldsymbol{x}'_i)) + \sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}(-\sigma_i)\mathcal{L}(y_i, f(\boldsymbol{x}_i))\right] \tag{119}$$

$$= \mathbb{E}_{\mathcal{S},\mathcal{S}'\sim P^N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}(y'_i, f(\boldsymbol{x}'_i)) + \sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}(y_i, f(\boldsymbol{x}_i))\right] \tag{120}$$

$$= 2\mathbb{E}_{\mathcal{S}\sim P^N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\mathcal{L}(y_i, f(\boldsymbol{x}_i))\right] \tag{121}$$

$$= 2\mathbb{E}_{\mathcal{S}\sim P^N}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\mathfrak{R}(\mathcal{G}\bullet\mathcal{S}) \tag{122}$$

, where (a) comes from the fact that $\sup_x \mathbb{E}[f(x)] \leq \mathbb{E}[\sup_x f(x)]$, (b) comes from the fact that $\mathcal{S}, \mathcal{S}' \sim P^N$, (c) comes from the fact that $\sup_x\{f(x)+g(x)\} \leq \sup_x f(x) + \sup_x g(x)$

Therefore, with probability at least $1-\delta$, we have

$$R(f) - \hat{R}_N(f) \leq 2\mathbb{E}_{\mathcal{S}\sim P^N}[\mathfrak{R}(\mathcal{G}\bullet\mathcal{S})] + LC\sqrt{\frac{\log\frac{1}{\delta}}{2N}} \tag{123}$$

For the second inequality, we use the union bound 18 over the two events that the inequalities (97) and (123) hold with probability at least $1-\delta/2$ each to obtain

$$R(f) - \hat{R}_N(f) \leq 2\mathfrak{R}(\mathcal{G}\bullet\mathcal{S}) + 3LC\sqrt{\frac{\log\frac{2}{\delta}}{2N}} \tag{124}$$

□

Lemma 22 (Talagrand's lemma from Ref. [13])**.** Let $\Phi_1, \ldots, \Phi_N$ be L-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$ and $(\sigma_1, \ldots, \sigma_N)$ be Rademacher random variables. Then, for any hypothesis set $\mathcal{F}$ of real-valued functions, the following inequality holds:

$$\frac{1}{N}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N}\sigma_i(\Phi_i\circ f)(\boldsymbol{x}_i)\right] \leq \frac{L}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N}\sigma_i f(\boldsymbol{x}_i)\right] = L\,\mathfrak{R}(\mathcal{F}\bullet\mathcal{S})$$

Remark 11. Let $\Phi_i = \mathcal{L}(y_i, \cdot)$ and suppose that $\mathcal{L}$ is L-Lipschitz w.r.t. the second argument. Then, we have

$$\mathfrak{R}(\mathcal{G} \bullet \mathcal{S}) \leq L\, \mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) \tag{125}$$

Thus, (124) can be rewritten as

$$R(f) - \hat{R}_N(f) \leq 2L\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) + 3LC\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \tag{126}$$

More generally, as an analogy to the definition of the Rademacher complexity of a class of functions, we can define the Rademacher complexity of a set $\mathcal{A} \subset \mathbb{R}^N$.

Definition 9 (Rademacher complexity of a set $\mathcal{A} \subset \mathbb{R}^N$)**.** Given a set $\mathcal{A} \subset \mathbb{R}^N$, the Rademacher complexity of $\mathcal{A}$ is defined as

$$\mathfrak{R}(\mathcal{A}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \in \{\pm 1\}^N} \left[\sup_{\boldsymbol{a} \in \mathcal{A}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i a_i \right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \in \{\pm 1\}^N} \left[\sup_{\boldsymbol{a} \in \mathcal{A}} \frac{\langle \boldsymbol{\sigma}, \boldsymbol{a} \rangle}{N} \right] \tag{127}$$

The following lemma gives an important upper bound on the Rademacher complexity of a set $\mathcal{A} \subset \mathbb{R}^N$ in terms of the cardinality of $\mathcal{A}$.

Lemma 23 (Massart's lemma)**.** Let $\mathcal{A}$ be a finite subset of $\mathbb{R}^N$. Then, the Rademacher complexity of $\mathcal{A}$ is upper bounded as

$$\mathfrak{R}(\mathcal{A}) \leq \max_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_2 \frac{\sqrt{2 \log |\mathcal{A}|}}{N} \tag{128}$$

Lemma 24 (Lemma 26.10 in Ref. [8])**.** Let $\mathcal{S} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ be vectors in Hilbert space. We define $\mathcal{F} \bullet \mathcal{S} = \{(\langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle, \ldots, \langle \boldsymbol{w}, \boldsymbol{x}_N \rangle) : \|\boldsymbol{w}\|_2 \leq c\}$ for some $c > 0$. Then, the Rademacher complexity of $\mathcal{F} \bullet \mathcal{S}$ is upper bounded as

$$\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) \leq \frac{c \max_i \|\boldsymbol{x}_i\|_2}{\sqrt{N}} \tag{129}$$

Proof.

$$\begin{aligned}
N\mathfrak{R}(\mathcal{H}_2 \circ \mathcal{S}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{a} \in \mathcal{H}_2 \circ \mathcal{S}} \sum_{i=1}^{N} \sigma_i a_i \right] && (130)\\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{w}: \|\boldsymbol{w}\|_2 \leq c} \sum_{i=1}^{N} \sigma_i \langle \boldsymbol{w} | \boldsymbol{x}_i \rangle \right] && (131)\\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{w}: \|\boldsymbol{w}\|_2 \leq c} \left\langle \boldsymbol{w} | \Sigma_{i=1}^{N} \sigma_i \boldsymbol{x}_i \right\rangle \right] && (132)\\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{w}: \|\boldsymbol{w}\|_2 \leq c} \|\boldsymbol{w}\| \left\| \sum_{i=1}^{N} \sigma_i \boldsymbol{x}_i \right\|_2 \right] \quad \text{(by Cauchy-Schwarz inequality)} && (133)\\
&\leq c \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left\| \sum_{i=1}^{N} \sigma_i \boldsymbol{x}_i \right\|_2 \right] && (134)
\end{aligned}$$

Next, using Jensen's inequality we have that

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left\| \sum_{i=1}^{N} \sigma_i \boldsymbol{x}_i \right\|_2 \right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left(\left\| \sum_{i=1}^{N} \sigma_i \boldsymbol{x}_i \right\|_2^2 \right)^{1/2} \right] \leq \left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left\| \sum_{i=1}^{N} \sigma_i \boldsymbol{x}_i \right\|_2^2 \right] \right)^{1/2} \tag{135}$$

And, since the variables $\sigma_1, \ldots, \sigma_m$ are independent we have

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{N}\sigma_i \boldsymbol{x}_i\right\|_2^2\right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_i\sigma_j \langle \boldsymbol{x}_i|\boldsymbol{x}_j\rangle\right] \tag{136}$$

$$= \sum_{i=1}^{N}\langle \boldsymbol{x}_i|\boldsymbol{x}_i\rangle \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_i^2] + \sum_{i\neq j}\langle \boldsymbol{x}_i|\boldsymbol{x}_j\rangle \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_i\sigma_j] \tag{137}$$

$$= \sum_{i=1}^{N}\|\boldsymbol{x}_i\|_2^2 \tag{138}$$

$$\leq N \max_i \|\boldsymbol{x}_i\|_2^2 \tag{139}$$

Therefore, we have

$$N\mathfrak{R}(\mathcal{F}\bullet\mathcal{S}) \leq c\sqrt{N\max_i\|\boldsymbol{x}_i\|_2^2} \iff \mathfrak{R}(\mathcal{F}\bullet\mathcal{S}) \leq \frac{c\max_i\|\boldsymbol{x}_i\|_2}{\sqrt{N}} \tag{140}$$

□

We can derive the upper bound of Rademacher complexity of quantum circuit learning model by using the same technique as in the proof of Lemma 24.

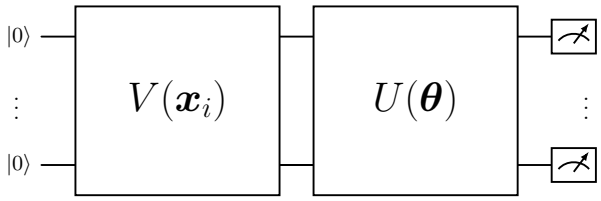

FIG. 10. Quantum circuit learning model

Theorem 3. The Rademacher complexity of quantum circuit learning model is upper bounded as

$$\mathfrak{R}(\mathcal{F}\bullet\mathcal{S}) \leq \|O\|_1\sqrt{\frac{2n\log 2}{N}} \tag{141}$$

$$\mathfrak{R}(\mathcal{F}\bullet\mathcal{S}) \leq \|O\|_2\sqrt{\frac{1}{N}} \tag{142}$$

$$\mathfrak{R}(\mathcal{F}\bullet\mathcal{S}) \leq \|O\|_\infty\sqrt{\frac{2^n}{N}} \tag{143}$$

Proof. We consider the Rademacher complexity of Quantum circuit learning models. Let $\mathcal{F}$ be a class of quantum machine learning models, and $\mathcal{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be a set of classical data. We define $\mathcal{F}\bullet\mathcal{S} = \{f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N) : f \in \mathcal{F}\}$. Then, the expectation of an observable O is given by

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \coloneqq \mathrm{Tr}\left[OU(\boldsymbol{\theta})V(\boldsymbol{x}_i)\,|\boldsymbol{0}\rangle\langle\boldsymbol{0}|\,V^\dagger(\boldsymbol{x}_i)U^\dagger(\boldsymbol{\theta})\right] \tag{144}$$

$$=: \mathrm{Tr}[O(\boldsymbol{\theta})\rho(\boldsymbol{x}_i)] \tag{145}$$

Firstly, we bound $\mathfrak{R}(\mathcal{F}\circ\mathcal{S})$ as follows:

$$N\mathfrak{R}(\mathcal{F}\circ\mathcal{S}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{a}\in\mathcal{F}\circ\mathcal{S}}\sum_{i=1}^{N}\sigma_i a_i\right] \tag{146}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{\theta}}\sum_{i=1}^{N}\sigma_i\,\mathrm{Tr}[O(\boldsymbol{\theta})\rho(\boldsymbol{x}_i)]\right] \tag{147}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\theta}} \mathrm{Tr}\left[O(\boldsymbol{\theta}) (\Sigma_{i=1}^{N} \sigma_i \rho(\boldsymbol{x}_i)) \right] \right] \tag{148}$$

$$\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\theta}} \|O(\boldsymbol{\theta})\|_p \left\| \Sigma_{i=1}^{N} \sigma_i \rho(\boldsymbol{x}_i) \right\|_q \right] \quad \text{(by Hölder's inequality)} \tag{149}$$

$$\leq \|O\|_p \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left\| \Sigma_{i=1}^{N} \sigma_i \rho(\boldsymbol{x}_i) \right\|_q \right] \quad (\because \|O(\boldsymbol{\theta})\|_p = \|O\|_p) \tag{150}$$

When $p = 1$ and $q = \infty$, we have

$$N\mathfrak{R}(\mathcal{F} \circ \mathcal{S}) \leq \|O\|_1 \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left\| \Sigma_{i=1}^{N} \sigma_i \rho(\boldsymbol{x}_i) \right\|_\infty \right] \tag{151}$$

$$\overset{(c)}{\leq} \|O\|_1 \sqrt{\left\| \Sigma_{i=1}^{N} \rho^2(\boldsymbol{x}_i) \right\|_\infty 2 \log 2^n} \tag{152}$$

$$= \|O\|_1 \sqrt{2n \log 2 \, \Sigma_{i=1}^{N} \|\rho^2(\boldsymbol{x}_i)\|_\infty} \tag{153}$$

$$\leq \|O\|_1 \sqrt{2Nn \log 2} \tag{154}$$

, where (c) comes from Theorem 4.6.1 in Ref. [14]. Therefore, we have

$$\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) \leq \|O\|_1 \sqrt{\frac{2n \log 2}{N}} \tag{155}$$

When $p = q = 2$, we have

$$N\mathfrak{R}(\mathcal{F} \circ \mathcal{S}) \leq \|O\|_2 \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left\| \Sigma_{i=1}^{N} \sigma_i \rho(\boldsymbol{x}_i) \right\|_2 \right] \tag{156}$$

$$\leq \|O\|_2 \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left\| \Sigma_{i=1}^{N} \sigma_i \rho(\boldsymbol{x}_i) \right\|_2^2 \right]} \quad \text{(by Jensen's inequality)} \tag{157}$$

$$\leq \|O\|_2 \sqrt{N \max_i \|\rho(\boldsymbol{x}_i)\|_2^2} \tag{158}$$

Since $\|\rho(\boldsymbol{x}_i)\|_2 = \sqrt{\mathrm{Tr}[(\rho(\boldsymbol{x}_i))^2]} = 1$, we have

$$\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) \leq \frac{\|O\|_2}{\sqrt{N}} = \sqrt{\frac{\mathrm{Tr}[O^2]}{N}} \tag{159}$$

When $p = \infty$ and $q = 1$, we have

$$N\mathfrak{R}(\mathcal{F} \circ \mathcal{S}) \leq \|O\|_\infty \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\left\| \Sigma_{i=1}^{N} \sigma_i \rho(\boldsymbol{x}_i) \right\|_1 \right] \tag{160}$$

$$\overset{(a)}{\leq} \|O\|_\infty \left\| \sqrt{\sum_{i=1}^{N} \rho^2(\boldsymbol{x}_i; \boldsymbol{\theta})} \right\|_1 \tag{161}$$

$$= \|O\|_\infty \, \mathrm{Tr}\left[\sqrt{\sum_{i=1}^{N} \rho^2(\boldsymbol{x}_i; \boldsymbol{\theta})} \right] \tag{162}$$

$$\overset{(b)}{\leq} \|O\|_\infty \sqrt{2^n N} \tag{163}$$

, where (a) comes from the operator Khintchine inequality and (b) comes from the fact that $(\mathrm{Tr}[\sqrt{A}])^2 \leq \dim(A) \, \mathrm{Tr}[A]$ for any operator A. Therefore, we have

$$\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) \leq \|O\|_\infty \sqrt{\frac{2^n}{N}} \tag{164}$$

□

Remark 12 (Theorem 3)**.**

- [15]
- This result may be the same as [16], which investigates the generalization error of quantum kernel methods. So, the generalization error of the quantum circuit learning model and the quantum kernel method may be the same.
- [17]
- This example seems inconsistent with the result of [18], in which the generalization error exhibits a U-shaped curve with respect to the number of parameters.

C. Rademacher complexity bound: One step discretization

If $|\mathcal{A}| = \infty$ (or $\mathcal{A}$ is a finite but very large set) then the bound from Massart's lemma 23 is not useful. We can overcome this problem by approximating the large set $\mathcal{A}$ with a much smaller set $\mathcal{C}$, which is an ε-covering net of $\mathcal{A}$.

Theorem 4. For $\mathcal{A} \subset \mathbb{R}^N$, the Rademacher complexity $\mathfrak{R}(\mathcal{A})$ is upper bounded by the covering number of $\mathcal{A}$ as

$$\mathfrak{R}(\mathcal{A}) = \inf_{\varepsilon>0} \left\{ \varepsilon + \max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_2 \frac{\sqrt{2\log N_{\mathrm{in}}(\mathcal{A}, \varepsilon, \|\cdot\|_{p,N})}}{N} \right\} \tag{165}$$

$$= \inf_{\varepsilon>0} \left\{ \varepsilon + \max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_2 \frac{\sqrt{2\log N(\mathcal{A}, \varepsilon/2, \|\cdot\|_{p,N})}}{N} \right\} \tag{166}$$

Proof. Let $\mathcal{C} \subset \mathcal{A}$ be an internal ε-covering net of $\mathcal{A}$ such that $|\mathcal{C}| = N_{\mathrm{in}}(\mathcal{A}, \varepsilon, \|\cdot\|_{p,N})$. For each $\boldsymbol{a} \in \mathcal{A}$, let $\pi(\boldsymbol{a}) = \boldsymbol{c}$ for $\boldsymbol{c} \in \mathcal{C}$ such that $\|\boldsymbol{a} - \boldsymbol{c}\|_{p,N} \leq \varepsilon$. By linearity of the inner product and Hölder's inequality 3, for any $\boldsymbol{a} \in \mathcal{A}$, $\boldsymbol{\sigma} \in \{\pm 1\}^N$ and $p, q \geq 1$ such that $1/p + 1/q = 1$, we have

$$\langle \boldsymbol{\sigma}, \boldsymbol{a} \rangle = \langle \boldsymbol{\sigma}, \pi(\boldsymbol{a}) \rangle + \langle \boldsymbol{\sigma}, \boldsymbol{a} - \pi(\boldsymbol{a}) \rangle \tag{167}$$

$$\leq \langle \boldsymbol{\sigma}, \pi(\boldsymbol{a}) \rangle + \|\boldsymbol{\sigma}\|_q \|a - \pi(\boldsymbol{a})\|_p \tag{168}$$

$$= \langle \boldsymbol{\sigma}, \pi(\boldsymbol{a}) \rangle + N^{\frac{1}{q}} \cdot N^{\frac{1}{p}} \varepsilon \tag{169}$$

$$= \langle \boldsymbol{\sigma}, \pi(\boldsymbol{a}) \rangle + N\varepsilon \tag{170}$$

Therefore,

$$\mathfrak{R}(\mathcal{A}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{a}\in\mathcal{A}} \frac{\langle \boldsymbol{\sigma}, \boldsymbol{a} \rangle}{N} \right] \tag{171}$$

$$\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{a}\in\mathcal{A}} \frac{\langle \boldsymbol{\sigma}, \pi(\boldsymbol{a}) \rangle + N\varepsilon}{N} \right] \tag{172}$$

$$= \varepsilon + \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{a}\in\mathcal{A}} \frac{\langle \boldsymbol{\sigma}, \pi(\boldsymbol{a}) \rangle}{N} \right] \tag{173}$$

$$= \varepsilon + \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{c}\in\mathcal{C}} \frac{\langle \boldsymbol{\sigma}, \boldsymbol{c} \rangle}{N} \right] \tag{174}$$

$$= \varepsilon + \mathfrak{R}(\mathcal{C}) \tag{175}$$

$$\leq \varepsilon + \max_{\boldsymbol{c}\in\mathcal{C}} \|\boldsymbol{c}\|_2 \frac{\sqrt{2\log|\mathcal{C}|}}{N} \tag{176}$$

$$\leq \varepsilon + \max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_2 \frac{\sqrt{2\log(N_{\mathrm{in}}(\mathcal{A}, \varepsilon, \|\cdot\|_{p,N}))}}{N} \tag{177}$$

$$\leq \varepsilon + \max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_2 \frac{\sqrt{2\log(N(\mathcal{A}, \varepsilon/2, \|\cdot\|_{p,N}))}}{N} \tag{178}$$

□

, where we used $N_{\mathrm{in}}(\mathcal{A}, \varepsilon, \|\cdot\|) \leq N(\mathcal{A}, \varepsilon/2, \|\cdot\|)$ in the last inequality.

D. Rademacher complexity bound: Chaining bound

You can get a tighter generalization bound by using the chaining method instead of the one-step discretization method.

Theorem 5 (Chaining bound)**.** Let $\mathcal{A} \subset \mathbb{R}^N$ be a set such that $\max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_p \leq r$. Then, the Rademacher complexity $\mathfrak{R}(\mathcal{A})$ is upper bounded by the covering number of $\mathcal{A}$ as

$$\mathfrak{R}(\mathcal{A}) \leq \inf_{\varepsilon\in(0,\frac{r}{2}]} \left\{ \frac{1}{N^{\frac{1}{p}}} \left(4\varepsilon + \frac{12}{\sqrt{N}} \int_{\varepsilon}^{\frac{r}{2}} \mathrm{d}\varepsilon' \sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_p)} \right) \right\} \tag{179}$$

Proof. For any $k \in \{0, 1, 2, \ldots, M\}$, let $\mathcal{C}_k \subset \mathbb{R}^N$ be an ε_k-covering net of $\mathcal{A}$ w.r.t. $\|\cdot\|_p$, where $\varepsilon_k \coloneqq r/2^k$ and $|\mathcal{C}_k| = N(\mathcal{A}, \varepsilon_k, \|\cdot\|_p)$. In particular, we can take $\mathcal{C}_0 \coloneqq \{\boldsymbol{0}\}$, because $\|\boldsymbol{a} - \boldsymbol{0}\|_p \leq r =: \varepsilon_0$ for all $\boldsymbol{a} \in \mathcal{A}$. For any k and any $\boldsymbol{a} \in \mathcal{A}$, let $\pi_k(\boldsymbol{a}) = \boldsymbol{c}$ such that $\boldsymbol{c} \in \mathcal{C}_k$ and $\|\boldsymbol{a} - \boldsymbol{c}\|_p \leq \varepsilon_k$. Furthermore, we define $\Delta_k(\boldsymbol{a}) = \pi_k(\boldsymbol{a}) - \pi_{k-1}(\boldsymbol{a})$. Then, for any $\boldsymbol{a} \in \mathcal{A}$, we have

$$\boldsymbol{a} = \boldsymbol{a} + \pi_0(\boldsymbol{a}) - \pi_M(\boldsymbol{a}) + \sum_{k=1}^{M} \Delta_k(\boldsymbol{a}) = \boldsymbol{a} - \pi_M(\boldsymbol{a}) + \sum_{k=1}^{M} \Delta_k(\boldsymbol{a}).$$

Hence,

$$\begin{aligned}
\mathfrak{R}(\mathcal{A}) &\coloneqq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{a}\in\mathcal{A}} \frac{\langle \boldsymbol{a}, \boldsymbol{\sigma} \rangle}{N} \right] && (180)\\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{a}\in\mathcal{A}} \left\{ \frac{\langle \boldsymbol{a} - \pi_M(\boldsymbol{a}), \boldsymbol{\sigma} \rangle}{N} + \frac{\langle \sum_{k=1}^{M} \Delta_k(\boldsymbol{a}), \boldsymbol{\sigma} \rangle}{N} \right\} \right] && (181)\\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{a}\in\mathcal{A}} \frac{\langle a - \pi_M(\boldsymbol{a}), \boldsymbol{\sigma} \rangle}{N} \right] + \sum_{k=1}^{M} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{a}\in\mathcal{A}} \frac{\langle \Delta_k(\boldsymbol{a}), \boldsymbol{\sigma} \rangle}{N} \right] \quad \because \sup \Sigma \leq \Sigma \sup && (182)\\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{a}\in\mathcal{A}} \frac{\langle a - \pi_M(\boldsymbol{a}), \boldsymbol{\sigma} \rangle}{N} \right] + \sum_{k=1}^{M} \mathfrak{R}(\Delta_k) && (183)
\end{aligned}$$

, where $\Delta_k \coloneqq \{\ \Delta_k(\boldsymbol{a}) \mid \boldsymbol{a} \in \mathcal{A}\ \}$.

For the first term, we use Hölder's inequality 3 to obtain

$$\langle \boldsymbol{a} - \pi_M(\boldsymbol{a}), \boldsymbol{\sigma} \rangle \leq \|\boldsymbol{a} - \pi_M(\boldsymbol{a})\|_p \|\boldsymbol{\sigma}\|_q \leq \varepsilon_M N^{\frac{1}{q}}.$$

Thus, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N} \left[\sup_{\boldsymbol{a}\in\mathcal{A}} \frac{\langle \boldsymbol{a} - \pi_M(\boldsymbol{a}), \boldsymbol{\sigma} \rangle}{N} \right] \leq \frac{\varepsilon_M}{N^{\frac{1}{p}}} \tag{184}$$

For the second term, we observe that

$$\begin{aligned}
|\Delta_k| &= |\{\pi_k(\boldsymbol{a}) - \pi_{k-1}(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{A}\}| \\
&\leq |\{\pi_k(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{A}\}| \cdot |\{\pi_{k-1}(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{A}\}| \\
&= N(\mathcal{A}, \varepsilon_k, \|\cdot\|_p) \cdot N(\mathcal{A}, \varepsilon_{k-1}, \|\cdot\|_p) \\
&\leq N(\mathcal{A}, \varepsilon_k, \|\cdot\|_p)^2
\end{aligned}$$

and for all $\boldsymbol{a} \in \mathcal{A}$,

$$\|\Delta_k(\boldsymbol{a})\|_p \leq \|\pi_k(\boldsymbol{a}) - \boldsymbol{a}\|_p + \|\boldsymbol{a} - \pi_{k-1}(\boldsymbol{a})\|_p \leq 3\varepsilon_k \tag{185}$$

and remember that $\|\boldsymbol{a}\|_p \le \|\boldsymbol{a}\|_2 \le N^{(1/2-1/p)}\|\boldsymbol{a}\|_p$ for all $1 \le r < p$, so we have

$$\|\Delta_k(\boldsymbol{a})\|_2 \le N^{(1/2-1/p)} \cdot 3\varepsilon_k$$

Thus, by Massart's lemma 23, we have

$$\mathfrak{R}(\Delta_k) \le \frac{N^{(1/2-1/p)} 3\varepsilon_k \sqrt{2\log|\Delta_k|}}{N} \le \frac{3\varepsilon_k}{N^{\frac{1}{p}}\sqrt{N}}\sqrt{2\log|\mathcal{C}_k|^2} = \frac{6\varepsilon_k}{N^{\frac{1}{p}}\sqrt{N}}\sqrt{\log N(\mathcal{A}, \varepsilon_k, \|\cdot\|_p)}$$

Therefore,

$$\sum_{k=1}^{M}\mathfrak{R}(\Delta_k) \le \sum_{k=1}^{M}\frac{6\varepsilon_k}{N^{\frac{1}{p}}\sqrt{N}}\sqrt{\log N(\mathcal{A}, \varepsilon_k, \|\cdot\|_p)} \tag{186}$$

$$= \frac{6}{N^{\frac{1}{p}}\sqrt{N}}\sum_{k=1}^{M}\varepsilon_k\sqrt{\log N(\mathcal{A}, \varepsilon_k, \|\cdot\|_p)} \tag{187}$$

$$= \frac{12}{N^{\frac{1}{p}}\sqrt{N}}\sum_{k=1}^{M}(\varepsilon_k - \varepsilon_{k+1})\sqrt{\log N(\mathcal{A}, \varepsilon_k, \|\cdot\|_p)} \tag{188}$$

$$= \frac{12}{N^{\frac{1}{p}}\sqrt{N}}\sum_{k=1}^{M}\int_{\varepsilon_{k+1}}^{\varepsilon_k}\mathrm{d}\varepsilon'\sqrt{\log N(\mathcal{A}, \varepsilon_k, \|\cdot\|_p)} \tag{189}$$

$$\le \frac{12}{N^{\frac{1}{p}}\sqrt{N}}\sum_{k=1}^{M}\int_{\varepsilon_{k+1}}^{\varepsilon_k}\mathrm{d}\varepsilon'\sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_p)} \tag{190}$$

$$= \frac{12}{N^{\frac{1}{p}}\sqrt{N}}\int_{\varepsilon_{M+1}}^{\varepsilon_1}\mathrm{d}\varepsilon'\sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_p)} \tag{191}$$

Finally, we have

$$\mathfrak{R}(\mathcal{A}) \le \frac{1}{N^{\frac{1}{p}}}\left(\varepsilon_M + \frac{12}{\sqrt{N}}\int_{\varepsilon_{M+1}}^{\frac{r}{2}}\mathrm{d}\varepsilon'\sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_p)}\right) \tag{192}$$

For any $\varepsilon \in (0, r/2]$, we can find an integer M such that $\varepsilon \le \varepsilon_{M+1} \le 2\varepsilon$ and thus $\varepsilon_M \le 4\varepsilon$. Then, we have

$$\mathfrak{R}(\mathcal{A}) \le \frac{1}{N^{\frac{1}{p}}}\left(4\varepsilon + \frac{12}{\sqrt{N}}\int_{\varepsilon}^{\frac{r}{2}}\mathrm{d}\varepsilon'\sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_p)}\right), \tag{193}$$

and the upper bound holds for all $\varepsilon \in (0, r/2]$, we can take the infimum over all $\varepsilon > 0$ to obtain the desired result. □

Remark 13. Since the above inequality holds for any $\varepsilon \in (0, r/2]$, if we take $\varepsilon = 0$, we have $\mathfrak{R}(\mathcal{A}) \le 2r/N^{\frac{1}{p}}$.

Theorem 6 (Chaining bound)**.** Let $\mathcal{A} \subset \mathbb{R}^N$ be a set such that $\max_{\boldsymbol{a}\in\mathcal{A}}\|\boldsymbol{a}\|_{p,N} \le r$. Then, the Rademacher complexity $\mathfrak{R}(\mathcal{A})$ is upper bounded by the covering number of $\mathcal{A}$ as

$$\mathfrak{R}(\mathcal{A}) \le \inf_{\varepsilon\in(0,\frac{r}{2}]}\left\{4\varepsilon + \frac{12}{\sqrt{N}}\int_{\varepsilon}^{\frac{r}{2}}\mathrm{d}\varepsilon'\sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_{p,N})}\right\} \tag{194}$$

Proof. For any $k \in \{0, 1, 2, \ldots, M\}$, let $\mathcal{C}_k \subset \mathbb{R}^N$ be an ε_k-covering net of $\mathcal{A}$ w.r.t. $\|\cdot\|_{p,N}$, where $\varepsilon_k := r/2^k$ and $|\mathcal{C}_k| = N(\mathcal{A}, \varepsilon_k, \|\cdot\|_{p,N})$. In particular, we can take $\mathcal{C}_0 := \{\boldsymbol{0}\}$, because $\|\boldsymbol{a} - \boldsymbol{0}\|_{p,N} \le r =: \varepsilon_0$ for all $\boldsymbol{a} \in \mathcal{A}$. For any k and any $\boldsymbol{a} \in \mathcal{A}$, let $\pi_k(\boldsymbol{a}) = \boldsymbol{c}$ such that $\boldsymbol{c} \in \mathcal{C}_k$ and $\|\boldsymbol{a} - \boldsymbol{c}\|_{p,N} \le \varepsilon_k$. Furthermore, we define $\Delta_k(\boldsymbol{a}) = \pi_k(\boldsymbol{a}) - \pi_{k-1}(\boldsymbol{a})$. Then, for any $\boldsymbol{a} \in \mathcal{A}$, we have

$$\boldsymbol{a} = \boldsymbol{a} + \pi_0(\boldsymbol{a}) - \pi_M(\boldsymbol{a}) + \sum_{k=1}^{M}\Delta_k(\boldsymbol{a}) = \boldsymbol{a} - \pi_M(\boldsymbol{a}) + \sum_{k=1}^{M}\Delta_k(\boldsymbol{a}).$$

Hence,

$$\mathfrak{R}(\mathcal{A}) = \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^N}\left[\sup_{\boldsymbol{a}\in\mathcal{A}}\frac{\langle \boldsymbol{a},\boldsymbol{\sigma}\rangle}{N}\right] \tag{195}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{\boldsymbol{a}\in\mathcal{A}}\left\{\frac{\langle a-\pi_M(\boldsymbol{a}),\boldsymbol{\sigma}\rangle}{N}+\frac{\langle\sum_{k=1}^M\Delta_k(\boldsymbol{a}),\boldsymbol{\sigma}\rangle}{N}\right\}\right] \tag{196}$$

$$\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{\boldsymbol{a}\in\mathcal{A}}\frac{\langle a-\pi_M(\boldsymbol{a}),\boldsymbol{\sigma}\rangle}{N}\right]+\sum_{k=1}^M\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{\boldsymbol{a}\in\mathcal{A}}\frac{\langle\Delta_k(\boldsymbol{a}),\boldsymbol{\sigma}\rangle}{N}\right] \quad \because \sup\Sigma\leq\Sigma\sup \tag{197}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{\boldsymbol{a}\in\mathcal{A}}\frac{\langle a-\pi_M(\boldsymbol{a}),\boldsymbol{\sigma}\rangle}{N}\right]+\sum_{k=1}^M\mathfrak{R}(\Delta_k) \tag{198}$$

, where $\Delta_k=\{\Delta_k(\boldsymbol{a}):\boldsymbol{a}\in\mathcal{A}\}$.

For the first term, we use Hölder's inequality 3 and the fact: $\|\boldsymbol{a}-\pi_M(\boldsymbol{a})\|_{p,N}\leq\varepsilon_M \iff \|\boldsymbol{a}-\pi_M(\boldsymbol{a})\|_p\leq N^{\frac{1}{p}}\varepsilon_M$ to obtain

$$\langle\boldsymbol{a}-\pi_M(\boldsymbol{a}),\boldsymbol{\sigma}\rangle\leq\|\boldsymbol{a}-\pi_M(\boldsymbol{a})\|_p\|\boldsymbol{\sigma}\|_q\leq N^{\frac{1}{p}}\varepsilon_M\cdot N^{\frac{1}{q}}=N\,\varepsilon_M.$$

Thus, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\in\{\pm1\}^N}\left[\sup_{\boldsymbol{a}\in\mathcal{A}}\frac{\langle\boldsymbol{a}-\pi_M(\boldsymbol{a}),\boldsymbol{\sigma}\rangle}{N}\right]\leq\varepsilon_M. \tag{199}$$

For the second term, we observe that

$$\begin{aligned}|\Delta_k| &= |\{\pi_k(\boldsymbol{a})-\pi_{k-1}(\boldsymbol{a}):\boldsymbol{a}\in\mathcal{A}\}|\\ &\leq|\{\pi_k(\boldsymbol{a}):\boldsymbol{a}\in\mathcal{A}\}|\cdot|\{\pi_{k-1}(\boldsymbol{a}):\boldsymbol{a}\in\mathcal{A}\}|\\ &= N(\mathcal{A},\varepsilon_k,\|\cdot\|_{p,N})\cdot N(\mathcal{A},\varepsilon_{k-1},\|\cdot\|_{p,N})\\ &\leq N(\mathcal{A},\varepsilon_k,\|\cdot\|_{p,N})^2\end{aligned}$$

and for all $\boldsymbol{a}\in\mathcal{A}$,

$$\|\Delta_k(\boldsymbol{a})\|_{p,N}\leq\|\pi_k(\boldsymbol{a})-\boldsymbol{a}\|_{p,N}+\|\boldsymbol{a}-\pi_{k-1}(\boldsymbol{a})\|_{p,N}\leq3\varepsilon_k \iff \|\Delta_k(\boldsymbol{a})\|_p\leq3N^{\frac{1}{p}}\varepsilon_k \tag{200}$$

and remember that $\|\boldsymbol{a}\|_p\leq\|\boldsymbol{a}\|_2\leq N^{(1/2-1/p)}\|\boldsymbol{a}\|_p$ for all $1\leq r<p$, so we have

$$\|\Delta_k(\boldsymbol{a})\|_2\leq N^{(1/2-1/p)}\cdot3N^{\frac{1}{p}}\varepsilon_k=3\sqrt{N}\varepsilon_k$$

Thus, by Massart's lemma 23, we have

$$\mathfrak{R}(\Delta_k)\leq\frac{3\sqrt{N}\varepsilon_k\sqrt{2\log|\Delta_k|}}{N}\leq3\varepsilon_k\sqrt{\frac{2\log|\mathcal{C}_k|^2}{N}}=6\varepsilon_k\sqrt{\frac{\log N(\mathcal{A},\varepsilon_k,\|\cdot\|_{p,N})}{N}}$$

Therefore,

$$\sum_{k=1}^M\mathfrak{R}(\Delta_k)\leq\sum_{k=1}^M6\varepsilon_k\sqrt{\frac{\log N(\mathcal{A},\varepsilon_k,\|\cdot\|_{p,N})}{N}} \tag{201}$$

$$=\frac{6}{\sqrt{N}}\sum_{k=1}^M\varepsilon_k\sqrt{\log N(\mathcal{A},\varepsilon_k,\|\cdot\|_{p,N})} \tag{202}$$

$$=\frac{12}{\sqrt{N}}\sum_{k=1}^M(\varepsilon_k-\varepsilon_{k+1})\sqrt{\log N(\mathcal{A},\varepsilon_k,\|\cdot\|_{p,N})} \tag{203}$$

$$=\frac{12}{\sqrt{N}}\sum_{k=1}^M\int_{\varepsilon_{k+1}}^{\varepsilon_k}\mathrm{d}\varepsilon'\sqrt{\log N(\mathcal{A},\varepsilon_k,\|\cdot\|_{p,N})} \tag{204}$$

$$\leq \frac{12}{\sqrt{N}} \sum_{k=1}^{M} \int_{\varepsilon_{k+1}}^{\varepsilon_k} \mathrm{d}\varepsilon' \sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_{p,N})} \tag{205}$$

$$= \frac{12}{\sqrt{N}} \int_{\varepsilon_{M+1}}^{\varepsilon_1} \mathrm{d}\varepsilon' \sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_{p,N})} \tag{206}$$

Finally, we have

$$\mathfrak{R}(\mathcal{A}) \leq \varepsilon_M + \frac{12}{\sqrt{N}} \int_{\varepsilon_{M+1}}^{\frac{r}{2}} \mathrm{d}\varepsilon' \sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_{p,N})} \tag{207}$$

For any $\varepsilon \in (0, r/2]$, we can find an integer M such that $\varepsilon \leq \varepsilon_{M+1} \leq 2\varepsilon$ and thus $\varepsilon_M \leq 4\varepsilon$. Then, we further bound $\mathfrak{R}(\mathcal{A})$ as

$$\mathfrak{R}(\mathcal{A}) \leq 4\varepsilon + \frac{12}{\sqrt{N}} \int_{\varepsilon}^{\frac{r}{2}} \mathrm{d}\varepsilon' \sqrt{\log N(\mathcal{A}, \varepsilon', \|\cdot\|_{p,N})}, \tag{208}$$

and the upper bound holds for all $\varepsilon \in (0, r/2]$, we can take the infimum over all $\varepsilon > 0$ to obtain the desired result. □

Remark 14. Since the above inequality holds for any $\varepsilon \in (0, r/2]$, if we take $\varepsilon = 0$, we have $\mathfrak{R}(\mathcal{A}) \leq 2r$.

VIII. Application to QML model

In this section, we derive the upper bound of the generalization error depending on the parameter-initialization strategy using covering number and Rademacher complexity.

A. Notation

- $\mathcal{X}, \mathcal{Y}$: input and output spaces
- $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$
- $\mathcal{S} \coloneqq \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\} \subset \mathcal{Z}$: : a set of N training samples
- $\Theta \coloneqq \{\ (\theta_1, \theta_2, \ldots, \theta_m) \mid \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R}\ \} \subset \mathbb{R}^m$: set of parameters of VQML
- $\mathcal{E}_{\boldsymbol{\theta}} : \rho(\boldsymbol{x}) \mapsto U(\boldsymbol{\theta})\rho(\boldsymbol{x})U^\dagger(\boldsymbol{\theta})$: parameterized quantum circuit channel
- O: observable
- $\rho(\boldsymbol{x})$: quantum input state
- $\mathcal{F} \coloneqq \{\ f_{\boldsymbol{\theta}} : (\boldsymbol{x}_i, y_i) \mapsto \mathrm{Tr}[O\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i))] \mid \boldsymbol{\theta} \in \Theta\ \} \subset [0, C]^{\mathcal{Z}}$: hypothesis class of VQML
- $\mathcal{L}(y_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$: loss function evaluated on a sample $(\boldsymbol{x}_i, y_i)$ which is L-Lipschitz continuous in the second argument.
- $\hat{R}_N(f_{\boldsymbol{\theta}}) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$: empirical risk of a hypothesis $f_{\boldsymbol{\theta}} \in \mathcal{F}$ over a sample $\mathcal{S}$
- $R(f_{\boldsymbol{\theta}}) \coloneqq \mathbb{E}_{(x,y)\sim P}[\mathcal{L}(y, f_{\boldsymbol{\theta}}(x))]$: expected risk of a hypothesis $f_{\boldsymbol{\theta}} \in \mathcal{F}$
- $\mathcal{G} \coloneqq \{(x, y) \mapsto \mathcal{L}(y, f_{\boldsymbol{\theta}}(x)) \mid f_{\boldsymbol{\theta}} \in \mathcal{F}\} \subseteq [0, LC]^{\mathcal{Z}}$: loss function class
- $\mathcal{F} \bullet \mathcal{S} \coloneqq \{\ (f_{\boldsymbol{\theta}}(\boldsymbol{x}_1), \ldots, f_{\boldsymbol{\theta}}(\boldsymbol{x}_N)) \mid f_{\boldsymbol{\theta}} \in \mathcal{F}, (\boldsymbol{x}_i, y_i) \in \mathcal{S}\ \} \subset [0, C]^N$: prediction evaluated on a sample $\mathcal{S}$
- $\mathcal{G} \bullet \mathcal{S} \coloneqq \{\ (\mathcal{L}(y_1, f_{\boldsymbol{\theta}}(\boldsymbol{x}_1)), \ldots, \mathcal{L}(y_N, f_{\boldsymbol{\theta}}(\boldsymbol{x}_N))) \mid f_{\boldsymbol{\theta}} \in \mathcal{F}, (\boldsymbol{x}_i, y_i) \in \mathcal{S}\ \} \subset [0, LC]^N$: loss function evaluated on a sample $\mathcal{S}$

Remember that from Lemma 21 and 22 the generalization error for any hypothesis $f_{\boldsymbol{\theta}} \in \mathcal{F}$ is bounded by the Rademacher complexity with probability at least $1-\delta$ as

$$R(f_{\boldsymbol{\theta}}) - \hat{R}_N(f_{\boldsymbol{\theta}}) \leq 2\,\mathfrak{R}(\mathcal{G} \bullet \mathcal{S}) + 3LC\sqrt{\frac{2\log(2/\delta)}{N}} \tag{209}$$

$$\leq 2L\,\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) + 3LC\sqrt{\frac{2\log(2/\delta)}{N}} \tag{210}$$

We can apply the one step discretization theorem 4 and the chaining bound 6 to bound the Rademacher complexity $\mathfrak{R}(\mathcal{F} \bullet \mathcal{S})$. Before applying these theorems, we first bound the covering number of $\mathcal{F} \bullet \mathcal{S}$ by the covering number of Θ.

Lemma 25. The covering number of $\mathcal{F} \bullet \mathcal{S}$ can be bounded by the covering number of Θ as

$$N(\mathcal{F} \bullet \mathcal{S}, \varepsilon, \|\cdot\|_{p,N}) \leq N(\Theta, \varepsilon/C, \|\cdot\|_1) \tag{211}$$

Proof. First, you can derive the following inequality for any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta$, using the same argument as in the proof of Theorem 1:

$$|f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - f_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{x}_i)| = \left|\mathrm{Tr}[O\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i))] - \mathrm{Tr}\left[O\mathcal{E}_{\tilde{\boldsymbol{\theta}}}(\rho(\boldsymbol{x}_i))\right]\right| \tag{212}$$

$$= \left|\mathrm{Tr}\left[O(\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i)) - \mathcal{E}_{\tilde{\boldsymbol{\theta}}}(\rho(\boldsymbol{x}_i)))\right]\right| \tag{213}$$

$$\leq \mathrm{Tr}\left|O(\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i)) - \mathcal{E}_{\tilde{\boldsymbol{\theta}}}(\rho(\boldsymbol{x}_i)))\right| \tag{214}$$

$$=: \left\|O(\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i)) - \mathcal{E}_{\tilde{\boldsymbol{\theta}}}(\rho(\boldsymbol{x}_i)))\right\|_1 \tag{215}$$

$$\leq \|O\|_\infty \left\|\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i)) - \mathcal{E}_{\tilde{\boldsymbol{\theta}}}(\rho(\boldsymbol{x}_i))\right\|_1 \quad \because \text{Hölder's inequality} \tag{216}$$

$$\leq \|O\|_\infty \left\|\mathcal{E}_{\boldsymbol{\theta}} - \mathcal{E}_{\tilde{\boldsymbol{\theta}}}\right\|_\diamond \tag{217}$$

$$\leq C\,2\sum_{l=1}^{L}\sum_{m=1}^{n} \|R_{l,m}(\theta_{l,m}) - R_{l,m}(\tilde{\theta}_{l,m})\|_\infty \tag{218}$$

$$\leq C\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1 \tag{219}$$

(Even if the classical input data and parameters are encoded alternately (data re-uploading), the above inequality holds.)

From this inequality, we have

$$\|\boldsymbol{f}_{\boldsymbol{\theta}} - \boldsymbol{f}_{\tilde{\boldsymbol{\theta}}}\|_{p,N} = \left(\frac{1}{N}\sum_{i=1}^{N}\left|\mathrm{Tr}[O\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i))] - \mathrm{Tr}\left[O\mathcal{E}_{\tilde{\boldsymbol{\theta}}}(\rho(\boldsymbol{x}_i))\right]\right|^p\right)^{\frac{1}{p}} \leq \left(\frac{1}{N}\sum_{i=1}^{N} C^p\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1^p\right)^{\frac{1}{p}} \leq C\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1 \tag{220}$$

, where $\boldsymbol{f}_{\boldsymbol{\theta}} = (f_{\boldsymbol{\theta}}(\boldsymbol{x}_1), \ldots, f_{\boldsymbol{\theta}}(\boldsymbol{x}_N)) \in \mathcal{F} \bullet \mathcal{S}$.

Thus, from Lemma 13, we have

$$N(\mathcal{F} \bullet \mathcal{S}, \varepsilon, \|\cdot\|_{p,N}) \leq N(\Theta, \varepsilon/C, \|\cdot\|_1) \tag{221}$$

□

Example 4. If $\Theta = B_1(\boldsymbol{\theta}_0, \gamma)$, then from Lemma 15, $N(\Theta, \varepsilon, \|\cdot\|_1) \leq (1 + 2\gamma/\varepsilon)^m \leq (3\gamma/\varepsilon)^m$, so we have

$$N(\Theta, \varepsilon/C, \|\cdot\|_1) \leq \left(\frac{3\gamma C}{\varepsilon}\right)^m. \tag{222}$$

Example 5. If $\Theta = \{\ (\theta_1, \theta_2, \ldots, \theta_m) \mid \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R}\ \} \subset \mathbb{R}^m$, then

$$N(\Theta, \varepsilon/C, \|\cdot\|_1) \leq N(\Theta, \varepsilon/Cm, \|\cdot\|_\infty) \tag{223}$$

$$= \prod_{i=1}^{T}\left\lceil\frac{Cmc_i}{\varepsilon}\right\rceil \tag{224}$$

$$\leq \prod_{i=1}^{T} \left(1 + \frac{Cmc_i}{\varepsilon}\right) \qquad (\because \lceil x \rceil \leq x + 1) \tag{225}$$

$$\leq \prod_{i=1}^{T} \left(\frac{1 + Cmc_i}{\varepsilon}\right) \qquad (\because 1 \leq 1/\varepsilon) \tag{226}$$

B. One step discretization

We can bound $\mathrm{Tr}[O\,\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i))]$ as

$$\mathrm{Tr}[O\,\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i))] \leq \mathrm{Tr}\,|O\,\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i))| = \|O\,\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i))\|_1 \leq \|O\|_\infty \|\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i))\|_1 = C \cdot 1 \tag{227}$$

, where we used Hölder's inequality 3 in the second inequality and the fact that the diamond norm of a quantum channel is 1 in the last equality.

So, we have

$$\max_{\boldsymbol{f}_{\boldsymbol{\theta}} \in \mathcal{F} \bullet \mathcal{S}} \|\boldsymbol{f}_{\boldsymbol{\theta}}\|_2 \leq \left(\sum_{i=1}^{N} \mathrm{Tr}[O\,\mathcal{E}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}_i))]^2\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{N} C^2\right)^{\frac{1}{2}} = N^{\frac{1}{2}} C \tag{228}$$

and we can apply Theorem 4 to obtain

$$\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) \leq \inf_{\varepsilon > 0} \left\{ \varepsilon + C \sqrt{\frac{2 \log N(\mathcal{F} \bullet \mathcal{S}, \varepsilon/2, \|\cdot\|_{p,N})}{N}} \right\}. \tag{229}$$

Using the covering number bound in Eq. 211, we have

$$\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) \leq \inf_{\varepsilon > 0} \left\{ \varepsilon + C \sqrt{\frac{2 \log N(\Theta, \varepsilon/2C, \|\cdot\|_1)}{N}} \right\} \quad \because \text{Eq.(211)} \tag{230}$$

$$= \inf_{\varepsilon > 0} C \left\{ \varepsilon + \sqrt{\frac{2 \log N(\Theta, \varepsilon/2, \|\cdot\|_1)}{N}} \right\} \quad \because \varepsilon \to C\varepsilon \tag{231}$$

If $\Theta = B_1(\boldsymbol{\theta}_0, \gamma)$, then from Eq.(222) we have

$$\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) \leq C \left\{ \varepsilon + \sqrt{\frac{2m \log(6\gamma C/\varepsilon)}{N}} \right\} \tag{232}$$

$$\leq C \left\{ \frac{1}{N} + \sqrt{\frac{2m \log(6\gamma NC)}{N}} \right\} \quad (\because \text{choose } \varepsilon = 1/N) \tag{233}$$

If $\Theta = \{\ (\theta_1, \theta_2, \ldots, \theta_m) \mid \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R}\ \} \subset \mathbb{R}^m$, then from Eq.(226) we have

$$\mathfrak{R}(\mathcal{F} \bullet \mathcal{S}) \leq \inf_{\varepsilon > 0} C \left\{ \varepsilon + \sqrt{\frac{2 \sum_{i=1}^{T} \log\left(1 + \frac{2Cmc_i}{\varepsilon}\right)}{N}} \right\} \tag{234}$$

$$\leq C \left\{ \frac{1}{N} + \sqrt{\frac{2 \sum_{i=1}^{T} \log(1 + 2NCmc_i)}{N}} \right\} \quad (\because \text{choose } \varepsilon = 1/N) \tag{235}$$

In one step discretization, the dependence on the number of data points N is $O(\sqrt{\log N/N})$. This can be improved to $O(\sqrt{1/N})$ by chaining the covering numbers.

Finally, we get the generalization error bound by substituting the upper bound of Rademacher complexity into Eq.(210).

When $\Theta = B_1(\boldsymbol{\theta}_0, \gamma)$, with probability at least $1-\delta$,

$$R(f_{\boldsymbol{\theta}}) - \hat{R}_N(f_{\boldsymbol{\theta}}) \leq 2LC\left\{\frac{1}{N} + \sqrt{\frac{2m\log(6\gamma NC)}{N}}\right\} + 3LC\sqrt{\frac{2\log(2/\delta)}{N}} \tag{236}$$

When $\Theta = \{\ (\theta_1, \theta_2, \ldots, \theta_m) \mid \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R}\ \} \subset \mathbb{R}^m$, with probability at least $1-\delta$,

$$R(f_{\boldsymbol{\theta}}) - \hat{R}_N(f_{\boldsymbol{\theta}}) \leq 2LC\left\{\frac{1}{N} + \sqrt{\frac{2\sum_{i=1}^{T}\log(1+2NCmc_i)}{N}}\right\} + 3LC\sqrt{\frac{2\log(2/\delta)}{N}} \tag{237}$$

C. Chaining bound

Since $\max_{\boldsymbol{f}_{\boldsymbol{\theta}} \in \mathcal{F}\bullet\mathcal{S}} \|\boldsymbol{f}_{\boldsymbol{\theta}}\|_{p,N} \leq \left(\frac{1}{N}\sum_{i=1}^{N} C^p\right)^{\frac{1}{p}} = C$, from the chaining bound in Theorem 6, we have

$$\mathfrak{R}(\mathcal{F}\bullet\mathcal{S}) \leq \inf_{\varepsilon \in (0, \frac{C}{2}]}\left\{4\varepsilon + \frac{12}{\sqrt{N}}\int_{\varepsilon}^{\frac{C}{2}} \mathrm{d}\varepsilon' \sqrt{\log N(\mathcal{F}\bullet\mathcal{S}, \varepsilon', \|\cdot\|_{p,N})}\right\} \tag{238}$$

By using $N(\mathcal{F}\bullet\mathcal{S}, \varepsilon, \|\cdot\|_{p,N}) \leq N(\Theta, \varepsilon/C, \|\cdot\|_1)$

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}\bullet\mathcal{S}) &\leq \inf_{\varepsilon \in (0, \frac{C}{2}]}\left\{4\varepsilon + \frac{12}{\sqrt{N}}\int_{\varepsilon}^{\frac{C}{2}} \mathrm{d}\varepsilon' \sqrt{\log N(\Theta, \varepsilon'/C, \|\cdot\|_1)}\right\} \quad \because \text{Eq.(211)} && (239)\\
&= \inf_{\varepsilon \in (0, \frac{C}{2}]}\left\{4\varepsilon + C\frac{12}{\sqrt{N}}\int_{\varepsilon/C}^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{\log N(\Theta, \varepsilon', \|\cdot\|_1)}\right\} \quad \because \varepsilon' \to C\varepsilon' && (240)\\
&= C\inf_{\varepsilon \in (0, \frac{1}{2}]}\left\{4\varepsilon + \frac{12}{\sqrt{N}}\int_{\varepsilon}^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{\log N(\Theta, \varepsilon', \|\cdot\|_1)}\right\} \quad \because \varepsilon \to C\varepsilon && (241)
\end{aligned}$$

If $\Theta = B_1(\boldsymbol{\theta}_0, \gamma)$, then from Eq.(222) we have

$$\mathfrak{R}(\mathcal{F}\bullet\mathcal{S}) \leq C\inf_{\varepsilon \in (0, \frac{1}{2}]}\left\{4\varepsilon + \frac{12}{\sqrt{N}}\int_{\varepsilon}^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{m\log\left(\frac{3\gamma C}{\varepsilon'}\right)}\right\} \tag{242}$$

If $\Theta = \{\ (\theta_1, \theta_2, \ldots, \theta_m) \mid \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R}\ \} \subset \mathbb{R}^m$, then from Eq.(226) we have

$$\mathfrak{R}(\mathcal{F}\bullet\mathcal{S}) \leq C\inf_{\varepsilon \in (0, \frac{1}{2}]}\left\{4\varepsilon + \frac{12}{\sqrt{N}}\int_{\varepsilon}^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{\log\prod_{i=1}^{T}\left(\frac{1+Cmc_i}{\varepsilon'}\right)}\right\} \tag{243}$$

By using Eq.(210), we have
When $\Theta = B_1(\boldsymbol{\theta}_0, \gamma)$, with probability at least $1-\delta$,

$$R(f_{\boldsymbol{\theta}}) - \hat{R}_N(f_{\boldsymbol{\theta}}) \leq 2LC\inf_{\varepsilon \in (0, \frac{1}{2}]}\left\{4\varepsilon + \frac{12}{\sqrt{N}}\int_{\varepsilon}^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{m\log\left(\frac{3\gamma C}{\varepsilon'}\right)}\right\} + 3LC\sqrt{\frac{2\log(2/\delta)}{N}} \tag{244}$$

When $\Theta = \{\ (\theta_1, \theta_2, \ldots, \theta_m) \mid \theta_i \in [a_i, a_i + c_i],\ a_i, c_i \in \mathbb{R}\ \} \subset \mathbb{R}^m$, with probability at least $1-\delta$,

$$R(f_{\boldsymbol{\theta}}) - \hat{R}_N(f_{\boldsymbol{\theta}}) \leq 2LC\inf_{\varepsilon \in (0, \frac{1}{2}]}\left\{4\varepsilon + \frac{12}{\sqrt{N}}\int_{\varepsilon}^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{\log\prod_{i=1}^{T}\left(\frac{1+Cmc_i}{\varepsilon'}\right)}\right\} + 3LC\sqrt{\frac{2\log(2/\delta)}{N}} \tag{245}$$

D. Evaluation of the integral

We consider the integral $\int \sqrt{a-\log x}\,\mathrm{d}x$. Let $u=\sqrt{a-\log x}$, then $x=e^{a-u^2}$ and $\mathrm{d}x=-2ue^{a-u^2}\,\mathrm{d}u$. Then, we have

$$\int \sqrt{a-\log x}\,\mathrm{d}x = -2\int u^2 e^{a-u^2}\,\mathrm{d}u$$

Since $\frac{\mathrm{d}}{\mathrm{d}u}\left(-ue^{a-u^2}\right) = -e^{a-u^2}+2u^2e^{a-u^2}$,

$$-2\int u^2 e^{a-u^2}\,\mathrm{d}u = ue^{a-u^2} - e^a\int e^{-u^2}\,\mathrm{d}u \tag{246}$$

$$= ue^{a-u^2} - e^a\frac{\sqrt{\pi}}{2}\operatorname{erf}(u) + \text{const.} \tag{247}$$

$$= x\sqrt{a-\log x} - e^a\frac{\sqrt{\pi}}{2}\operatorname{erf}(\sqrt{a-\log x}) + \text{const.} \tag{248}$$

$$= x\sqrt{a-\log x} + e^a\frac{\sqrt{\pi}}{2}\operatorname{erfc}(\sqrt{a-\log x}) + \text{const.} \quad (\operatorname{erfc}(x) \coloneqq 1-\operatorname{erf}(x)) \tag{249}$$

We define $w(\varepsilon) \coloneqq 4\varepsilon + \frac{12}{\sqrt{N}}\int_\varepsilon^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{\log\prod_{i=1}^T \left(\frac{1+Cmc_i}{\varepsilon'}\right)}$ and $a \coloneqq \log\prod_{i=1}^T (1+Cmc_i)^{\frac{1}{T}}$. Convensionally, ε is taken as 0, so we have

$$w(0) = 12\sqrt{\frac{T}{N}}\int_0^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{a-\log\varepsilon'} \tag{250}$$

$$= 12\sqrt{\frac{T}{N}}\left[\frac{1}{2}\sqrt{a+\log 2} + \frac{\sqrt{\pi}}{2}e^a \operatorname{erfc}(\sqrt{a+\log 2})\right] \tag{251}$$

$$\le 12\sqrt{\frac{T}{N}}\left[\frac{1}{2}\sqrt{a+\log 2} + \frac{\sqrt{\pi}}{2}e^a\, e^{-a-\log 2}\right] \qquad (\operatorname{erfc}(x) \le e^{-x^2}) \tag{252}$$

$$= 6\sqrt{\frac{1}{N}\sum_{i=1}^T \log(1+Cmc_i) + \frac{T}{N}\log 2} + 3\sqrt{\pi}\sqrt{\frac{T}{N}} \tag{253}$$

Thus, using Eq.(210), we have

$$R(f_{\boldsymbol{\theta}}) - \hat{R}_N(f_{\boldsymbol{\theta}}) \le 12LC\sqrt{\frac{1}{N}\sum_{i=1}^T \log(1+Cmc_i) + \frac{T}{N}\log 2} + 6LC\sqrt{\pi}\sqrt{\frac{T}{N}} + 3LC\sqrt{\frac{2\log(2/\delta)}{N}} \tag{254}$$

To get a more refined bound, we evaluate the minimum of $w(\varepsilon)$. The derivative of $w(\varepsilon)$ is

$$w'(\varepsilon) = 4 - \frac{12}{\sqrt{N}}\sqrt{\log\prod_{i=1}^T \left(\frac{1+Cmc_i}{\varepsilon}\right)} \tag{255}$$

The minimum of $w(\varepsilon)$ is obtained when $w'(\varepsilon)=0$, i.e.,

$$\sqrt{\log\prod_{i=1}^T \left(\frac{1+Cmc_i}{\varepsilon}\right)} = \frac{\sqrt{N}}{3} \iff \varepsilon = e^{a-\frac{N}{9T}} \quad (\le 1/2) \tag{256}$$

Thus, the minimum of $w(\varepsilon)$ is

$$w\left(e^{a-\frac{N}{9T}}\right) = 4e^{a-\frac{N}{9T}} + \frac{12}{\sqrt{N}}\int_{e^{a-\frac{N}{9T}}}^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{\log\prod_{i=1}^T \left(\frac{1+Cmc_i}{\varepsilon'}\right)} \tag{257}$$

The integral can be evaluated as

$$\int_{e^{a-\frac{N}{9T}}}^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{\log \prod_{i=1}^{T} \left(\frac{1+Cmc_i}{\varepsilon'}\right)} = \sqrt{T} \int_{e^{a-\frac{N}{9T}}}^{\frac{1}{2}} \mathrm{d}\varepsilon' \sqrt{a - \log \varepsilon'} \qquad (a = \log \prod_{i=1}^{T} (1+Cmc_i)^{\frac{1}{T}}) \tag{258}$$

$$= \sqrt{T} \left[\varepsilon' \sqrt{a - \log \varepsilon'} + \frac{\sqrt{\pi}}{2} e^a \operatorname{erfc}(\sqrt{a - \log \varepsilon'}) \right]_{e^{a-\frac{N}{9T}}}^{\frac{1}{2}} \tag{259}$$

Thus,

$$R(f_{\boldsymbol{\theta}}) - \hat{R}_N(f_{\boldsymbol{\theta}}) \leq 2LC \left\{ 4\varepsilon + 12\sqrt{\frac{T}{N}} \left[\varepsilon' \sqrt{a - \log \varepsilon'} + \frac{\sqrt{\pi}}{2} e^a \operatorname{erfc}(\sqrt{a - \log \varepsilon'}) \right]_{\varepsilon}^{\frac{1}{2}} \right\} + 3LC\sqrt{\frac{2\log(2/\delta)}{N}} \tag{260}$$

, where $\varepsilon = \min\left\{e^{a-\frac{N}{9T}}, \frac{1}{2}\right\}$.

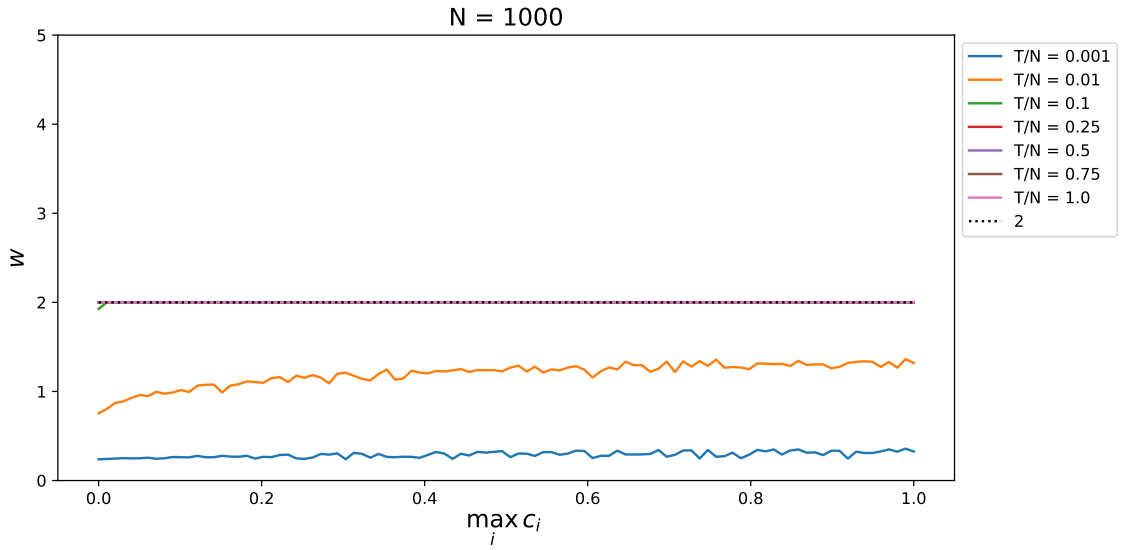

FIG. 11. Numerical evaluation of $\inf w(\varepsilon)$ for $N = 1000$, $m = 2T$, $T = 1, 10, 100, 250, 500, 750, 1000$

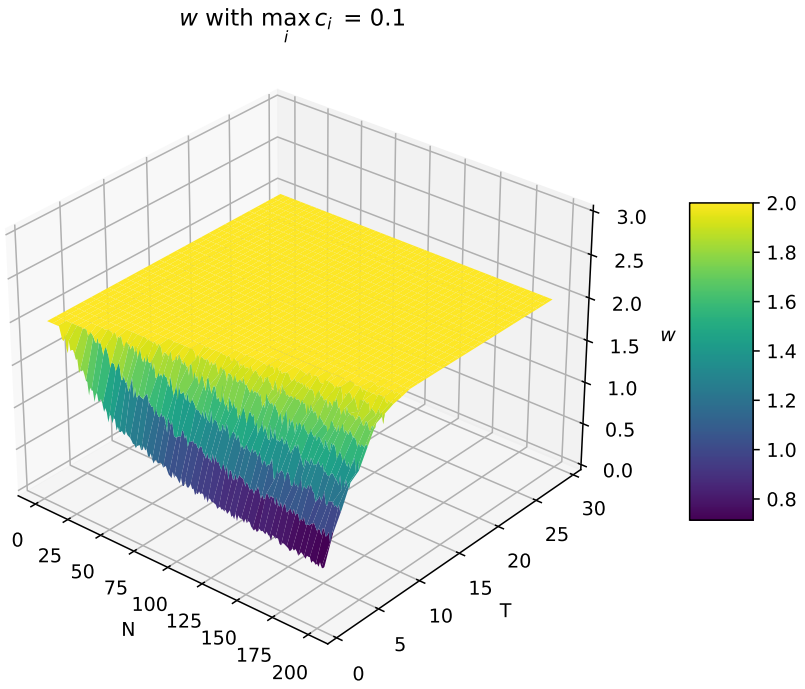

FIG. 12. Numerical evaluation of $\inf w(\varepsilon)$ for N and T with $m = 2T, c_i = 0.1$

[1] M. Imaizumi, Analysis of deep learning principles: From the perspective of generalization error (in japanese), Journal of the Japan Statistical Society **50**, 257 (2021).
[2] M. Imaizumi, Deep Learning Theory (2021), accessed: 2023-10-27.
[3] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, Generalization in quantum machine learning from few training data, Nature communications **13**, 4919 (2022).
[4] Y. Wang and B. Qi, Enhanced generalization of variational quantum learning under reduced-domain initialization, in *2023 42nd Chinese Control Conference (CCC)* (2023) pp. 6771–6776.
[5] M. M. Wolf, Mathematical foundations of supervised learning, (2023).
[6] J. Shafer, Unit 7 covering numbers and chaining, course: CS 294-220, Spring 2021.
[7] T. Barthel and J. Lu, Fundamental limitations for measurements in quantum many-body systems, Physical Review Letters **121**, 080406 (2018).
[8] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms* (Cambridge university press, 2014).
[9] C. Scott, Rademacher complexity, course: EECS 598: Statistical Learning Theory, Winter 2014.
[10] G. Brown and R. Ali, Bias/variance is not the same as approximation/estimation, Transactions on Machine Learning Research (2024).
[11] M. Kliesch and I. Roth, Theory of quantum system certification, PRX quantum **2**, 010201 (2021).
[12] I. Nechita, Z. Puchała, Ł. Pawela, and K. Życzkowski, Almost all quantum channels are equidistant, Journal of Mathematical Physics **59** (2018).
[13] T. Jo, Machine learning foundations, Supervised, Unsupervised, and Advanced Learning. Cham: Springer International Publishing **6**, 8 (2021).
[14] J. A. Tropp *et al.*, An introduction to matrix concentration inequalities, Foundations and Trends® in Machine Learning **8**, 1 (2015).
[15] M. Schuld, Supervised quantum machine learning models are kernel methods, arXiv preprint arXiv:2101.11020 (2021).
[16] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning, Nature communications **12**, 2631 (2021).
[17] L. Banchi, J. L. Pereira, S. T. Jose, and O. Simeone, Statistical complexity of quantum learning, Advanced Quantum Technologies **8**, 2300311 (2025).
[18] Y. Du, Y. Yang, D. Tao, and M.-H. Hsieh, Problem-dependent power of quantum neural networks on multiclass classification, Physical Review Letters **131**, 140601 (2023).