Leibniz Universität Hannover
Institut für Theoretische Physik

Bachelor Thesis

PAC Bounds for Quantum Neural Networks

Robin Syring
student number: 10031533

supervised by
Prof. Dr. Tobias J. Osborne

date: 31.01.2023

Contents

1 Introduction

People have thought of machines being able to think for decades. Today *artificial intelligence* (AI) is a field with many applications in a variety of different fields. One might think about computers being able to recognize handwritten digits or intelligent translators coming up with recognizing and translating several languages [4]. However, this field is also highly relevant for industrial purposes in the way that many companies are using *machine learning* to analyse data.

One might wonder how to program such intelligent systems. Take an example the task to recognize handwritten digits. Humans have different styles of handwriting and finding rules to characterize digits is not that easy. But look at how kids learn at school how to identify digits. They see examples and other people like teachers, parents or friends tell them what the names of the digits are and finally they learn from this data. So the idea behind machine learning is to set up some procedure that will "learn" by an input of examples - like handwritten digits with the name of each digit - and eventually outputs an algorithm being able to predict the name for future examples [13].

There are different kinds of machine learning but in this work we will focus on *inductive supervised learning*. In supervised learning we are having a labelled data set (*training set*) and are aiming to find a classifier function which predicts the label for unseen data. Inductive means that we can use the classifier function - in theory - infinitely many times and the test instances and their outcomes do not influence each other. This is commonly used for tasks like recognizing handwritten digits [7].

Most machine learning algorithms are based on minimizing some sort of *cost function* or sometimes called *empirical risk* evaluating how well each program candidate behaves for the given training set. Empirical indicates that it is of course an empirical quantity only evaluated on the given examples. However, these examples are just samples of an unknown probability distribution. But it could be that the training data does not represent a fair sample from the probability distribution. For example imagine just feeding our program for recognizing handwritten digits 0's and 1's, meaning that the training set just includes these digits. It could be that our program would work for these two examples, but obviously it would not work for other digits. To have control over such cases, there is a theoretical quantity - the *risk* - using the whole probability distribution in order to rate the performance of the candidate program. Since the probability distribution is unknown, the risk in practice cannot be determined exactly. But it is clear, that for a growing number of training examples,

the probability that the risk and the empirical risk are approximately the same grows. This circumstance is formulated with the help of *PAC bounds*. These provide a lower bound on the number of training data we need, such that using the empirical risk instead of the risk in the minimization process is *probably approximately correct* (PAC) [13].

In the current days another discipline is influencing the field of machine learning, namely *quantum mechanics*. *Moore's law* seemed to come to an end in the first two decades of the twenty-first century, because conventional methods for the fabrication of computer technology began to run up against fundamental limitations of size. When electronic devices become smaller, quantum effects begin to interfere in the functioning. So people wonder if there is a way to hold More's law true. One possibility might be to use quantum mechanics to perform calculations, instead of classical physics. This is known as *quantum computing* or *quantum information processing*. It turns out that it might offer a crucial speed up, especially for problems conventional algorithms have problems with [8].

In this work we focus on transferring the concept of PAC learning from the theory of classical inductive supervised learning to quantum learning theory. Therefore, we will look at several quantum scenarios and we are aiming to find PAC bounds.

We will start with a recap on the relevant aspects of classical inductive supervised learning in section (2). In (2.1) we will stack out the mathematical framework i.e. define the terms mentioned above like learning algorithm, training data, risk and empirical risk. Afterwards we will look at the exact definition of PAC bounds in (2.2). In order to find explicit proofs we need complexity measures introduced in (2.3). Here the Rademacher complexity is introduced before we will come to the concept of packing and covering in (2.3.1). Finally (2.3.2) presents helpful tools in order to gain PAC bounds, which will be used later.
Since we want to transfer concepts from classical learning to quantum machine learning, section (3) gives a brief recap of the basics of quantum mechanics. It starts with a reminder of Hilbert spaces and operators in (3.1). After this we talk about the concept of preparation and measurement (3.2) and the quantum bit in (3.3). We will look at how composite systems (3.4) and time evolution (3.5) are defined in quantum mechanics. In order to gain PAC bounds we need some distance measure, so we look on different norms in (3.6). Finally (3.7) will introduce the basic concepts of variational quantum algorithms.

In (4), the main results will be discussed, so this section is about PAC bounds for

quantum neural networks. In (4.1) we will look at complexity measures for some VQA structure before we will look at some real QNN's in (4.2). Thereby, (4.2.1) sets up the mathematical framework and in (4.2.2) and (4.2.3) we gain PAC bound for the structure. In (4.3), we will look at a different structure of QNN's, namely dissipative QNN's, and in (4.3.1) we gain a PAC bound for these networks.

2 Classical Learning Theory

In this section, we want to give a brief summary of the mathematical structure of classical inductive supervised machine learning. Unless mentioned otherwise we will focus mainly on the lecture notes "Mathematical Foundations of supervised Learning" by Michael Wolf [13].

2.1 Framework

The general framework for supervised learning scenarios is that we have an input set $\mathcal{X}$, an output set $\mathcal{Y}$ and some probability measure $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$. $\mathcal{P}$ determines the probability with which pairs of elements of $\mathcal{X}$ and $\mathcal{Y}$ come up. We aim for a map h giving the most likely element of $\mathcal{Y}$ upon input of an element of $\mathcal{X}$.

However, we have no direct access to the probability measure $\mathcal{P}$. But it induces pairs $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ which are treated as values of identically and independently distributed random variables, according to $\mathcal{P}$ and together are called *training data* $S = ((x_1, y_1), ..., (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})$. S contains the whole information about $\mathcal{P}$ that is accessible to us, so we want to learn from it labeling all data that is distributed according to $\mathcal{P}$. Thus S is the input of the learning algorithm.
The learning algorithm outputs a map $h : \mathcal{X} \to \mathcal{Y}$, called *hypothesis*, that aims to predict the right label $y \in \mathcal{Y}$ for an arbitrary $x \in \mathcal{X}$, in particular for those not contained in the training data. Therefore, a learning algorithm can formally be described by a map

$$\mathcal{A} : \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}$$

where $\mathcal{Y}^X$ denotes the set of maps from $\mathcal{X}$ to $\mathcal{Y}$. The range of $\mathcal{A}$ is called *hypothesis class* $\mathcal{F}$ and contains all possible hypotheses.

The goal of the learning algorithm is to find a good hypothesis for predicting the output for every input. Therefore we do need a *loss function* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ measuring how far $h(x)$ is from the particular y. Thus, the loss function may be seen as some kind of distance measure. We still need a way to estimate how good a hypothesis really is. Therefore, we take the average loss, called *risk*, which is given by

$$R(h) := \int_{\mathcal{X} \times \mathcal{Y}} L(y, h(x)) d\mathcal{P}(x, y).$$

Here the error for every pair (x, y) is weighted with its probability to occur. The smaller the risk, the better the hypothesis so by minimizing the risk we get the optimal hypothesis.

Since the exact probability measure $\mathcal{P}$ is unknown, we cannot calculate the risk and therefore cannot find the right hypothesis by minimizing $R(h)$. We therefore have to choose some different approach. One possibility is to use the information available to us, namely the training data. Instead of integrating the loss function over the whole set $\mathcal{X} \times \mathcal{Y}$ w.r.t. $\mathcal{P}$, we average the loss function of elements y and $h(x)$ over the training set

$$\hat{R}(h) := \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i)).$$

$\hat{R}(h)$ is called *empirical risk* and the process of minimizing is called *empirical risk minimization*.
Just by definition, it is intuitively clear, that the more training data available, the better the empirical risk approximates the true risk. To be more precise about how many training data we need so that using the empirical risk instead of the risk in the minimization process is approximately correct, we need *PAC-bounds*.

2.2 PAC Bounds

The last section shows that we are not able to use the true risk

$$R(h) = \int_{\mathcal{X} \times \mathcal{Y}} L(y, h(x)) d\mathcal{P}(x, y)$$

in the minimization process, since the probability measure $\mathcal{P}$ is unknown. Instead we are using the empirical risk

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i)).$$

Therefore, it is really important to know how much they differ from each other and to give an upper bound on their distance $|R(h) - \hat{R}(h)|$. Since the training data is considered to be random, we have to take into account the possibilities to have either a fair sample or not, so the bounds have to be probabilistic. Hence we need bounds of the form

$$\mathbb{P}_S[|R(h) - \hat{R}(h)| \geq \epsilon] \leq \delta(n, \epsilon) \quad \forall h \in \mathcal{F}, \tag{1}$$

where $\mathbb{P}_S$ denotes the probability of some event when S is considered a random variable with values distributed according to $\mathcal{P}^n$.

Equation (1) is true for all $h \in \mathcal{F}$, so that the bounds are only helpful to approximate the risk when we know the empirical risk. However, what we really want, is to justify the use of the empirical risk instead of the risk in the minimization process. So useful

bounds are of the form

$$\mathbb{P}_S[\exists h \in \mathcal{F} : |R(h) - \hat{R}(h)| \geq \epsilon] \leq \delta. \tag{2}$$

An equivalent formulation would be:

For $n \geq n_0$ we have with probability of at least $1 - \delta$ that $|R(h) - \hat{R}(h)| \leq \epsilon \ \forall h \in \mathcal{F}$.

Apparently, for small ϵ and δ, the probability is close to 1 that empirical and true risk are nearly the same. Since the probability measure $\mathcal{P}$ is unknown it is important, that ϵ and δ do not dependent on $\mathcal{P}$. Bounds of this form are called *PAC bounds*. They tell us if it is *probably approximately correct* to use the empirical risk instead of the risk in the minimization process.

We will find them with the help of different complexity measures of the hypothesis class $\mathcal{F}$ which we will introduce in the next section.

2.3 Complexity Measures

In this section, we discuss several complexity measures for function classes $\mathcal{F} \in \mathcal{Y}^{\mathcal{X}}$, where the sets $\mathcal{X}, \mathcal{Y}$ may be arbitrary. Our aim is to develop PAC bounds using them.

The easiest complexity measure is the cardinality $|\cdot|$, telling us how many elements are in one set. However if the function class is not countable, we do need another approach. In this work we discuss two possible concepts namely the *Rademacher complexity* and the *covering* or *packing number*.

Definition 2.1 (Rademacher Complexity)**.** *Consider a set of real-valued functions $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ and a vector $z \in \mathcal{Z}^n$. The empirical Raedemacher complexity of $\mathcal{G}$ w.r.t z is defined as*

$$\hat{\mathfrak{R}}(\mathcal{G}) := \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right].$$

Here, $\mathbb{E}_\sigma$ denotes the expectation value w.r.t. a uniform distribution of $\sigma \in \{-1, 1\}^n$.

Further, assume that the z_i's are the values of a vector of identically and independent distributed random variables $Z := (Z_1, ..., Z_n)$, each of them distributed according to some probability measure $\mathcal{P}$ on $\mathcal{Z}$. Then the Rademacher Complexities of $\mathcal{G}$ w.r.t. $\mathcal{P}$ are defined as

$$\mathfrak{R}_n := \mathbb{E}_Z \left[\hat{\mathfrak{R}}(\mathcal{G}) \right].$$

For the moment we put function classes away and apply the concept of the Rademacher complexity to a subset A of $\mathbb{R}^n$ in the following way:

$$\mathfrak{R}_n(A) := \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{x\in A}\langle\sigma, x\rangle\right]$$

This definition leads to the following Lemma.

Lemma 2.2 (Massart's Lemma)**.** *Consider A as a finite subset of $\mathbb{R}^n$ contained in an euclidean Ball of radius r. Then*

$$\mathfrak{R}_n(A) \leq \frac{r}{n}\sqrt{2\ln|A|}. \tag{3}$$

Proof. At first, we notice that Eq.(3) is unaffected by a translation, so we can assume that the center of the ball is at the origin. Therefore, we upper bound the rescaled set λA, where $\lambda > 0$ is a parameter to be choosen later.

$$\begin{aligned}
\mathfrak{R}_n(\lambda A)\cdot n = \mathbb{E}_\sigma\left[\max_{a\in\lambda A}\sum_{i=1}^n \sigma_i a_i\right] &\leq \mathbb{E}_\sigma\left[\ln\sum_{a\in\lambda A} e^{\sigma a}\right] \\
&\leq \ln\mathbb{E}_\sigma\left[\sum_{a\in\lambda A} e^{\sigma a}\right] = \ln\sum_{a\in\lambda A}\prod_{i=1}^n \frac{e^{a_i}+e^{-a_i}}{2} \\
&\leq \ln\sum_{a\in\lambda A}\exp(\|a\|_2^2/2) \\
&\leq \ln\left(e^{\frac{r^2\lambda^2}{2}}\cdot|A|\right) \quad = \quad \frac{1}{2}r^2\lambda^2 + \ln|A|
\end{aligned}$$

The first step transfers to the statement that the maximum over positive numbers can be upper bounded by their sum, when taking the exponential on both sides. The second inequality follows from concavity of the logarithm and Jensen's inequality (A.1). The third step exploits that the σ_i's are independently and uniformly distributed. The fourth step follows from $e^x + e^{-x} \leq 2e^{\frac{x^2}{2}}$, which holds for all $x \in \mathbb{R}$. Finally the last inequality bounds the sum by its maximal element multiplied by the number of terms. The claimed result can be obtained by inserting $\lambda = \sqrt{2\ln|A|}/r$ into

$$\mathfrak{R}_n(A) = \frac{1}{n}\mathbb{E}_\sigma\left[\max_{a\in A}\sum_{i=1}^n \sigma_i a_i\right] \leq \frac{1/2r^2\lambda^2 + \ln|A|}{\lambda} = \frac{r}{n}\sqrt{2\ln|A|}$$

□

2.3.1 Packing-Covering

In this part, we focus on covering numbers to measure the complexity of function classes.

Even though the Rademacher complexity was some sort of discretization, this idea

is expressed more explicitly in the following approach. Both the covering number and the packing number quantify the complexity of a function class in terms of the minimal number of discretization points to approximate any function in its class to a given degree.

Definition 2.3 (Covering)**.** *Let $(\mathcal{H}, d)$ be a pseudometric space, $A, B \subseteq \mathcal{H}$ and $\epsilon > 0$. If for all $b \in B$ there exists $a \in A$ such that $d(a, b) \leq \epsilon$, A is called an ϵ-cover of B. If further $A \subseteq B$, A is called an internal ϵ-cover of B.*

The ϵ-covering number of B, $\mathcal{N}(B, \epsilon, d)$, is the smallest cardinality of any ϵ-cover of B. Analogously, the internal ϵ-covering number, $\mathcal{N}_{in}(B, \epsilon, d)$, is the smallest cardinality of any internal ϵ-cover of B.

Thus the ϵ-covering number determines how many ϵ-balls are at least required to cover a set. Obviously, there are more restrictions for an internal ϵ-cover, so

$$\mathcal{N}(B, \epsilon, d) \leq \mathcal{N}_{in}(B, \epsilon, d).$$

Definition 2.4 (Packing)**.** *Let $(\mathcal{H}, d)$ be a pseudometric space, $A, B \subseteq \mathcal{H}$ and $\epsilon > 0$. If for all $a_1, a_2 \in A$, $d(a_1, a_2) > \epsilon$ holds, then $A \subseteq$ is called an ϵ-packing of B.*

The ϵ-packing number of B, $\mathcal{M}(B, \epsilon, d)$ is the largest cardinality of any ϵ-packing of B.

The ideas behind the definitions are visualized graphically in Fig.(1).

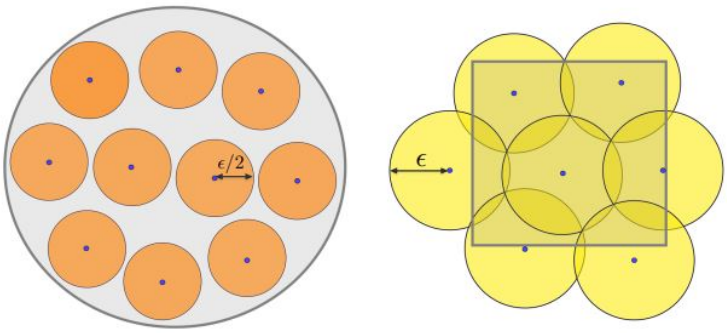

Figure 1: On the left: The set of blue points form an ϵ-packing of the gray disk, since the $\epsilon/2$-balls are non-intersecting. On the right: The set of blue points form an ϵ-covering of the gray square. Since not all balls lie in the square the cover is not internal.[13]

One might ask about the relation between covering and packing number. This is presented in the following proposition.

Proposition 2.5 (Relation between Covering and Packing). *For an arbitrary pseudometric space $(\mathcal{H}, d)$ and $B \subseteq \mathcal{H}$ the following relation holds.*

$$\mathcal{N}_{in}(B, \epsilon, d) \leq \mathcal{M}(B, \epsilon, d) \leq \mathcal{N}(B, \epsilon/2, d)$$

If there exists a bi-lipschitz function between two metric spaces, we can set their covering numbers in relation. The more accurate formulation is presented in Lem.(2.6).

Lemma 2.6 (Covering for two Metric Spaces). *Let $(\mathcal{H}_1, d_1)$ and $(\mathcal{H}_2, d_2)$ be metric spaces and $f : \mathcal{H}_1 \to \mathcal{H}_2$ be bi-Lipschitz such that*

$$d_2(f(x), f(y)) \leq K d_1(x, y) \quad \forall x, y \in \mathcal{H}_1 \tag{4}$$

$$d_2(f(x), f(y)) \geq k d_1(x, y) \quad \forall x, y \in \mathcal{H}_1 \quad with \quad d_1(x, y) \leq r. \tag{5}$$

Then the covering numbers obey

$$\mathcal{N}(\mathcal{H}_1, \frac{2\epsilon}{k}, d_1) \leq \mathcal{N}(\mathcal{H}_2, \epsilon, d_2) \leq \mathcal{N}(H_1, \frac{\epsilon}{K}, d_1). \tag{6}$$

Proof. Let Q be an $\frac{\epsilon}{K}$-covering for $(\mathcal{H}_1, d_1)$, i.e. $\forall z \in \mathcal{H}_1, \exists x \in Q s.t. d_1(x, z) \leq \frac{\epsilon}{K}$. Because f is bi-Lipschitz for all $z' \in \mathcal{H}_1$ there exists a $x' \in f(Q)$ s.t $d_2(x', z') \leq K d_1(x, z)$ which is due to the covering upper bounded by ϵ. Here x and z are chosen such that $f(x) = x'$ and $f(z) = z'$. This shows that $f(Q)$ is an ϵ-covering of $(\mathcal{H}_2, d_2)$ and

$$\mathcal{N}(\mathcal{H}_2, \epsilon, d_2) \leq \mathcal{N}(\mathcal{H}_1, \epsilon/K, d_1).$$

Let B be a $2\epsilon/k$-packing for $(\mathcal{H}_1, d_1)$, i.e. $d_1(x, y) > \frac{2\epsilon}{k} \quad \forall x \neq y \in B$. Because f is bi-Lipschitz we have for all $x' \neq y' \in f(B)$ it is $d_2(x', y') \geq k d_1(x, y) \geq 2\epsilon$. Again x and y are chosen such that $f(x) = x'$ and $f(y) = y'$. Further we assumed that $d_1(x, y) \leq r$ for the relevant x and y.
This shows that $f(B)$ is an 2ϵ-packing for $(\mathcal{H}_2, d_2)$. The packing numbers for the two metric spaces fulfill $\mathcal{M}(\mathcal{H}_2, 2\epsilon, d_2) \geq \mathcal{M}(\mathcal{H}_1, 2\epsilon/k, d_1)$. Application of Prop.(2.5) leads to

$$\mathcal{N}(\mathcal{H}_2, \epsilon, d_2) \geq \mathcal{N}(\mathcal{H}_1, 2\epsilon/k, d_1).$$

□

In this work, we need to measure the complexity of some function class. Doing that we will focus on the covering number, so we need some pseudometric on function spaces.
In some cases in learning theory it is sufficient to consider an arbitrary set $\mathcal{Z}$ and a function class $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ because in order to obtain PAC bounds one might consider the

function class of the loss function

$$\mathcal{G} = \{g \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}} | \exists h \in \mathcal{F} : g(x,y) = L(x,h(x)) \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}\}.$$

We define the $\|\cdot\|_{p,z}$-norm on $\mathcal{G}$'s linear span with respect to some $z \in \mathcal{Z}$ and $p \in (1,\infty)$ as

$$\|g\|_{p,z} = \left(\frac{1}{n}\sum_{i=1}^{n}|g(z_i)|^p\right)^{\frac{1}{p}} \quad with \quad \|g\|_{\infty,z} = \max_i |g(z_i)| \tag{7}$$

Its corresponding pseudometric is defined as $d(g_1,g_2) = \|g_1 - g_2\|_{p,z}$. However, there are other norms on function spaces presented in section (3.6).

All norms above depend on the choice of $z \in \mathcal{Z}$, so of course the covering number with respect to the specific norm also depends on z. One way to get rid of these dependency is to take the maximum over all $z \in \mathcal{Z}$ which leads to the uniform covering number.

Definition 2.7 (Uniform Covering Number)**.** *Consider the function class $\mathcal{G} \in \mathbb{R}^{\mathcal{Z}}$ with an arbitrary set $\mathcal{Z}$. Then the uniform ϵ-covering numbers are defined by*

$$\Gamma_p(n,\epsilon,\mathcal{G}) := \max\{\mathcal{N}_{in}(\epsilon,\mathcal{G},\|\cdot\|_{p,z}) | z \in \mathcal{Z}^n\}.$$

2.3.2 Inequalities to gain PAC Bounds

In this section we will look at relevant inequalities which we will use later in order to gain PAC bounds. These inequalities use the complexity measures introduced before.

In order to gain PAC bounds for some function class $\mathcal{F}$, Thm.(2.8) is an extremely helpful tool which uses the uniform covering number of the function class of the loss function.

Theorem 2.8 (PAC-Bound using Uniform Covering Numbers)**.** *Let $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ be an arbitrary function class and $L : \mathcal{Y}\times\mathcal{Y} \to [0,c]$ a loss function. We define $\mathcal{G} := \{g : \mathcal{X}\times\mathcal{Y} \to [0,c] \,|\, \exists h \in \mathcal{F} : g(x,y) = L(y,h(x))\}$. Then for any $\epsilon > 0$ and any probability measure P on $\mathcal{X}\times\mathcal{Y}$ the following inequality holds.*

$$\mathbb{P}_{S\sim P^n}\left[\exists h \in \mathcal{F} : \, |R(h) - \hat{R}(h)| \geq \epsilon\right] \leq 4\Gamma_1(2n,\epsilon/8,\mathcal{G})\exp\left(-\frac{n\epsilon^2}{32c^2}\right) \tag{8}$$

Proof. Consider an i.i.d. copy $S' \in (\mathcal{X}\times\mathcal{Y})^n$ of S. For every value of S' we denote by $\hat{R}(h)$ the corresponding empirical risk of a hypothesis $h \in \mathcal{F}$.
If we assume $|R(h) - \hat{R}(h)| > \epsilon$ and $|R(h) - \hat{R}'(h)| < \frac{\epsilon}{2}$, the triangle inequality implies

$|\hat{R}'(h) - \hat{R}(h)| > \frac{\epsilon}{2}$. Expressed in terms of the indicator function we get

$$\mathbb{1}_{|R(h)-\hat{R}(h)|>\epsilon}\mathbb{1}_{|R(h)-\hat{R}'(h)|<\frac{\epsilon}{2}} \leq \mathbb{1}_{|\hat{R}'(h)-\hat{R}(h)|>\frac{\epsilon}{2}}. \tag{9}$$

Now, we take the expectation value w.r.t. S'. Obviously this just effects the second and third term. At first we focus on the former one and just insert the definitions.

$$\begin{aligned}\mathbb{E}_{S'}\left[\mathbb{1}_{|R(h)-\hat{R}'(h)|<\frac{\epsilon}{2}}\right] &= \mathbb{P}_{S'}\left[|R(h) - \hat{R}'(h)| < \frac{\epsilon}{2}\right] \\ &= 1 - \mathbb{P}_{S'}\left[|R(h) - \hat{R}'(h)| \geq \frac{\epsilon}{2}\right]\end{aligned}$$

We define $Z_i = \frac{L(y_i,h(x_i))}{n}$ as random variables whose values are contained in an interval of length $\frac{c}{n}$. Just by definition one easily recognizes $\sum_{i=1}^{n} Z_i = \hat{R}(h)$ and $\sum_{i=1}^{n} \mathbb{E}_{S'}[Z_i] = R'(h)$. Now the last term can be upper bounded using Hoeffding's inequality (A.2)

$$\begin{aligned}\mathbb{E}_{S'}\left[\mathbb{1}_{|R(h)-\hat{R}'(h)|<\frac{\epsilon}{2}}\right] &\leq 1 - 2\exp\left(-\frac{2(\frac{\epsilon}{2})^2}{\sum_{i=1}^{n}(\frac{c}{n})^2}\right) \\ &= 1 - 2\exp\left(-\frac{\epsilon^2 n}{2c^2}\right) \leq \frac{1}{2},\end{aligned}$$

where the last step follows if we assume $n \geq 4c^2\epsilon^{-2}\ln 2$.

Now we take a look on the last term in Eq.(9).

$$\mathbb{E}_{S'}\left[\mathbb{1}_{|\hat{R}'(h)-\hat{R}(h)|>\frac{\epsilon}{2}}\right] = \mathbb{P}_{S'}\left[|\hat{R}'(h) - \hat{R}(h)| > \frac{\epsilon}{2}\right] \leq \mathbb{P}_{S'}\left[\exists h \in \mathcal{F} : |\hat{R}'(h) - \hat{R}(h)| > \frac{\epsilon}{2}\right]$$

Inserting both bounds back into Eq.(9), when the expectation value w.r.t. S' has already been taken, gives

$$\mathbb{1}_{|R(h)-\hat{R}(h)|>\epsilon} \leq 2\mathbb{P}_{S'}\left[\exists h \in \mathcal{F} : |\hat{R}'(h) - \hat{R}(h)| > \frac{\epsilon}{2}\right]$$

Surely this holds for all $h \in \mathcal{F}$, thus we can replace the left hand side by $\mathbb{1}_{\exists h\in\mathcal{F}:|R(h)-\hat{R}(h)|>\epsilon}$. Taking the expectation value w.r.t. S leads to

$$\mathbb{P}_S\left[\exists h \in \mathcal{F} : |R(h) - \hat{R}(h)| > \epsilon\right] \leq 2\mathbb{P}_{S,S'}\left[\exists h \in \mathcal{F} : |\hat{R}'(h) - \hat{R}(h)| > \frac{\epsilon}{2}\right].$$

By exploiting the definition we can further write

$$
\begin{aligned}
&2\mathbb{P}_{S,S'}\left[\exists h \in \mathcal{F} : |\hat{R}'(h) - \hat{R}(h)| > \frac{\epsilon}{2}\right] \\
&= 2\mathbb{E}_{S,S'}\left[\exists h \in \mathcal{F} : \frac{1}{n}|\sum_{i=1}^{n}\big(L(Y_i', h(X_i')) - L(Y_i, h(X_i))\big)| > \frac{\epsilon}{2}\right] \\
&= 2\mathbb{E}_{S,S'}\left[\mathbb{P}_\sigma\left[\exists h \in \mathcal{F} : \frac{1}{n}|\sum_{i=1}^{n}\sigma_i \cdot \big(L(Y_i, h(X_i)) - L(Y_i', h(X_i'))\big)| > \frac{\epsilon}{2}\right]\right] \\
&= 2\mathbb{E}_{S,S'}\left[\mathbb{P}_\sigma\left[\exists g \in \mathcal{G} : \frac{1}{n}|\sum_{i=1}^{n}\sigma_i \cdot (g(X_i, Y_i) - g(X_i', Y_i'))| > \frac{\epsilon}{2}\right]\right].
\end{aligned}
$$

The first step uses the definition of $\hat{R}(h)$ and $\hat{R}'(h)$ and the definition of the expectation value. where $\sigma_i \in \{\pm 1\}$ are i.i.d. random variables. The equation is true to the fact, that multiplication with $\sigma_i = -1$ just contributes to interchanging $(X_i, Y_i) \leftrightarrow (X_i', Y_i')$ which has no effect since we consider i.i.d. random variables. For the sake of clarity, we sum up the last steps.

$$
\begin{aligned}
&\mathbb{P}_S\big[\exists h \in \mathcal{F} : |R(h) - \hat{R}(h)| > \epsilon\big] \\
&\qquad \leq 2\mathbb{E}_{S,S'}\left[\mathbb{P}_\sigma\big[\exists g \in \mathcal{G} : \frac{1}{n}\big|\sum_{i=1}^{n}\sigma_i \cdot (g(X_i, Y_i) - g(X_i', Y_i'))\big| > \frac{\epsilon}{2}\big]\right]
\end{aligned} \tag{10}
$$

The aim of the next step is to approximate g by an element g' of an internal $\frac{\epsilon}{8}$-covering $\mathcal{G}'$ of $\mathcal{G}$ w.r.t. the norm $\|\cdot\|_{1,S\cup S'}$. If we regard g and σ' as vectors in $\mathbb{R}^{2n}$, with the latter ± 1-valued, then we can rewrite the upper condition inside the expectation value as

$$
\begin{aligned}
\frac{\epsilon}{2} &< \frac{1}{n}\big|\sum_{i=1}^{n}\sigma_i(g(X_i, Y_i) - g(X_i', Y_i'))\big| = \frac{1}{n}|\langle\sigma', g\rangle| \\
&= \frac{1}{n}\big|\langle\sigma', g'\rangle + \langle\sigma', g - g'\rangle\big| \leq \frac{1}{n}|\langle\sigma', g'\rangle| + \frac{1}{n}|\langle\sigma', g - g'\rangle| \\
&\leq \frac{1}{n}|\langle\sigma', g'\rangle| + \frac{1}{n}\|\sigma'\|_{1,S\cup S'} \cdot \|g - g'\|_{1,S\cup S'} \\
&\leq \frac{1}{n}|\langle\sigma', g'\rangle| + \frac{\epsilon}{4}.
\end{aligned}
$$

Using the upper inequality we can look back at Eq.(10) and upper bound the right hand side by

$$
\begin{aligned}
&\mathbb{P}_\sigma\left[\exists g \in \mathcal{G} : \frac{1}{n}\big|\sum_{i=1}^{n}\sigma_i \cdot (g(X_i, Y_i) - g(X_i', Y_i'))\big| > \frac{\epsilon}{2}\right] \\
&\qquad \leq \mathbb{P}_\sigma\left[\exists g' \in \mathcal{G}' : \frac{1}{n}\big|\sum_{i=1}^{n}\sigma_i(g'(X_i, Y_i) - g'(X_i', Y_i'))\big| > \frac{\epsilon}{4}\right].
\end{aligned}
$$

By interpreting $Z_i := \sigma_i(g'(X_i, Y_i) - g'(X_i', Y_i'))$ as random variables, with zero mean (by varying over σ_i) and range in $[-c, c]$, we are finally able to finish the proof.

$$\begin{aligned}
&\mathbb{P}_S\Big[\exists h \in \mathcal{F} : |R(h) - \hat{R}(h)| \geq \epsilon\Big] \\
&\qquad \leq 2\mathbb{E}_{S,S'}\mathbb{P}_\sigma\left[\exists g' \in \mathcal{G}' : \frac{1}{n}\Big|\sum_{i=1}^n \sigma_i(g'(X_i, Y_i) - g'(X_i', Y_i'))\Big| > \frac{\epsilon}{4}\right] \\
&\qquad \leq 2\mathbb{E}_{S,S'}\sum_{g' \in \mathcal{G}'}\mathbb{P}_\sigma\left[\frac{1}{n}\Big|\sum_{i=1}^n \sigma_i(g'(X_i, Y_i) - g'(X_i', Y_i'))\Big| > \frac{\epsilon}{4}\right] \\
&\qquad = 2\mathbb{E}_{S,S'}\sum_{g' \in \mathcal{G}'}\mathbb{P}_Z\left[\Big|\sum_{i=1}^n Z_i\Big| > \frac{n\epsilon}{4}\right] \\
&\qquad \leq 4\mathcal{N}_{in}(\frac{\epsilon}{8}, \mathcal{G}, \|\cdot\|_{1,S\cup S'})\exp\left(-\frac{2(\frac{n\epsilon}{4})^2}{\sum_{i=1}^n (2c)^2}\right) \\
&\qquad \leq 4\Gamma_1(2n, \frac{\epsilon}{8}, \mathcal{G})\exp\left(-\frac{n\epsilon^2}{32c^2}\right)
\end{aligned}$$

The first inequality follows from inserting the former inequality into Eq.(10), the second is an application of the union bound, the third term follows from inserting the Z_i's, the forth is an application of the union bound and follows from $\sum_{g' \in \mathcal{G}'} = |\mathcal{G}|$, the last follows by definition of the uniform covering number. □

When using the foregoing theorem it might be a bit tedious to introduce $\mathcal{G}$. In some cases one can use the covering number of the function class $\mathcal{F}$ directly and the detour via $\mathcal{G}$ is not necessary. The exact formulation is presented in the following Lemma.

Lemma 2.9 (Relation between Uniform Covering Numbers)**.** *For $\mathcal{Y}, \tilde{\mathcal{Y}} \subseteq \mathbb{R}$, function class $\mathcal{F} \subseteq \tilde{\mathcal{Y}}^{\mathcal{X}}$ and Loss function $L : \mathcal{Y} \times \tilde{\mathcal{Y}} \to [0, c]$ define $\mathcal{G} := \{g : \mathcal{X} \times \tilde{\mathcal{Y}} \to [0, c] | \exists h \in \mathcal{F} : g(x, y) = L(y, h(x))\}$. If there exists an $l \in \mathbb{R}$, such that for all $\tilde{y}_1, \tilde{y}_2 \in \tilde{\mathcal{Y}}$ and all $y \in \mathcal{Y}$, $|L(y, \tilde{y}_1) - L(y, \tilde{y}_2)| \leq l \cdot |\tilde{y}_1 - \tilde{y}_2|$ yields, then $\forall p \in [1, \infty]$, $\epsilon > 0$ and every $n \in \mathbb{N}$*

$$\Gamma_p(n, \epsilon, \mathcal{G}) \leq \Gamma_p(n, \frac{\epsilon}{l}, \mathcal{F}). \tag{11}$$

Proof. Let $\tilde{\mathcal{F}}$ be an $\frac{\epsilon}{l}$-cover of $\mathcal{F}$, i.e. for all $f \in \mathcal{F}$ there exists an $\tilde{f} \in \tilde{\mathcal{F}} : |f - \tilde{f}| \leq \frac{\epsilon}{l}$. The Lipschitz assumption implies that every $\frac{\epsilon}{l}$-cover of $\mathcal{F}$ becomes an ϵ-cover of $\mathcal{G}$ since we have:

$$\left(\frac{1}{n}\sum_{i=1}^n |L(y_i, f(x_i)) - L(y_i, \tilde{f}(x_i))|^p\right)^{\frac{1}{p}} \leq l\left(\frac{1}{n}\sum_{i=1}^n |f(x_i) - \tilde{f}(x_i)|^p\right)^{\frac{1}{p}} \leq \epsilon$$

□

We discussed two different complexity measures so far, namely the Rademacher complexity and the concept of covering and packing numbers. Dudley's Theorem(2.10)

presents a relation between them.

Theorem 2.10 (Dudley's Theorem)**.** *Let $z \in \mathcal{Z}^n$ be a fixed vector and let $\mathcal{G}$ be a subset of the pseudometric space $(\mathbb{R}^{\mathcal{Z}}, \|\cdot\|_{2,z})$. Set $\gamma_0 := sup_{g\in\mathcal{G}}\|g\|_{2,z}$. Then the emperical Rademacher complexity of $\mathcal{G}$ w.r.t. z fullfils*

$$\hat{\mathfrak{R}}(\mathcal{G}) \leq \inf_{\alpha\in[0,\gamma_0/2]}\left(4\alpha + \frac{12}{\sqrt{n}}\int_{\alpha}^{\gamma_0}\sqrt{\ln \mathcal{N}(\beta,\mathcal{G})}d\beta\right). \tag{12}$$

Proof. Define $\gamma_j := 2^{-j}\gamma_0$ for $j \in \mathbb{N}$. Let $G_j \subseteq \mathbb{R}^{\mathcal{Z}}$ be a minimal γ_j-cover of $\mathcal{G}$, which means $|G_j| = \mathcal{N}(\gamma_j,\mathcal{G})$ and for every $g \in \mathcal{G}$ there exists a $g_j \in G_j$, such that $\|g - g_j\|_{2,z} \leq \gamma_j$. This inequality still holds for $j = 0$, if one sets $g_0 := 0$.

For some $m \in \mathbb{N}$ which will be determined later, one inserts $g = g - g_m + \sum_{j=1}^{m}(g_j - g_{j-1})$, which is true due to $g_0 = 0$, into the definition of the empirical Rademacher complexity (2.1). One obtains

$$\begin{aligned}\hat{\mathfrak{R}}(\mathcal{G}) &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i\left(g(z_i) - g_m(z_i) + \sum_{j=1}^{m}g_j(z_i) - g_{j-1}(z_i)\right)\right]\\ &\leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i\left(g(z_i) - g_m(z_i)\right)\right] + \frac{1}{n}\sum_{j=1}^{m}\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i\left(g_j(z_i) - g_{j-1}(z_i)\right)\right].\end{aligned}$$

Further we bound the two terms separately. First, we consider the first term

$$\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i\left(g(z_i) - g_m(z_i)\right)\right] \leq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\|\sigma_i\|\cdot\|g(z_i) - g_m(z_i)\|\right] \leq \gamma_m.$$

We exploited the Cauchy-Schwarz-inequality and used the definition of γ_m.
The second term can be upper bounded by Massart's Lemma (2.2)

$$\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g_j(z_i) - g_{j-1}(z_i)\right)\right] \leq \sup_{g\in\mathcal{G}}\frac{\|g_j - g_{j-1}\|_{2,z}\sqrt{2\ln\left(|G_j|\cdot|G_{j-1}|\right)}}{\sqrt{n}},$$

where the Factor $\sqrt{n}$ comes from the normalization of the 2-norm. By the triangle inequality we are able to bound

$$\begin{aligned}\frac{1}{\sqrt{n}}\sup_{g\in\mathcal{G}}\|g_j - g_{j-1}\|_{2,z} &\leq \frac{1}{\sqrt{n}}\left(\sup_{g\in\mathcal{G}}\|g_j - g\|_{2,z} + \sup_{g\in\mathcal{G}}\|g - g_{j-1}\|_{2,z}\right)\\ &\leq \frac{\gamma_j + \gamma_{j-1}}{\sqrt{n}}\\ &= \frac{\gamma_j + 2\gamma_j}{\sqrt{n}} \quad = \quad \frac{3\gamma_j}{\sqrt{n}}.\end{aligned}$$

So we can finally merge the results.

$$\begin{aligned}\hat{\mathfrak{R}}(\mathcal{G}) &\leq \gamma_m + \frac{3}{\sqrt{n}} \sum_{j=1}^{m} \gamma_j \sqrt{2 \ln \left(|G_j| \cdot |G_{j-1}| \right)} \\ &\leq \gamma_m + \frac{6}{\sqrt{n}} \sum_{j=1}^{m} \gamma_j \sqrt{\ln \mathcal{N}(\gamma_j, \mathcal{G})} \\ &= \gamma_m + \frac{12}{\sqrt{n}} \sum_{j=1}^{m} (\gamma_j - \gamma_{j+1}) \sqrt{\ln \mathcal{N}(\gamma_j, \mathcal{G})} \\ &\leq \gamma_m + \frac{12}{\sqrt{n}} \int_{\gamma_{m+1}}^{\gamma_0} \sqrt{\ln \mathcal{N}(\beta, \mathcal{G})} d\beta\end{aligned}$$

The first inequality is just a combination of the foregoing bounds, the second uses $|G_{j-1}| \leq |G_j$ and that G_j is a minimal γ_j-covering, the equality comes from the definition of γ_j and the last inequality uses that the integral is lower bounded by its lower Riemann sum.

Finally, we need to choose m. Therefore, for any fixed $\alpha \in [0, \gamma_0/2]$ one chooses m such that $\alpha < \gamma_{m+1} < 2\alpha$ holds. Because of the definition of γ_j, we have $\gamma_j \leq 4\alpha$ which terminates the proof. □

Another very helpful inequality we will use later is presented in the following proposition.

Proposition 2.11 (PAC Bound using the Rademacher Complexity)**.** *Assume the loss l is L_1-Lipschitz and upper bounded by C_1. With probability at least $1 - \delta$ over a sample $\mathcal{S}$ of size n, every $h \in \mathcal{H}_{QNN}$ satisfies*

$$R(\mathcal{A}(S)) \leq \hat{R}_S(\mathcal{A}(S)) + 2L_1 \mathfrak{R}(\mathcal{F}) + 3C_1 \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

3 Quantum Computation

Quantum computation, in contrast to classical computer science, uses the phenomena of quantum mechanics to perform calculations. Instead of bits and electrical circuits, quantum mechanical systems are used [8].
Therefore, we will have a brief recap of the basics of quantum theory in this chapter in which the important concepts are motivated by the postulates of quantum mechanics. For this section we will focus mainly on the lecture notes from Reinhard Werner on "Mathematical methods of quantum information theory" [9],[10],[11],[12] unless mentioned otherwise.

3.1 Operators on Hilbert Spaces

The first step is to stake out the mathematical background.

Postulate 1

Every isolated quantum mechanical system is assigned a *Hilbert space* $\mathcal{H}$.

A Hilbert space is a complex vector space with a scalar product $\langle\cdot,\cdot\rangle : \mathcal{H}\times\mathcal{H}\to\mathbb{C}$ which is complete w.r.t. the norm $\|\cdot\|^2 = \langle\cdot,\cdot\rangle$ (induced by the scalar product).
Considering several quantum mechanical systems respectively their Hilbert spaces, the scalar product or norm will often be indexed with the corresponding space.

In QM, we deal with (linear) *operators*, i.e. linear maps $A : \mathcal{H}_1 \to \mathcal{H}_2$. QM is by construction a local theory and since linearity is strongly related to locality, linearity is always needed [14]. Therefore, we will just write operators instead of linear operators.

We say an operator is *bounded*, and write $A \in \mathcal{B}(\mathcal{H}_1,\mathcal{H}_2)$, if there exists a constant $c \geq 0$ s.t.

$$\|A\psi\|_{\mathcal{H}_2} \leq c\cdot\|\psi\|_{\mathcal{H}_1} \quad \forall\psi\in\mathcal{H}_1.$$

If $A : \mathcal{H}\to\mathcal{H}$ is bounded on $\mathcal{H}$, we write $A \in \mathcal{B}(\mathcal{H})$. For $A \in \mathcal{B}(\mathcal{H}_1,\mathcal{H}_2)$ the *adjoint* $A^\dagger \in \mathcal{B}(\mathcal{H}_2,\mathcal{H}_1)$ is defined by

$$\langle\phi, A^\dagger\psi\rangle_{\mathcal{H}_1} = \langle A\phi,\psi\rangle_{\mathcal{H}_2} \quad \forall\phi\in\mathcal{H}_1,\psi\in\mathcal{H}_2.$$

An operator $U \in \mathcal{B}(\mathcal{H}_1,\mathcal{H}_2)$ is unitary, if $UU^\dagger = \mathbb{1}_{\mathcal{H}_2}$ and $U^\dagger U = \mathbb{1}_{\mathcal{H}_1}$.

We say an operator is *positive (semi definite)* $(A \geq 0)$ if $\langle\phi, A\phi\rangle \geq 0 \quad \forall\phi\in\mathcal{H}$.

Now we have some basic definitions and will go over to the description of simple physical experiments.

3.2 Preparation and Measurement

Physical experiments can be split into a preparation and a measurement phase.

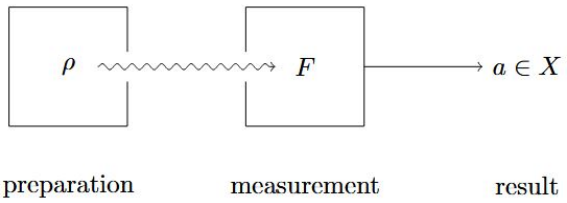

Figure 2: Principle of preparation and measurement. A state $\rho \in \mathcal{D}(\mathcal{H})$ is prepared and gets measured by an observable $F = \{F_a\}_{a \in X}$ on $\mathcal{H}$. The result of this process is $a \in X$ with probability $\mathbb{P}(a) = \mathrm{Tr}(\rho F_a)$.

First we look at the preparation.

Postulate 2

In quantum mechanics, the states $\rho \in \mathcal{D}(\mathcal{H})$ are given by density operators. They describe the preparation in all facets that can be measured afterwards.

A density operator is a positive operator $\rho \in \mathcal{T}(\mathcal{H})$ with $\mathrm{Tr}(\rho) = 1$. The set of density operators on a Hilbert space $\mathcal{H}$ therefore is

$$\mathcal{D}(\mathcal{H}) := \{\rho \in \mathcal{T}(\mathcal{H}) | \rho \geq 0, \mathrm{Tr}(\rho) = 1\}.$$

Now we look at the measurement.

Postulate 3

A detector, i.e. a measurement apparatus with outcomes "a" or "not a" is described by an operator $0 \leq F_a \leq \mathbb{1}$, $F : \mathcal{H} \to \mathcal{H}$.

Postulate 4

The probability of the outcome $a \in X$ is given by $\mathbb{P}(a) = \mathrm{Tr}(\rho \cdot F_a)$.

To be more precise, for every possible outcome $a \in X$ there is assigned a so called *effect operator* $F_a \in \mathcal{B}(\mathcal{H})$. The assumption $0 \leq F_a \leq \mathbb{1}$ ensures that the probabilities are contained in the interval $[0, 1]$, since density operators have trace 1 by definition. Because $\rho \in \mathcal{T}(\mathcal{H})$ and the trace class is an ideal, the probabilities are well defined. Obviously the probability of all possible outcomes should sum up to 1 which requires

$$\sum_{a \in X} F_a = \mathbb{1}.$$

In case X is not countable the sum is replaced by an integral.
An *observable* or *POVM (positive operator valued measurement)* is a labelled collection $F = \{F_a\}_{a \in X}$.

3.3 Qubits

The fundamental concept of classical computation is the bit. It takes values 0 and 1 to store information. The quantum mechanical analogue is the *Qubit* or Quantum bit. Here the basis vectors can have labels 0 or 1. The mathematical equivalent are 2- dimensional Hilbert spaces. These are isomorphic to $\mathbb{C}^2$ i.e. $\mathcal{H} \simeq \mathbb{C}^2$.

3.4 Composite Systems

Now we look at the description of composite systems in quantum mechanics.

Postulate 5

If we have two quantum mechanical systems described by $\mathcal{H}_A$ and $\mathcal{H}_B$, the joint system is described by $\mathcal{H}_A \otimes \mathcal{H}_B$, where $\otimes$ denotes the tensor product.

The space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$ is again a Hilbert space spanned by vectors $\phi \otimes \psi$, with $\phi \in \mathcal{H}_A$, $\psi \in \mathcal{H}_B$. The scalar product is defined by

$$\langle \phi \otimes \psi, \xi \otimes \nu \rangle = \langle \phi, \xi \rangle_A \langle \psi, \nu \rangle_B.$$

3.5 Time Evolution

Up to this point, we just described preparation and measurement in the quantum mechanic case. However, we did not consider that the system evolves. Assume that we prepare a state, let the system evolve or do some operation on it and then measure. By this procedure, we could even end up in a different Hilbert space.

Postulate 6

Every system is further characterised by a Hamiltonian $H : \mathcal{H} \to \mathcal{H}$, which is a self-adjoined operator on $\mathcal{H}$. The Hamiltonian generates time-evolution via

$$U_t := \exp\left(-\tfrac{it}{\hbar} H\right).$$

If one applies a measurement apparatus F after t units of time passed on the prepared state ρ we obtain "yes" with probability

$$\mathbb{P}[\text{"yes after } t\text{"}] = \mathrm{Tr}(U_t \rho U_t^\dagger F) = \mathrm{Tr}(\rho U_t^\dagger F U_t)$$

The last equality follows from $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. Both sides are obviously equivalent but correspond to different perceptions of time evolution.
On the left hand side we introduce a time-dependent density matrix $\rho_t = U_t \rho U_t^\dagger$ and consider the observable as time- independent. This is called the *Schrödinger picture*.
In contrast on the right hand side we introduce a time-dependent observable $F_t = U_t^\dagger F U_t$ and consider the density matrix as time independent. This is called the *Heisenberg picture*.

3.6 Norms and Fidelity

As we have seen in Section (2), it is essential for this work, to have a proper distance measure whether it is to quantify errors in order to find a loss function s.t. the empirical risk can be minimized or to gain PAC bounds. To extend this onto quantum computation, we introduce suitable norms on $\mathcal{B}(\mathcal{H})$.

The *operator norm* is defined by

$$\|A\| = \sup_{\psi\in\mathcal{H},\|\psi\|_{\mathcal{H}}=1} \|A\psi\|_{\mathcal{H}}$$

or equivalently

$$\|A\| = \sup_{\psi\in\mathcal{H},\psi\neq 0} \frac{\|A\psi\|_{\mathcal{H}}}{\|\psi\|_{\mathcal{H}}} = \sup_{\psi\in\mathcal{H},\|\psi\|_{\mathcal{H}}\leq 1} \|A\psi\|_{\mathcal{H}}$$

for $A \in \mathcal{B}(\mathcal{H})$. A family of norms that we could use and includes the operator norm are the *Schatten-p-norms*, for $A \in \mathcal{B}(\mathcal{H})$ defined by

$$\|A\|_p = (\operatorname{Tr}|A|^p)^{\frac{1}{p}},$$

for $1 \leq p \leq \infty$, where

$$|A| = (A^\dagger A)^{\frac{1}{2}}.$$

In addition

$$\|A\|_\infty = \|A\|$$

yields, with $\|\cdot\|$ being the operator norm.

Another useful tool, when it comes to distance measure, is the *fidelity*. If one wants to measure the distance of a pure state ρ to a pure state $|\psi\rangle\langle\psi|$, the fidelity is defined by

$$F(\psi,\rho) = \langle\psi|\rho\psi\rangle\,.$$

The fidelity takes values between 0 and 1 and is optimal, obviously if the state ρ is equal to the pure state, where it takes the value 1. In order to minimize errors, one has, in contrast to norms, to maximize the fidelity.
The fidelity for two mixed states is defined as

$$F(\rho_1,\rho_2) = \left(\operatorname{Tr}((\sqrt{\rho_1}\rho_2\sqrt{\rho_1})^{1/2})\right), \tag{13}$$

with $\rho_1,\rho_2 \in \mathcal{D}(\mathcal{H})$. [6].

3.7 Variational Quantum Algorithm (VQA)

One aim of quantum computation is to develop learning methods that are superior to classical methods. A candidate to achieve this goal are *Variational Quantum Algorithm's* (VQA's). Therefore, we want to look at the basic concepts of such algorithms and focus on [3].

Specifically, VQA contains two subroutines, an N-qudit quantum circuit and a classical optimizer. In the training stage it follows an iterative approach. The output of the quantum circuit is permanently used by the optimizer to update the trainable parameters of the adopted approaches. Of course, the goal is to minimize the loss function L.

We define the N-qubit input quantum state as $\rho \in \mathcal{B}(\mathbb{C}^{d^N}) = \mathbb{C}^{d^N \times d^N}$, the quantum observable as $O \in \mathbb{C}^{d^N \times d^N}$ and the applied approaches as $\hat{U}(\theta) = \prod_{l=1}^{N_g} \hat{u}_l(\theta) \in \mathcal{U}(d^N)$. Hereby $\theta \in \Theta$ are trainable parameters living in the parameter space Θ which will be updated by the optimizer, $\hat{u}_l(\theta) \in \mathcal{U}(d^k)$ refers to the l-th quantum gate operated with at most k-qudits (with $k < N$), and $\mathcal{U}(d^N)$ refers to the unitary group in d^N dimension.

We are able to specify the applied approaches $\hat{U}(\theta)$ even further. In general, not all quantum gates are trainable such that there are N_{gt} trainable and $N_g - N_{gt}$ fixed gates.
Under these circumstances the output of the quantum circuit is

$$h(\theta^{(t)}) := \mathrm{Tr}\left(\hat{U}(\theta^{(t)})^\dagger O \hat{U}(\theta^{(t)})\rho\right). \tag{14}$$

The aim of VQA's is like in classical scenarios to find a good hypothesis which is an element of the hypothesis class

$$\mathcal{F} = \left\{\mathrm{Tr}\left(\hat{U}(\theta)^\dagger O \hat{U}(\theta)\rho\right) \middle| \theta \in \Theta\right\}. \tag{15}$$

4 PAC Bounds for Quantum Neural Networks

In this section we will look at quantum neural networks (QNN's), more precisely we will find PAC bounds for them. In order to do this, in section(4.1) we will bound the covering number of the hypothesis space of the VQA's introduced previously. This is just some preliminary work for the last two sections. Here, we will look on two different structures of QNN's and finally we will be able to find suitable PAC bounds for them. This part roughly follows [3].

4.1 VQA's

In this section we will focus on the structures introduced in section(3.7). We will quantify the expressivity of VQA's by the complexity of its hypothesis space $\mathcal{F}$ using the covering number(2.3) as a complexity measure.

In order to bound the covering number of the hypothesis space
$\mathcal{F} = \left\{\mathrm{Tr}\left(\hat{U}(\theta)^\dagger O \hat{U}(\theta)\rho\right) \middle| \theta \in \Theta\right\}$, we need to introduce the operator group

$$\mathcal{F}_{circ} := \left\{\hat{U}(\theta)^\dagger O \hat{U}(\theta) | \theta \in \Theta\right\}, \tag{16}$$

where $\hat{U}(\theta) = \prod_{i=1}^{N_g} \hat{u}_i(\theta_i)$ and only $N_{gt} < N_g$ quantum gates are trainable. Later, we will use this covering number to upper bound $\mathcal{F}$'s covering number and eventually gain PAC bounds. The following proposition is a necessary step to reach this goal.

Proposition 4.1 (Covering Number of the Unitary Group)**.** *For $\epsilon > 0$, the ϵ-covering number for the unitary group $\mathcal{U}(d^k)$ obeys*

$$\left(\frac{3}{4\epsilon}\right)^{d^{2k}} \leq \mathcal{N}(\mathcal{U}(d^k), \epsilon, \|\cdot\|) \leq \left(1 + \frac{2\pi}{\epsilon}\right)^{d^{2k}} \tag{17}$$

where $\|O\|$ denotes the operator norm.

Using this result we are able to upper bound the covering number of the operator group $\mathcal{F}_{circ}$.

Lemma 4.2 (Covering Number of $\mathcal{F}_{circ}$)**.** *The ϵ-covering number for the operator group $\mathcal{F}_{circ}$ obeys*

$$\mathcal{N}(\mathcal{F}_{circ}, \epsilon, \|\cdot\|) \leq \left(1 + \frac{4\pi N_{gt}\|O\|}{\epsilon}\right)^{d^{2k} N_{gt}} \tag{18}$$

Proof. The operator group is defined as $\mathcal{F}_{circ} = \left\{\hat{U}(\theta)^\dagger O \hat{U}(\theta) | \theta \in \Theta\right\}$. Here $\hat{U}(\theta) = \prod_{i=1}^{N_g} \hat{u}_i(\theta_i)$ is the trainable unitary consisting of N_{gt} trainable gates and $N_g - Ngt$

fixed gates. To determine the covering number of $\mathcal{F}_{circ}$ one considers a fixed minimal ϵ-covering $\mathcal{S}$ for the set of all possible gates. Furthermore define the set

$$\tilde{\mathcal{S}} = \left\{ \prod_{j\in\{N_g - N_{gt}\}} \hat{u}_j \prod_{i\in\{N_{gt}\}} \hat{u}_i(\theta_i) | \hat{u}_i(\theta_i) \in S \right\}, \tag{19}$$

where $\hat{u}_j$ corresponds to the fixed and $\hat{u}_i(\theta_i)$ to the trainable gates.
For any circuit $\hat{U}(\Theta) = \prod_{i=1}^{N_g} \hat{u}_i(\theta_i)$ one always finds a $\hat{U}_\epsilon(\theta) \in \tilde{S}$, where each trainable gate $\hat{u}_i(\theta_i)$ is replaced with the nearest element in S (the covering set).
Now we examine the discrepancy. For better overview the θ-dependencies are suppressed.

$$\begin{aligned}
\|\hat{U}^\dagger O\hat{U} - \hat{U}_\epsilon^\dagger O\hat{U}_\epsilon\| =& \|\hat{U}^\dagger O(\hat{U} - \hat{U}_\epsilon) + (\hat{U}^\dagger - \hat{U}_\epsilon^\dagger)O\hat{U}_\epsilon\| \\
\leq& \|\hat{U}O(\hat{U} - \hat{U}_\epsilon^\dagger)\| + \|(\hat{U}^\dagger - \hat{U}_\epsilon^\dagger)O\hat{U}_\epsilon\| \\
=& \|O(\hat{U} - \hat{U}_\epsilon)\| + \|(\hat{U}^\dagger - \hat{U}_\epsilon^\dagger)O\| \\
\leq& \|O\|\|\hat{U} - \hat{U}_\epsilon\| + \|O\|\|\hat{U}^\dagger - \hat{U}_\epsilon^\dagger\| \\
=& 2\|O\|\|\hat{U} - \hat{U}_\epsilon\| \\
\leq& 2N_{gt}\epsilon\|O\|
\end{aligned}$$

The first inequality uses the Cauchy-Schwarz inequality, the second inequality uses the sub-multiplicity of the operator norm. The third inequality uses the ϵ-covering, more precisely, that $\|\hat{u}_i - \hat{u}_i^{(\epsilon)}\| \leq \epsilon$ implies $\|\hat{U} - \hat{U}_\epsilon\| \leq N_{gt}\epsilon$ with $\hat{U}_\epsilon = \prod_{i\in\{N_{gt}\}} \hat{u}_i^{(\epsilon)}$.

Thus, by Def.(2.3) we know $\tilde{\mathcal{S}}$ is a $2N_{gt}\|O\|\epsilon$-covering set for $\mathcal{F}_{circ}$.

The upper bound in Prop.(4.1) implies $|S| \leq \left(1 + \frac{2\pi}{\epsilon}\right)^{d^{2k}}$, since S is the covering set of all possible gates. There are $|S|^{N_{gt}}$ combinations for the gates in $\tilde{\mathcal{S}}$, because we consinder N_{gt} trainable gates, which implies $\tilde{\mathcal{S}} \leq \left(1 + \frac{2\pi}{\epsilon}\right)^{d^{2k}N_{gt}}$. Thus the covering number of $\mathcal{F}_{circ}$ is upper bounded by

$$\mathcal{N}(\mathcal{F}_{circ}, 2N_{gt}\|O\|\epsilon, \|\cdot\|) \leq \left(1 + \frac{2\pi}{\epsilon}\right)^{d^{2k}N_{gt}}$$

which is equivalent to

$$\mathcal{N}(\mathcal{F}_{circ}, \epsilon, \|\cdot\|) \leq \left(1 + \frac{4\pi N_{gt}\|O\|}{\epsilon}\right)^{d^{2k}N_{gt}}. \tag{20}$$

□

Using this result we are finally able to bound the covering number of the hypothesis space $\mathcal{F}$.

Theorem 4.3 (Covering Number of the $\mathcal{F}$)**.** *The covering number of the hypothesis space $\mathcal{F}$ in Eq.(15) yields*

$$\mathcal{N}(\mathcal{F}, \epsilon, |\cdot|) \leq \left(1 + \frac{4\pi N_{gt}\|O\|}{\epsilon}\right)^{d^{2k} N_{gt}}. \tag{21}$$

Proof. The idea is to use Lem.(2.6) to upper bound the covering number of $\mathcal{F}$ by the covering number of $\mathcal{F}_{circ}$. Therefore one sets in the mentioned lemma $\mathcal{H}_2$ as hypothesis space $\mathcal{F}$ and $\mathcal{H}$ as the operator group $\mathcal{F}_{circ}$. We take a covering set and define $\hat{U}_\epsilon$ as quantum circuit, where each of the gates is replaced with the nearest element in the covering set.
With this, we are able to use Lem. (2.6) and derive the constant K.

$$\begin{aligned} d_2(\mathrm{Tr}(\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon \rho), \mathrm{Tr}(\hat{U}^\dagger O \hat{U} \rho)) &= |\mathrm{Tr}(\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon \rho) - \mathrm{Tr}(\hat{U}^\dagger O \hat{U} \rho)| \\ &= |\mathrm{Tr}\left((\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon - \hat{U}^\dagger O \hat{U})\rho\right)| \\ &\leq \|\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon - \hat{U}^\dagger O \hat{U}\| \,\mathrm{Tr}(\rho) \\ &= d_1(U_\epsilon^\dagger O \hat{U}_\epsilon, \hat{U}^\dagger O \hat{U}) \end{aligned}$$

The first equality follows from the explicit form of the hypothesis space $\mathcal{F}$, the second uses that the trace is a linear map, the first inequality follows from the cauchy schwarz inequality, and the last equality uses $\mathrm{Tr}(\rho) = 1$ and the explicit form of the operator group $\mathcal{F}_{circ}$, i.e

$$\|\hat{U}_\epsilon^\dagger O \hat{U}_\epsilon - \hat{U}^\dagger O \hat{U}\| = d_1(U_\epsilon^\dagger O \hat{U}_\epsilon, \hat{U}^\dagger O \hat{U}).$$

The upper equation determines K=1. Using Lemma (4.2) we obtain

$$\mathcal{N}(\mathcal{F}, \epsilon, |\cdot|) \leq \mathcal{N}(\mathcal{F}_{circ}, \epsilon, \|\cdot\|) \leq \left(1 + \frac{4\pi N_{gt}\|O\|}{\epsilon}\right)^{d^{2k} N_{gt}}.$$

□

4.2 Quantum Neural Networks

In the previous section we gained a bound for the covering number of the hypothesis space $\mathcal{F}$. We are going to use this bound in the context of QNN's to find PAC bounds. There are several possibilities to implement different QNN structures. We will look at the QNN's presented in [3] and begin with a recap on the general setting.

4.2.1 Framework

The aim of quantum machine learning as well as in classical machine learning is devising an algorithm $\mathcal{A}$ such that with a given training set S, $\mathcal{A}$ is able to infer a hypothesis h from its hypothesis space to minimize the risk, or the empirical risk.

We will employ quantum neural networks (QNN) to implement the learning algorithm $\mathcal{A}_{QNN}$ to minimize the empirical risk. As mentioned above there are more than one type of QNN's. In this section we will focus on [3] .

The mechanism is as follows. Given a classical parameter $x^{(i)}$, an input state ρ is prepared that contains $x^{(i)}$ in some way. After the preparation is done, the unitaries $U(\hat{\theta}^t)$ are applied to the state. Then finally the state is measured with a predefined quantum measurement O.
Hence, the explicit form of a hypothesis is

$$h_{\mathcal{A}_{QNN}(S)}(x^{(i)}) = \mathrm{Tr}\left(\hat{U}(\theta^{(t)\dagger} O \hat{U}(\theta)^{(t)} \rho_{x^{(i)}}\right),$$

where $\theta^{(t)}$ are the updated parameters. The hypothesis space is

$$\mathcal{H}_{QNN} = \left\{h_{A_{QNN}(S)}(\cdot) \middle| \theta \in \Theta\right\}$$

Because of the explicit structure of the hypothesis space, we can use the bounds on the covering number of $\mathcal{F}$ and $\mathcal{F}_{circ}$ gained before.

4.2.2 PAC Bounds via Dudley's Theorem

Now we want to find an estimation for the difference between the true and the empirical risk. In this section we exploit Prop.(2.11) in order to obtain such a PAC bound.
Obviously before we can make use of Prop.(2.11) we need some upper bound on the Rademacher complexity of $\mathcal{H}_{QNN}$. However we can use Dudley's Theorem (2.10) to gain control.

Lemma 4.4 (Rademacher Complexity of $\mathcal{H}_{QNN}$)**.** *The Rademacher complexity* $\mathfrak{R}(\mathcal{H}_{QNN})$ *with* $\mathcal{H}_{QNN}$ *as before is given by*

$$\begin{aligned}\mathfrak{R}(\mathcal{H}_{QNN}) \leq \frac{4}{\sqrt{n}} + \frac{12}{\sqrt{n}} d^{2k} N_{gt}\big[&(1 + 4\pi\sqrt{n}N_{gt}\|O\|) \cdot \ln(1 + 4\pi\sqrt{n}N_{gt}\|O\|) \\ &- (\frac{1}{\sqrt{n}} + 4\pi\sqrt{n}N_{gt}\|O\|) \cdot \ln(1 + 4\pi n N_{gt}\|O\|) + 4\pi\sqrt{n}N_{gt}\|O\| \ln(\sqrt{n})\big].\end{aligned}$$

Proof. As mentioned before, we use Dudley's Theorem (2.10), which connects the Rademacher complexity with the covering number and therefore allows us to use the bounds gained before on the covering number.

We shortly recall the inequality

$$\hat{\mathfrak{R}}(\mathcal{H}_{QNN}) \leq \inf_{\alpha\in[0,\gamma_0/2]} \left(4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{\gamma_0} d\beta \sqrt{\ln \mathcal{N}((\mathcal{H}_{QNN})_{|S}, \beta, \|\cdot\|_2)}\right).$$

To bound $\mathcal{N}(\mathcal{H}_{QNN_{|S}}, \beta, \|\cdot\|_2)$, consider a minimal $\frac{\beta}{\sqrt{n}}$-covering $\mathcal{S}$ of $\mathcal{H}_{QNN_{|x^{(i)}}}$. Then for any $h \in \mathcal{H}_{QNN_{|x^{(i)}}}$ there exists an $h' \in \mathcal{S}$ such that

$$|h_{\mathcal{A}_{QNN}(S)}(x^{(i)}) - h'_{\mathcal{A}_{QNN}(S)}(x^{(i)})| \leq \frac{\beta}{\sqrt{n}} \quad \forall i \in [n].$$

Now look at the discrepancy

$$\begin{aligned}
&\left\| \left[h_{\mathcal{A}_{QNN}(S)}(x^{(i)})\right]_{i=1:n} - \left[h'_{\mathcal{A}_{QNN}(S)}(x^{(i)})\right]_{i=1:n} \right\|_2 \\
&= \left\| \begin{pmatrix} h_{\mathcal{A}_{QNN}(S)}(x^{(1)}) \\ \dots \\ h_{\mathcal{A}_{QNN}(S)}(x^{(n)}) \end{pmatrix} - \begin{pmatrix} h'_{\mathcal{A}_{QNN}(S)}(x^{(1)}) \\ \dots \\ h'_{\mathcal{A}_{QNN}(S)}(x^{(n)}) \end{pmatrix} \right\|_2 \\
&= \sqrt{\sum_{i=1}^{n} |h_{A\text{-}QNN(S)}(x^{(i)}) - h'_{A_{QNN}(S)}(x^{(i)})|^2} \\
&\leq \beta
\end{aligned}$$

The first two equations exploit the definition and the inequality follows directly from $\mathcal{S}$ being an $\frac{\beta}{\sqrt{n}}$-covering.

One defines $\mathcal{S}' := \left\{ \begin{pmatrix} h'_{\mathcal{A}_{QNN}(S)}(x^{(1)}) \\ \dots \\ h'_{\mathcal{A}_{QNN}(S)}(x^{(n)}) \end{pmatrix} h' \in \mathcal{S} \right\}$ and see directly that $|S'| = |S|$ holds.

Thus S' is an β-covering for $\mathcal{H}_{QNN_S}$, which gives

$$\begin{aligned}
\ln\left(\mathcal{N}(\mathcal{H}_{QNN_{|S}}, \beta, \|\cdot\|_2)\right) &\leq \ln\left(\mathcal{N}(\mathcal{H}_{QNN_{|x^{(i)}}}, \frac{\epsilon}{\sqrt{n}}, |\cdot|)\right) \\
&\leq \ln\left(\mathcal{N}(\mathcal{F}_{circ}, \frac{\beta}{\sqrt{n}}, \|\cdot\|)\right) \\
&\leq \ln\left(1 + \frac{4\pi\sqrt{n}N_{gt}\|O\|}{\beta}\right)^{d^{2k}N_{gt}} \\
&= d^{2k}N_{gt} \cdot \ln\left(1 + \frac{4\pi\sqrt{n}N_{gt}\|O\|}{\beta}\right).
\end{aligned}$$

Using this and realising that $\gamma_0 = 1$ which is true because the restriction on the allowed evolution of quantum systems that ensures the sum of probabilities of all possible outcomes of any event always equals 1. Therefore, one is able to bound the second term in Eq.(4.2.2).

$$\begin{aligned}
&\frac{12}{\sqrt{n}}\int_\alpha^1 d\beta\sqrt{\ln\mathcal{N}((\mathcal{H}_{QNN})_{|S}, \beta, \|\cdot\|_2)} \\
&\leq \frac{12}{\sqrt{n}}\int_\alpha^1 d\beta\sqrt{d^{2k}N_{gt}\ln\left(1 + \frac{4\pi\sqrt{n}N_{gt}\|O\|}{\beta}\right)} \\
&\leq \frac{12}{\sqrt{n}}\int_\alpha^1 d\beta N_{gt}d^{2k}\ln\left(1 + \frac{4\pi\sqrt{n}N_{gt}\|O\|}{\beta}\right) \\
&= \frac{12}{\sqrt{n}}d^{2k}N_{gt}\left[(4\pi\sqrt{n}N_{gt}\|O\| + \beta)\cdot\ln(1 + \frac{4\pi\sqrt{n}N_{gt}\|O\|}{\beta}) + 4\pi\sqrt{n}N_{gt}\|O\|\ln(\beta)\right]_\alpha^1 \\
=&\frac{12}{\sqrt{n}}d^{2k}N_{gt}\Bigg((4\pi\sqrt{n}N_{gt}\|O\| + 1)\cdot\ln(1 + 4\pi\sqrt{n}N_{gt}\|O\|) \\
&- (4\pi\sqrt{n}N_{gt}\|O\| + \alpha)\cdot\ln\left(1 + \frac{4\pi\sqrt{n}N_{gt}\|O\|}{\alpha}\right) - 4\pi\sqrt{n}N_{gt}\|O\|\ln(\alpha)\Bigg)
\end{aligned}$$

Here, the first inequality follows from exploiting the upper bound of the covering number of $\mathcal{H}_{QNN}$ and the second uses the monotony of the integral.

For simplicity, we set $\epsilon = \frac{1}{\sqrt{n}}$. Then the Rademacher complexity is upper bounded by

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}_{QNN}) \leq &\frac{4}{\sqrt{n}} + \frac{12}{\sqrt{n}}d^{2k}N_{gt}\big[(1 + 4\pi\sqrt{n}N_{gt}\|O\|)\cdot\ln(1 + 4\pi\sqrt{n}N_{gt}\|O\|) \\
&- (\frac{1}{\sqrt{n}} + 4\pi\sqrt{n}N_{gt}\|O\|)\cdot\ln(1 + 4\pi nN_{gt}\|O\|) + 4\pi\sqrt{n}N_{gt}\|O\|\ln(\sqrt{n})\big]
\end{aligned}$$

□

Theorem 4.5 (PAC Bound for $\mathcal{H}_{QNN}$)**.** *Assume the loss L is L_1-Lipschitz and upper bounded by C_1. With probability at least $1-\delta$ with $\delta \in (0,1)$ we have*

$$\begin{aligned}\mathcal{R}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S)) \leq \\ 2L_1 \cdot \Bigg(\frac{4}{\sqrt{n}} + \frac{12}{\sqrt{n}} d^{2k} N_{gt} \big[(1 + 4\pi\sqrt{n} N_{gt} \|O\|) \cdot \ln(1 + 4\pi\sqrt{n} N_{gt}\|O\|) \\ - (\frac{1}{\sqrt{n}} + 4\pi\sqrt{n} N_{gt}\|O\|) \cdot \ln(1 + 4\pi n N_{gt}\|O\|) + 4\pi\sqrt{n} N_{gt}\|O\| \ln(\sqrt{n}) \big] \Bigg) \\ + 3C_1 \sqrt{\frac{\ln(2/\delta)}{2n}}.\end{aligned}$$

Proof. For the proof we just insert the bound for the Rademacher complexity gained in Lem.(4.4) into the inequality in Prop.(2.11). Then we gain the stated term. □

4.2.3 PAC Bounds via Uniform Covering Number

Consider $\mathcal{F} = \left\{ h : \mathcal{D}(\mathcal{H}) \to \mathbb{R}, h(\rho) = \mathrm{Tr}\left(\hat{U}(\theta)^\dagger O \hat{U}(\theta)\rho\right) | \theta \in \Theta \right\}$ and $\mathcal{F}_{circ} = \left\{ h : \mathcal{D}(\mathcal{H}) \to \mathcal{D}(\mathcal{H}) | h(\rho) \to \hat{U}(\theta)^\dagger O \hat{U}(\theta) | \theta \in \Theta \right\}$ where $\rho \in \mathcal{D}(\mathcal{H})$ is an element of the unbounded operators.

In the previous section we gained a PAC bound by using Lem.(2.11). Now we follow a different path, more precisely we will use Thm.(2.8). Obviously this requires knowledge of the uniform covering number of some space $\mathcal{G}$ we have not defined yet. However, we will develop ways to get around these issues in such a manner that previous results can be used.

First of all, we present an inequality connecting the uniform covering number of $\mathcal{F}$ with the covering number of $\mathcal{F}_{circ}$.

Lemma 4.6 (Relation between Uniform and Covering Number)**.** *The relation between the uniform covering number of $\mathcal{F}$ and the ϵ-covering number of $\mathcal{F}_{circ}$ is given by*

$$\Gamma_1(n, \epsilon, \mathcal{F}) \leq \mathcal{N}(\mathcal{F}_{circ}, \epsilon, \|\cdot\|). \tag{22}$$

Proof. At first, notice that for all $f \in \mathcal{F}$ there exists a $f_{circ} \in \mathcal{F}_{circ}$ such that $f(\rho) = \mathrm{Tr}(f_{circ}(\rho))$, $\forall \rho \in \mathcal{D}(\rho)$. Furthermore, let $\tilde{\mathcal{F}}_{circ}$ be a minimal ϵ-covering. Define

$\tilde{\mathcal{F}} = \left\{\tilde{f} : \mathcal{D}(\mathcal{H}_{Hil}) \to \mathbb{R}\right\}$ and look at the discrepancy between $f \in \mathcal{F}$ and $\tilde{f} \in \tilde{\mathcal{F}}$.

$$\begin{aligned}
\|f - \tilde{f}\|_{1,(\rho_-)_{i=1}^n} &= \frac{1}{n}\sum_{i=1}^{n} |f(\rho_i) - \tilde{f}(\rho_i)| \\
&= \frac{1}{n}\sum_{i=1}^{n} |\operatorname{Tr}\left(f_{circ}(\rho_i) - \tilde{f}_{circ}(\rho_i)\right)| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \|U(\theta)^\dagger O U - \tilde{U}(\theta)^\dagger O \tilde{U}(\theta)\| \operatorname{Tr}(\rho_i) \\
&= \|U(\theta)^\dagger O U - \tilde{U}(\theta)^\dagger O \tilde{U}(\theta)\| \quad \leq \epsilon
\end{aligned}$$

The first step is just the definition, the second exploits the precondition, the third uses the Hölder inequality and structure of $\mathcal{F}_{circ}$ and the last follows because $\tilde{\mathcal{F}}_{circ}$ is an ϵ-covering.

The Uniform covering number yields

$$\begin{aligned}
\Gamma_1(n, \epsilon, \mathcal{F}) &= \max_{(\rho_i)_{i=1}^n \in (\mathcal{D})^n} \mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_{1,\rho_{i=1}^n}) \\
&\leq \max_{(\rho_i)_{i=1}^n \in (\mathcal{D})^n} \mathcal{N}(\mathcal{F}_{circ}, \epsilon, \|\cdot\|) \\
&= \mathcal{N}(\mathcal{F}_{circ}, \epsilon, \|\cdot\|).
\end{aligned}$$

□

Finally, we have everything together to use Lem.(2.8) in order to gain a PAC bound.

Theorem 4.7 (PAC Bound for the QNN's)**.** *Consider $\mathcal{F}$ as before. Assume that the loss Function $L : \mathcal{X} \times \mathcal{Y} \to [0, c]$ is lipschitz continuous with constant l. Then for any $\epsilon > 0$ and any probability measure P on $\mathcal{X} \times \mathcal{Y}$ the following inequality yields*

$$\mathbb{P}_{S\sim\mathcal{F}^n}[\exists h \in \mathcal{F} : |R(h) - \hat{R}(h)| \geq \epsilon] \leq 4 \cdot \left(1 + \frac{32\pi l N_{gt} \|O\|}{\epsilon}\right)^{d^{2k} N_{gt}} \exp\left(-\frac{n\epsilon}{32c^2}\right). \quad (23)$$

Proof. We define $\mathcal{G} := \{g : \mathcal{X} \times \mathcal{Y} \to [0, c] | \exists h \in \mathcal{F} : g(x, y) = L(h(x), y)\}$ and recognize that all conditions in order to use Thm.(2.8) are fulfilled. Hence for any $\epsilon > 0$ and

any probability measure P on $\mathcal{X} \times \mathcal{Y}$ we have

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{F}^n}[\exists h \in \mathcal{F} : |R(h) - \hat{R}(h)| \geq \epsilon] &\leq 4\Gamma_1(2n, \epsilon/8, \mathcal{G}) \exp\left(-\frac{n\epsilon}{32c^2}\right) \\ &\leq 4\Gamma_1(2n, \frac{\epsilon}{8l}, \mathcal{F}) \exp\left(-\frac{n\epsilon}{32c^2}\right) \\ &\leq 4\mathcal{N}(\mathcal{F}_{circ}, \frac{\epsilon}{8l}, \|\cdot\|) \exp\left(-\frac{n\epsilon}{32c^2}\right) \\ &\leq 4\left(1 + \frac{32\pi l N_{gt}\|O\|}{\epsilon}\right)^{d^{2k}N_{gt}} \exp\left(-\frac{n\epsilon}{32c^2}\right).\end{aligned}$$

For the second step, we notice that $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$ and use that the loss function L is lipschitz continuous such that we can use Lem.(2.9). The third step uses Lem.(4.6) and the last exploits the bound on the covering number of $\mathcal{F}_{circ}$ as presented in Lem.(4.2). □

4.3 Dissipative Quantum Neural Networks

In this section we will focus on [2]. Here another approach for QNN's, namely dissipative QNN's, are discussed and again our aim is to find PAC bounds for its hypothesis class.

The hypothesis class is

$$\mathcal{J} := \left\{\mathcal{D}(\mathcal{H}_{in}) \to \mathcal{D}(\mathcal{H}_{out}) : h(\rho^{in}) = \mathrm{Tr}_{in,hid}(\mathcal{U}(\rho^{in} \otimes |0\cdots 0\rangle_{hid,out}\langle 0\cdots 0|)\mathcal{U}^\dagger)\right\}, \quad (24)$$

where $\mathcal{U} = U^{out}U^L \cdots U^1$ is the QNN quantum circuit, $U^l\mathcal{U}(d^{m_l+m_{l-1}})$ are the layer unitaries covering the action of a product of quantum perceptrons on the qubits in layers l and $l-1$ and $|0\cdots 0\rangle_{hid,out}\langle 0\cdots 0|$ represents the ground state of the layers.

In the case that the output state σ^{out} is a pure state we have

$$\begin{aligned}\mathcal{G} = \Big\{&\mathcal{D}(\mathcal{H}_{in}) \times \mathcal{D}(\mathcal{H}_{out}) \to \mathbb{R} : \\ &g(\rho^{in}, \sigma^{out}) = \mathrm{Tr}\left((\mathbb{1}^{in,hid} \otimes \sigma^{out})\mathcal{U}(\rho^{in} \otimes |0\cdots 0\rangle_{hid,out}\langle 0\cdots 0|)\mathcal{U}^\dagger\right)\Big\},\end{aligned}$$

where g is the fidelity eq.(13).

4.3.1 PAC Bounds

Our goal is to find PAC-bounds for the hypothesis class $\mathcal{J}$. Therefore, we will take a similar approach to the one we took before and use Theorem (2.8). Hence, we need

to take some preparatory steps. One may compare this with the previous section. At first, we need a complexity measure namely the covering number of this class. In order to find it we introduce the operator group, as before

$$\mathcal{J}_{circ} := \left\{\mathcal{D}(\mathcal{H}_{in}) \to \mathcal{D}(\mathcal{H}_{out}) : h_{circ}(\rho^{in}) = \mathcal{U}(\rho^{in} \otimes |0\cdots 0\rangle_{hid,out} \langle 0\cdots 0|)\mathcal{U}^\dagger\right\} \tag{25}$$

and determine its ϵ-covering number.

Lemma 4.8 (Covering Number of $\mathcal{J}_{circ}$)**.** *The ϵ-covering number for the operator group $\mathcal{J}$ obeys*

$$\mathcal{N}(\mathcal{J}_{circ}, \epsilon, \|\cdot\|) \leq \prod_l \left(1 + \frac{4(L+1)\pi}{\epsilon}\right)^{2(m_l+m_{l-1})^{2m_l}}.$$

Proof. The proof is quite similar to Lem.(4.2). Consider S_l as minimal ϵ-coverings of $\mathcal{U}(d^{m+m_{l+1}})$, such that for all u_l there is an $u_l^\epsilon \in S_l$ with $\|u_l - u_l^\epsilon\| \leq \epsilon$. So there are L ϵ-covering sets and we define $U_\epsilon = u_{L+1}^\epsilon \cdots u_1^\epsilon$.

For simplicity, we will identify $|0\rangle\langle 0|$ with $|0\cdots 0\rangle_{hid,out}\langle 0\cdots 0|$ for the next calculation. Now, we look at the discrepancy

$$\begin{aligned}
&\|U(\rho^{in} \otimes |0\rangle\langle 0|)U^\dagger - U_\epsilon(\rho^{in} \otimes |0\rangle\langle 0|)U_\epsilon^\dagger\| \\
&\quad = \|U(\rho^{in} \otimes |0\rangle\langle 0|)(U^\dagger - U_\epsilon)^\dagger) + (U - U_\epsilon)(\rho^{in} \otimes |0\rangle\langle 0|)U_\epsilon^\dagger\| \\
&\quad \leq \|U(\rho^{in} \otimes |0\rangle\langle 0|)(U^\dagger - U_\epsilon)^\dagger)\| + \|(U - U_\epsilon)(\rho^{in} \otimes |0\rangle\langle 0|)U_\epsilon^\dagger\| \\
&\quad \leq 2\|U - U_\epsilon\| \\
&\quad \leq 2(L+1)\epsilon
\end{aligned}$$

Thus we now S_L is a $2(L+1)\epsilon$-covering for $\mathcal{J}_{circ}$. Considering the number of trainable unitaries Prop.(4.1) gives

$$|S_L| \leq \prod_l \left(1 + \frac{2\pi}{\epsilon}\right)^{2(m_l+m_{l-1})^{2m_l}}.$$

This leads to

$$\mathcal{N}(\mathcal{J}_{circ}, 2(L+1)\epsilon, \|\cdot\|) \leq \prod_l \left(1 + \frac{2\pi}{\epsilon}\right)^{2(m_l+m_{l-1})^{2m_l}}$$

and finally

$$\mathcal{N}(\mathcal{J}_{circ}, \epsilon, \|\cdot\|) \leq \prod_l \left(1 + \frac{4(L+1)\pi}{\epsilon}\right)^{2(m_l+m_{l-1})^{2m_l}}.$$

□

In the following, we use this result to upper bound $\mathcal{G}$'s ϵ-covering number.

Lemma 4.9 (Covering Number for $\mathcal{G}$). *The ϵ-covering number for the hypothesis space $\mathcal{G}$ obeys*

$$\mathcal{N}(\mathcal{G}, \epsilon, |\cdot|) \leq \prod_l \left(1 + \frac{4\pi(L+1)m_0 \cdots m_L}{\epsilon}\right)^{2(m_L+m_{L-1})^{2m_l}}. \tag{26}$$

Proof. The idea is to use the previous Lemma(4.8) together with Lemma(2.6) to upper bound the covering number of $\mathcal{J}$. For simplicity we will identify $|0\rangle\langle 0|$ with $|0\cdots 0\rangle_{hid,out}\langle 0\cdots 0|$ for the next calculation.

$$\begin{aligned}
&d_2\Big((\mathbb{1}^{in,hid}\otimes\sigma^{out})\mathcal{U}(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}^\dagger, (\mathbb{1}^{in,hid}\times\sigma^{out})\mathcal{U}_\epsilon(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}_\epsilon^\dagger\Big)\\
&= \Big|\operatorname{Tr}[(\mathbb{1}^{in,hid}\otimes\sigma^{out})\mathcal{U}(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}^\dagger - (\mathbb{1}^{in,hid}\times\sigma^{out})\mathcal{U}_\epsilon(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}_\epsilon^\dagger]\Big|\\
&= \Big|\operatorname{Tr}[(\mathbb{1}^{in,hid}\otimes\sigma^{out})(\mathcal{U}(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}^\dagger - \mathcal{U}_\epsilon(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}_\epsilon^\dagger)]\Big|\\
&\leq \Big|\operatorname{Tr}(\mathbb{1}^{in,hid}\otimes\sigma^{out})\Big|\cdot\|\mathcal{U}(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}^\dagger - \mathcal{U}_\epsilon(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}_\epsilon^\dagger\|\\
&= \Big|\operatorname{Tr}(\mathbb{1}^{in,hid})\operatorname{Tr}(\sigma^{out})\Big|\cdot\|\mathcal{U}(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}^\dagger - \mathcal{U}_\epsilon(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}_\epsilon^\dagger\|\\
&= m_0\cdots m_L\cdot d_1(\mathcal{U}(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}^\dagger, \mathcal{U}_\epsilon(\rho^{in}\otimes|0\rangle\langle 0|)\mathcal{U}_\epsilon^\dagger)
\end{aligned}$$

Therefore the conditions for Lemma(2.6) are fulfilled, the constant is determined and the relation between the covering numbers is given by

$$\mathcal{N}(\mathcal{G}, \epsilon, |\cdot|) \leq \mathcal{N}(\mathcal{J}_{circ}, \frac{\epsilon}{m_0\cdots m_L}, \|\cdot\|) \leq \prod_l \left(1 + \frac{4\pi(L+1)m_0\cdots m_L}{\epsilon}\right)^{2(m_L+m_{L-1})^{2m_l}}.$$

□

Finally, all preconditions are fulfilled in order to apply Thm.(2.8) and gain a PAC bound for the dissipative quantum neural networks.

Theorem 4.10 (PAC Bound for Dissipative QNN's). *Consider $\mathcal{J}$ as before. Assume the output state σ^{out} is a pure state. Then for any $\epsilon > 0$ and any probability measure P on $\mathcal{X}\times\mathcal{Y}$ the following inequality yields*

$$\begin{aligned}
&\mathbb{P}_{S\sim P^n}\Big[\exists h\in\mathcal{J} : |R(h) - \hat{R}(h)| \geq \epsilon\Big]\\
&\qquad\leq 4\prod_l \left(1 + \frac{32\pi C(L+1)m_0\cdots m_L}{\epsilon}\right)^{2(m_l+m_{l-1})^{2m_l}} \cdot \exp\left(-\frac{n\epsilon^2}{32c^2}\right)
\end{aligned} \tag{27}$$

Proof. In order to proof this we will use Thm.(2.8). We define $\mathcal{G} := \{g : \mathcal{X}\times\mathcal{Y} \to [0,c] | \exists h\in\mathcal{J} : g(x,y) = L(h(x),y))\}$ and recognize that all conditions for using the Thm. mentioned above are fulfilled.
In addition we notice that for any $\epsilon > 0$ and any probability measure P on $\mathcal{X}\times\mathcal{Y}$ we

have

$$\begin{aligned}
\mathbb{P}_{S\sim P^n}[\exists h \in \mathcal{J} : |R(h) - \hat{R}(h)| \geq \epsilon] &\leq 4\Gamma_1(2n, \epsilon/8, \mathcal{G}) \cdot \exp\left(-\frac{n\epsilon^2}{32c^2}\right) \\
&\leq 4\Gamma_1(2n, \frac{\epsilon}{8C}, \mathcal{J}) \cdot \exp\left(-\frac{n\epsilon^2}{32c^2}\right) \\
&\leq 4\mathcal{N}(\mathcal{J}_{circ}, \frac{\epsilon}{8C}, \|\cdot\|) \cdot \exp\left(-\frac{n\epsilon^2}{32c^2}\right) \\
&\leq 4\prod_l \left(1 + \frac{32\pi C(L+1)m_0 \cdots m_L}{\epsilon}\right)^{2(m_l+m_{l-1})^{2m_l}} \\
&\quad \cdot \exp\left(-\frac{n\epsilon^2}{32c^2}\right).
\end{aligned}$$

The first step uses Thm.(2.8). The second step exploits Lem.(2.9) with C being the Lipschitz constant. The third step uses Lem(4.6). Finally, the last step uses Lemma(4.8). □

5 Conclusion

In this work, we aimed at transferring the concept of PAC bounds from classical learning theory to quantum learning theory.

We started with a recap of the basic concepts of the classical scenario in the case of inductive supervised learning. Hereby, important complexity measures were introduced and we looked at a PAC bound (2.8) which we used in later sections.

To create a transition to the quantum scenario, we had a brief summary of quantum mechanics and quantum computation. Here, the structure of VQA's was presented.

We looked at two quantum neural networks. First we focused on the QNN's described in "Efficient Measure for the Expressivity of Variational Quantum Algorithms" [3]. A PAC bound was already known, but we were able to formulate it more accurately. Furthermore, we were able to present another PAC bound for the same QNN. Future research could examine in which scenarios, which bound makes better predictions.

Further, we looked at "Training deep quantum neural networks" [2] where dissipative QNN's have been introduced. The biggest success of this work is that we found a PAC bound for this specific structure of QNN's, since no bound existed previously. When using these networks and therefore using empirical risk minimization it is extremely important to have knowledge on the distance between the risk and the empirical risk. This job can be done by the bound calculated in this work.

However, it remains an open task to find suitable bounds for more general scenarios.

A Appendix

A.1 Jensen's Inequality

Jensen's inequality relates the value of a convex function of an integral to the integral of the convex function. In the context of probability theory it takes the following form.

Lemma A.1 (Jensen's Inequality)**.** *Let X be a real-valued random variable. For any convex function $f : \mathbb{R} \to \mathbb{R}$,*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

yields if both expectation values exist. If f is concave the inequality is reversed.

A.2 Hoeffding' Inequality

The following inequality is a very helpful tool used in this work.

Lemma A.2 (Hoeffding's Inequality)**.** *Let $Z_1,...,Z_n$ be real independent random variables whose values are contained in intervals $[a_i, b_i]$. Then for every $\epsilon > 0$*

$$\mathbb{P}\left[\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq \epsilon\right] \leq \exp\left[-\frac{2\epsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right]$$
$$\mathbb{P}\left[\left|\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right| \geq \epsilon\right] \leq 2\exp\left[-\frac{2\epsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right].$$

The proof can be found in [5].

A.3 Covering the Unitary Group

We want to bound the covering number of the unitary group in n dimensions. Therefore, we will divide the proof into steps. We will follow the proof presented in [1].

Lemma A.3 (Covering Number of a Norm Ball)**.** *Let $A \subset \mathbb{R}$ be a domain in space and B the unit norm ball. Then the following relation holds.*

$$\left(\frac{1}{\epsilon}\right)^d \frac{vol(A)}{vol(B)} \overset{(a)}{\leq} \mathcal{N}(A, \|\cdot\|, \epsilon) \overset{Def}{\leq} \mathcal{M}(A, \|\cdot\|, \epsilon) \overset{(b)}{\leq} \frac{vol(A + \frac{\epsilon}{2}B)}{vol(\frac{\epsilon}{2}B)} \tag{28}$$

Proof. (a):

We consider a minimal covering, $A \subset \bigcup_{i=1}^{\mathcal{N}}(X_i, \epsilon)$. Therefore we have

$$vol(A) \leq vol(\bigcup_{i=1}^{\mathcal{N}(\epsilon)} B(X_i, \epsilon)) \leq \sum_{i=1}^{\mathcal{N}(\epsilon)} vol(B(X_i, \epsilon)) = \mathcal{N}(\epsilon) \cdot \epsilon^d vol(B)$$

The first inequality follows by definition of the covering, the second is an application of the union bound and the equality follows from $vol(B(X_i, \epsilon)) = \epsilon^d vol(B)$.
(b):
We consider an ϵ-packing, so the Balls $B(A_i, \frac{\epsilon}{2})$ are disjoint, and $\bigcup_{i=1}^{\mathcal{M}(\epsilon)} B(A_i, \frac{\epsilon}{2}) \subset A + \frac{\epsilon}{2}B$. Taking the volume gives

$$vol(A + \frac{\epsilon}{2}B) \geq vol(\bigcup_{i=1}^{\mathcal{M}(\epsilon)} B(A_i, \frac{\epsilon}{2})) = \mathcal{M}(\epsilon) vol(\frac{\epsilon}{2}B)$$

The first inequality follows from the definition of packing, and the equality uses that the sets are disjoint.

□

Lemma A.4. *In the special case of A being a norm ball with radius R we get*

$$\left(\frac{R}{\epsilon}\right)^d \leq \mathcal{N}(B_R, \|\cdot\|, \epsilon) \leq \left(1 + \frac{2R}{\epsilon}\right)^d. \tag{29}$$

Lemma A.5 (Writing Unitary via Exponential Map)**.** *The unitary group $U(n)$ can be expressed via*

$$U(n) = \exp[B_\pi(u(n))]. \tag{30}$$

Proof. It's a well known result that there is a correspondence between a Lie-Algebra and it's Lie Group and vice versa via the exponential map, which maps the Lie Algebra to its Lie Group.
We consider the relation

$$U(n) = \exp(u(n)),$$

where U(n) is unitary, so an element of the Lie Algebra, and u(n) is skew-hermitian, so an element of the Lie Group.

Unitary matrices are normal with eigenvalues $e^{i\lambda_k}$ having absolute value 1. We can write $U = V^\dagger diag(e^{i\lambda_1}, ..., e^{i\lambda_n})V$ with unitary V. Furthermore $U = \exp(J)$ with $J = V^\dagger diag(i\lambda_1, ..., i\lambda_n)V \in u(n)$.
One can choose $|\lambda_k| \leq \pi$ such that $\|J\| \leq \pi$.

□

Lemma A.6 (Bounding Distance between two Skew-hermitian Matrices)**.** *Let X, Y*∈

u(n) be two skew-hermitian matrices. Then we have

$$(2-e^r)\|X-Y\| \overset{(a)}{\leq} \|e^X - e^Y\| \overset{(b)}{\leq} \|X-Y\|, \tag{31}$$

where (a) is valid for $\|X\|, \|Y\| \leq r$.

Proof. (a):
Applying the definition of the exponential map and using the reverse triangle inequality it follows

$$\|e^X - e^Y\| \geq \|X-y\| - \|\sum_{k\geq 2} \frac{1}{k!}(X^k - Y^k)\|$$

We further consider the second term

$$\begin{aligned}
\|\sum_{k\geq 2} \frac{1}{k!}(X^k - Y^k)\| &= \|\sum_{k\geq 2}\sum_{l=1}^{k} \frac{1}{k!} X^{l-1}(X-Y)Y^{k-l}\| \\
&\leq \sum_{k\geq 2} \frac{1}{(k-1)!} r^{k-1}\|X-Y\| \\
&= (e^r - 1)\|X-Y\|.
\end{aligned}$$

The first equality follows from writing $X^k - Y^k$ as a telescope sum. The first inequality uses the triangle inequality and the assumption $\|X\|, \|Y\| \leq r$. The last equality follows by an index shift and definition of the exponential map. Thus we proofed (a).
(b):
For the upper bound we write $e^X - e^Y$ as a telescope sum

$$\begin{aligned}
\|e^X - e^Y\| &= \|\lim_{m\to\infty} \sum_{k=1}^{m} e^{(k-1)\frac{X}{m}}(e^{\frac{X}{m}} - e^{\frac{Y}{m}})e^{(m-k)\frac{Y}{m}}\| \\
&\leq \lim_{m\to\infty} \sum_{k=1}^{m} \|e^{(k-1)\frac{X}{m}}(e^{\frac{X}{m}} - e^{\frac{Y}{m}})e^{(m-k)\frac{Y}{m}}\| \\
&= \lim_{m\to\infty} \sum_{k=1}^{m} \|e^{\frac{X}{m}} - e^{\frac{Y}{m}}\| \\
&= \lim_{m\to\infty} m\|e^{\frac{X}{m}} - e^{\frac{Y}{m}}\| \\
&= \|X-Y\|.
\end{aligned}$$

The first inequality follows from the triangle inequality, the first equality follows because the operator norm is unitary invariant and the last equality follows from the definition of the exponential map. □

Now we are ready to bound the covering number for the unitary group $\mathcal{U}(n)$.

Lemma A.7. *For $\epsilon > 0$, the ϵ-covering number for the unitary group $\mathcal{U}(n)$ obeys*

$$\left(\frac{3}{4\epsilon}\right)^{n^{2k}} \leq \mathcal{N}(\mathcal{U}(n^k), \epsilon, \|\cdot\|) \leq \left(1 + \frac{2\pi}{\epsilon}\right)^{n^{2k}} \tag{32}$$

where $\|\cdot\|$ denotes the operator norm.

Proof. The idea of the proof is straightforward. Eqs.(28, 30, 31) secure the preconditions to apply Lem.(2.6).

Therefore one chooses $H_1 = \mathcal{B}_\pi(u(n))$, $H_2 = \mathcal{U}(n)$, d_1 and d_2 as the operator norm and $f(X) = \exp(X)$.

Applying Theorem (2.6) to Eq.(31), sets the constant $K = 1$. If one chooses $r = 2/5$ the lower bound of eq.(31) gives $2 - e^r > \frac{1}{2} =: k$.
Since $u(n)$ is isomorphic to an n^2-dimensional real vector space, one specifies $D = n^2$ and $R = \pi$, when applying 28 for $\mathcal{B}_\pi(u(n))$. Theorem 2.6 ensures

$$\mathcal{N}(\mathcal{B}_\pi(u(n)), \|\cdot\|, 4\epsilon) \leq \mathcal{N}(U(n), \|\cdot\|, \epsilon) \leq \mathcal{N}(\mathcal{B}_\pi(u(n)), \|\cdot\|, \epsilon) \tag{33}$$

Equation (28) yields

$$\left(\frac{\pi}{4\epsilon}\right)^{n^2} \leq \mathcal{N}(\mathcal{U}(n), \|\cdot\|, \epsilon) \leq \left(1 + \frac{2\pi}{\epsilon}\right)^{n^2} \tag{34}$$

□

Bibliography

[1] Thomas Barthel and Jianfeng Lu. "Fundamental Limitations for Measurements in Quantum Many-Body Systems". In: *Physical Review Letters* 121.8 (Aug. 2018). DOI: 10.1103/physrevlett.121.080406. URL: https://doi.org/10.1103%2Fphysrevlett.121.080406.

[2] Kerstin Beer et al. "Training deep quantum neural networks". In: *Nature Communications* 11.1 (Feb. 2020). DOI: 10.1038/s41467-020-14454-2. URL: https://doi.org/10.1038%2Fs41467-020-14454-2.

[3] Yuxuan Du et al. "Efficient Measure for the Expressivity of Variational Quantum Algorithms". In: *Physical Review Letters* 128.8 (Feb. 2022). DOI: 10.1103/physrevlett.128.080506. URL: https://doi.org/10.1103%2Fphysrevlett.128.080506.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[5] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30. ISSN: 01621459. URL: http://www.jstor.org/stable/2282952 (visited on 10/26/2022).

[6] Richard Jozsa. "Fidelity for Mixed Quantum States". In: *Journal of Modern Optics* 41.12 (Dec. 1994), pp. 2315–2323. DOI: 10.1080/09500349414552171.

[7] Alex Monràs, Gael Sentıs, and Peter Wittek. "Inductive Supervised Quantum Learning". In: *Phys. Rev. Lett.* 118 (19 May 2017), p. 190503. DOI: 10.1103/PhysRevLett.118.190503. URL: https://link.aps.org/doi/10.1103/PhysRevLett.118.190503.

[8] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010. DOI: 10.1017/CBO9780511976667.

[9] Werner, R. *Mathematical methods of quantum information theory, Lecture 1*. 2017. URL: https://www.youtube.com/watch?v=vb0ZEsATUcw&list=PLDfPUNusx1EoBAn8vXYjcF95R7mI_eR6o&index=1.

[10] Werner, R. *Mathematical methods of quantum information theory, Lecture 2*. 2017. URL: https://www.youtube.com/watch?v=MD-QKVkgaqY&list=PLDfPUNusx1EoBAn8vXYjcF95R7mI_eR6o&index=2.

[11] Werner, R. *Mathematical methods of quantum information theory, Lecture 3*. 2017. URL: https://www.youtube.com/watch?v=m26XXxqNSWM&list=PLDfPUNusx1EoBAn8vXYjcF95R7mI_eR6o&index=3.

[12] Werner, R. *Mathematical methods of quantum information theory, Lecture 4*. 2017. URL: https://www.youtube.com/watch?v=R5wcAxYW36M&list=PLDfPUNusx1EoBAn8vXYjcF95R7mI_eR6o&index=4.

[13] Michael M. Wolf. *Mathematical Foundations of Supervised Learning*. 2021. URL: https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MA4801_2021S/ML.pdf.

[14] Michael M. Wolf. *Quantum Channels & Operations*. 2012. URL: http://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MichaelWolf/QChannelLecture.pdf.

Declaration of Authorship

I hereby declare that the work was completed independently, no sources or means other than those indicated were used, all passages of the work that make reference to other sources, whether through direct quotation or paraphrasing, have been indicated accordingly, and the work has not previously been submitted to an examining authority in the same or a similar form.

31.01.23, Hannover

date, place

signature