Chaos and Complexity from Quantum Neural Network

A study with Diffusion Metric in Machine Learning

Sayantan Choudhury[1,2,3,‡, §],
Ankan Dutta[4] and Debisree Ray[5]

[1] National Institute of Science Education and Research, Bhubaneswar, Odisha - 752050, India
[2] Homi Bhabha National Institute, Training School Complex, Anushakti Nagar, Mumbai - 400085, India
[3] Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, 14476 Potsdam-Golm, Germany.
[4] Department of Mechanical Engineering, Jadavpur University, Kolkata - 700032, India
[5] Department of Physics and Astronomy, Mississippi State University, 355 Lee Boulevard Mississippi State, MS 39762, United States

Abstract

In this work, our prime objective is to study the phenomena of quantum chaos and complexity in the machine learning dynamics of *Quantum Neural Network* (QNN). A *Parameterized Quantum Circuits* (PQCs) in the hybrid quantum-classical framework is introduced as a universal function approximator to perform optimization with *Stochastic Gradient Descent* (SGD). We employ a statistical and differential geometric approach to study the learning theory of QNN. The evolution of parametrized unitary operators is correlated with the trajectory of parameters in the Diffusion metric. We establish the parametrized version of Quantum Complexity and Quantum Chaos in terms of physically relevant quantities, which are not only essential in determining the stability, but also essential in providing a very significant lower bound to the generalization capability of QNN. We explicitly prove that when the system executes limit cycles or oscillations in the phase space, the generalization capability of QNN is maximized. Moreover, a lower bound on the optimization rate is determined using the well known *Maldacena Shenker Stanford* (MSS) bound on the Quantum Lyapunov exponent.

Keywords: Neural Network, Quantum Complexity & Chaos, Machine Learning.

[‡] *Corresponding author, E-mail* : `sayanphysicsisi@gmail.com`

[§] *NOTE: This project is the part of the non-profit virtual international research consortium "Quantum Aspects of Space-Time & Matter" (QASTM) .*

Contents

1 Introduction

The advent of machine learning research started with the proposal of the basic framework of the neural network, perceptron by Rosenblatt in 1958. Soon in 1975, Werbos developed the back-propagation, a learning algorithm that can train multi-level perceptron. But it took almost four decades to implement a 'deep' neural network on a large industrial scale. Since then, classical machine learning has been an indispensable tool in various fields ranging from healthcare to robotics. The most crucial factor that directed the mammoth success of classical machine learning is the advancement of computational power. Back in the 1980s, Feynman proposed utilizing the quantum mechanical power of nature to compute, which led to the idea of quantum computers. But to avoid the decoherence effect in qubits, low temperature is required to sustain the coherence property of qubits. Due to these engineering issues, it took nearly three decades to build a practical quantum computer[1–5]. Recently, Google has been able to simulate quantum chemistry reactions using only 12 hydrogen atoms representing 12 qubits of information in these quantum computers [1, 6]. Quantum computer researchers lately have shown interest in quantum machine learning, performing machine learning tasks using the computational power of the quantum mechanical world. There has been a lot of focus recently on the proposal of a quantum neural network as a quantum analog of classical neural network with back-propagation as its learning algorithm [7–10].

Similar to classical machine learning, quantum machine learning can be categorized into supervised, unsupervised, and semi-supervised learning like reinforcement learning. The paper focuses on the supervised learning of quantum neural networks. In a classical supervised learning algorithm, the neural network is provided with two sets of data, a training set, and a testing set. The dataset comprises ordered pairs of input and desired output, and the neural network samples these ordered pairs to the training set under a fixed probability distribution. The neural network learns from the labeled data of the training set and predicts the unlabelled data of the testing set with given accuracy and a confidence parameter. We define a loss function f, which represents the training error between the predicted output by the neural network and actual output in the labeled data. The neural network optimizes the loss function during the training period. In the quantum supervised learning scenario, there can be three different possible data-algorithm pairs in terms of quantum and classical nature of data and learning algorithms. In this paper, we use a hybrid quantum-classical neural network where we train quantum data with classical learning algorithms like stochastic gradient descent. In the quantum-classical hybrid framework proposed by Mitarai, a *Parametrized Quantum Circuits* (PQCs) is introduced by [11, 12]. The quantum circuit is characterized by parameters that are optimized using a classical learning algorithm to optimize the loss function and simultaneously reducing the testing error. A gradient-based learning algorithm is used for back-propagation. In this paper, we have used *Stochastic Gradient Descent* (SGD) as our learning algorithm [13–15].

SGD performs gradient descent in batches rather than on the complete training set. Remarkable progress has been made in executing quantum neural networks, but there has been no significant progress in developing the quantum analogy of the statistical learning theory of neural networks. The learning theory of classical neural networks has been an important tool for computer scientists to decipher the black-box of information processing in neural networks [16–21]. Learning theory can be analyzed from two different perspectives: statistical approach and information-theoretic approach. For classical neural networks, the work by [16, 17] analyzed the learning dynamics using the evolution of mutual information between the output and the layers of the neural network. From a quantum analogy of the information-theoretical approach, the works by [22, 23] correlated the dynamical behavior of the tripartite information with the loss function in the training process. In this paper, we analyze from a statistical and differential geometric perspective of evaluating chaos and complexity in QNN. One of the most important aspects of learning theory is the determination of stability in the learning process of the QNN. In classical neural networks, the work [20] showed the learning trajectory of deep linear networks is exponentially stable. On the other hand, there can be neural network systems with limit cycles in the phase portrait [21]. Moreover, it is argued by [24] that the coherent systems with oscillations or limit cycles are essential in the stability of continuous memory in the human brain. The stability of the learning trajectory in classical neural networks motivated us to have a look from the QNN perspective. On the other hand, over the years, scientists searched for chaos in various fields ranging from population models to black holes. The search for chaotic patterns in neuroscience has been analyzed since the inception of the Hodgkin-Huxley model [25]. The work is further extended with experiments performed by the work [26]. The chaos in neural networks has been well studied over the past years [27–29]. The quantum processing in neurons proposed by [30] signifies that its equivalence with the human brain, which justifies the analysis of the learning theory of quantum neural network. Yet, there has been no significant progress in the chaos in quantum neural networks. Recently, [22] analyzed chaos from the notion of information scrambling in quantum neural networks. In this paper, we approach to analyze the chaos and complexity in a quantum neural network from a statistical perspective. The analysis will form a basis to further co-relate the chaos in a quantum neural network with other known physical models like black holes or economic models. This would give a deep insight into how different the human brain motivated quantum neural networks behaves from the rest of the physical models.

A recent research direction in the classical machine learning context has been the possibility of searching for optimal neural architecture [31–34]. The ability to search for optimal architectures given the training dataset gives us the power to predict an optimal neural network theoretically prior to the long training periods. The optimality is defined on two competing terms. Firstly, the time constant of the decay of the training error, better neural network shows a larger time constant. On the other hand, one also needs to take care of the generalization error. When the neural network attains more number of minima in the learning manifold, the probability of reaching a wide minima increases, which would result in minimum generalization error. But for a larger time constant, the number of minima attended decreases, thus a trade-off resulting in higher generalization error. To attain this optimal ability for the neural network to reach a large number of minima with an optimal time constant, the paper attempts to map the learning trajectory or unitary evolution of QNN to the trajectory of parameters using a Riemannian manifold called diffusion metric. The diffusion metric is introduced by Foressi in [35] and it is constructed by perturbing the flat Minkowski space by the magnitude of the noise in the gradient of the loss function. We can observe how the manifold changes in changing the neural architecture. In doing so, we will be able to correlate the optimal unitary evolution of QNN with optimal i.e. stationary action path of particles in diffusion metric. The correlation has two implications, firstly searching for an optimal QNN architecture as mentioned before, another coming from a more theoretical high energy physics perspective. The optimal unitary evolution denotes the minimum number of computations required to generate the final unitary $\mathcal{U}_f$ from the initial unitary $\mathcal{U}_i$ which in other words, the relative complexity [36–38] between the initial and final unitary, $\mathcal{C}(\mathcal{U}_f,\mathcal{U}_i)$. This measure of complexity in unitary space is directly correlated with the optimal trajectory evolution of parameters or geodesic in the parameter space. The correlation establishes complexity as a function of the parameters of the QNN. After establishing the complexity, an extensive study of quantum chaos has been studied [36, 37, 39–44]. The stability of the neural network in terms of Lyapunov stability and its evolution establishes how the neural architecture governs the stability of the neural network. Rather than an extensive study of the Lyapunov evolution, an extremal study in terms of the growth of the complexity has been carried out. The Maldacena, Shenker, Stanford (MSS) bound [39, 45] on the Quantum Lyapunov exponent, given by $\lambda \leq 2\pi/\beta$ with β inverse equilibrium temperature, puts forward an interesting limit on the optimization rate. The QNN cannot optimize its parameters completely, as there will always be a minimum deviation from the optimal parameters, which hold equality for maximal complexity systems. In this connection, the out-of-time-correlator (OTOC) has also been calculated using the universality relation between complexity and OTOC, given by $\mathcal{C} = -\log(\text{OTOC})$ [¶].

The paper considers a hybrid quantum-classical neural network framework based on PQCs [11, 12] optimizing quantum data with classical gradient-based algorithms like stochastic gradient descent (SGD). Throughout the paper, we have assumed that the length of the training dataset is large enough for the loss function to stabilize i.e. the loss function doesn't change as we increase the length of the training set. This led us to

[¶] The concept of out-of-time-correlator (OTOC) is treated as a very important probe to quantify the amplitude of quantum chaos in terms of Quantum Lyapunov exponent. In this paper we have explicitly computed the expression for the OTOC from the first principle, rather using the universality relation between complexity and OTOC we have determined OTOC in terms of complexity. So this implies once the expression for the complexity can be computed from the present set-up the connection with quantum chaos can be very easily established using the mentioned universality relationship.

avoid fluctuations due to sampling. The assumption of a large training dataset and the stabilizing of the loss function is inspired by Bialek *et al* [46] corresponding to quantum computation at thermal equilibrium [47]. The paper established the behavior of noise in SGD, which is governed by the neural architecture and dataset of QNN using the diffusion metric. The parameterized complexity of QNN is established by corresponding with the geodesic of parameter trajectories in the diffusion metric. The paper further analyses the stability of QNN using the Lyapunov exponent as a function of the neural architecture and dataset.

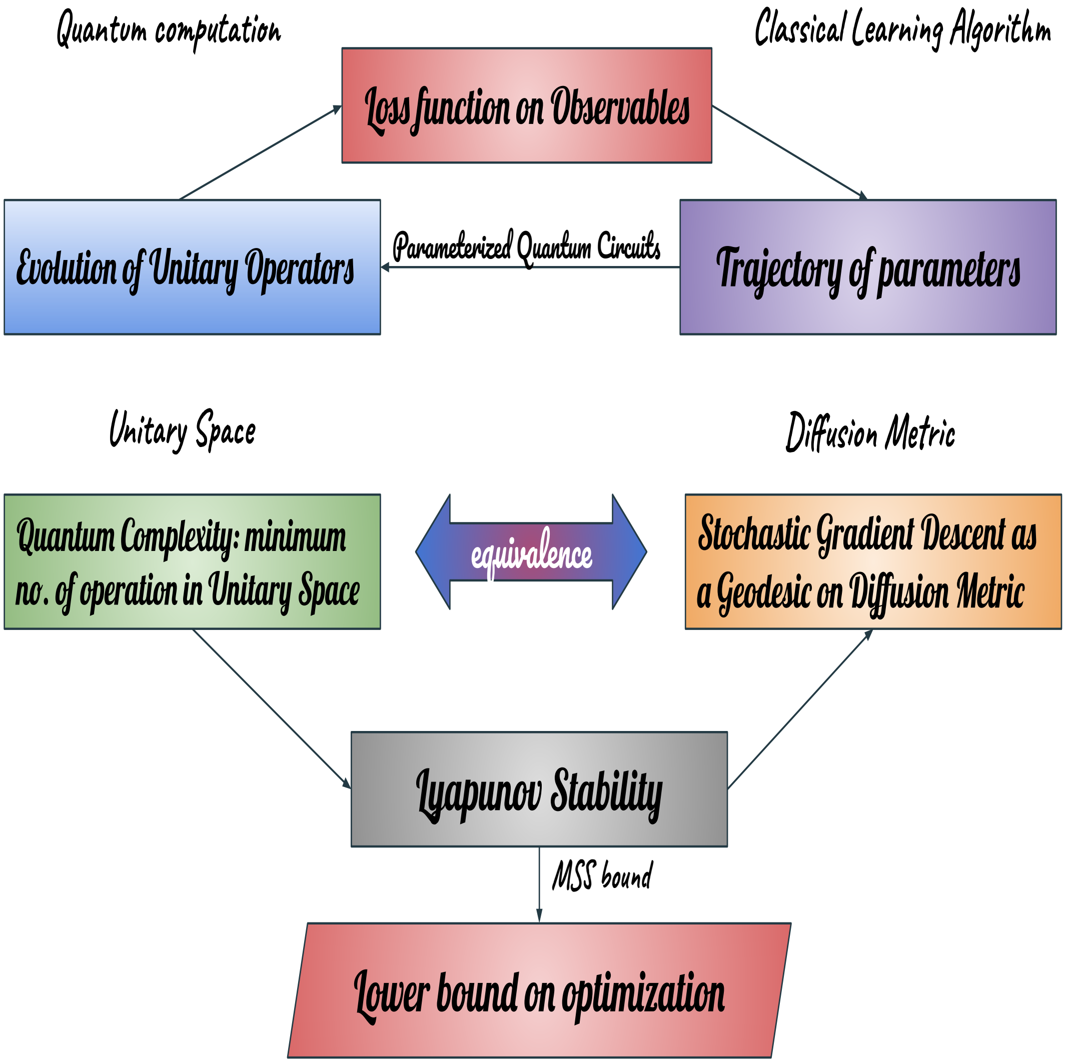
Quantum computation
Classical Learning Algorithm
Loss function on Observables
Evolution of Unitary Operators
Parameterized Quantum Circuits
Trajectory of parameters
Unitary Space
Diffusion Metric
Quantum Complexity: minimum no. of operation in Unitary Space
equivalence
Stochastic Gradient Descent as a Geodesic on Diffusion Metric
Lyapunov Stability
MSS bound
Lower bound on optimization

Figure 1.1: Roadmap of the paper

The paper is divided into three sections, building up the mathematical background in Section 2 to the analysis of stability using the Lyapunov exponent in Section 4. In Section 2, Parameterized Quantum Circuits were introduced as a universal function approximator [48] and analysis were performed as a quantum analog of the statistical learning theory based on [46]. The diffusion metric [35] is introduced in Section 3, correlating the learning trajectory of QNN to the evolution of noise in SGD during training. After establishing the fundamental and mathematical concepts in Mathematical Background and Diffusion Metric, the paper determines complexity as a function of parameters in Parameterized Complexity. The complexity of QNN determines the Lyapunov exponents and thus the stability analysis was established in Quantum Lyapunov exponents.

2 Mathematical Background

We focus on executing supervised learning tasks using *Parameterized Quantum Circuit* framework. In classical supervised learning, the model learns to map from an input dataset $\{x_i\}$ to the output $\{\hat{y}_i\}$. The map represents a $w-$parameterized function $y_i = m(x_i; w)$, which is optimized to be close to the output $\hat{y}_i$ for all data index i belonging to the training dataset. The metric used to define the closeness of the parameterized function and output is called the loss function. The loss function can represent any metric like cross-entropy, likelihood loss, log loss, etc [13]. Here, we consider the mean square error loss. The main objective of the model is to optimize the loss function using a certain learning algorithm. These algorithms updates the parameters w of the parameterized function $m(x_i; w)$ to optimize the loss function. Here, the model optimizes the loss function employing stochastic gradient descent (SGD) as its learning algorithm. SGD is an iterative method for optimizing the loss function by using the gradient of the loss function calculated from a randomly selected subset of the training dataset [13, 14]. This supervised learning scenario is valid for both classical and quantum neural network. In the quantum neural network context, the initial density matrix $\rho_{in}(x_i)$ is created by encoding the input data stream $\{x_i\}$ onto the *Encoder Circuit* Unitary $\mathcal{U}_\phi(x_i)$, which acts on the ground state $|0\rangle$ [11, 12]. The Unitary $\mathcal{U}_\phi$ can be represented as a sum of a linear combination of the basis operators α spanning over $K-$dimensional space with basis functions $\phi(x)$ as coefficients. The *Encoder Circuit* Unitary $\mathcal{U}_\phi$ is characterized by basis functions $\phi_\mu(x)$, where x is sampled independently and identically from the training dataset under a fixed probability distribution $P(x)$ with variance σ_η^2 [15, 46]. Mathematically the unitary $\mathcal{U}_\phi$ can be represented by:

$$\boxed{\textbf{\textit{Encoder Circuit Unitary :}}\ \mathcal{U}_\phi(x) = \sum_{\mu=1}^{K} \phi_\mu(x)\alpha^\mu}, \tag{2.1}$$

using which the input density matrix ρ_{in} is defined as:

$$\boxed{\boxed{\textbf{\textit{Input Density Matrix :}}\ \rho_{\text{in}}(x_i) = \mathcal{U}_\phi(x_i)\left|0\right\rangle\left\langle 0\right|\mathcal{U}_\phi^\dagger(x_i)}}. \tag{2.2}$$

Here the input density matrix ρ_{in} is created from the input dataset using the equations 2.1-2.2. The quantum neural network (QNN) applies a parameterized Unitary operator $\mathcal{U}_\theta$ on the input density matrix to produce an output density matrix ρ_{out} at every epoch (or iteration). Here, it is important to note that similar to the universal approximation theorem in artificial neural networks [48], there always exists a quantum circuit that can represent a target function within an arbitrarily small error. The parameterized quantum circuit learning will always be able to optimize to any arbitrary small error but the depth or complexity of the circuit increases. This optimization also doesn't guarantee the generalization capability of the quantum neural network which would result in a high testing error. Motivated by the notion of deep neural networks and the unitary arrangement proposed by Beer *et. al.* [7], we use the quantum neural network architecture of stacked unitary operators with a L number of layers. The unitary operator of the whole quantum circuit parameterized by θ is given by:

$$\boxed{\boxed{\mathcal{U}_\theta = \prod_{i=1}^{L} \mathcal{U}_i}}, \tag{2.3}$$

where $\mathcal{U}_i$ is the unitary at ith layer parameterized by weights w. The unitary at ith layer can be expressed as the sum of the linear combination of the basis operators σ with w as its coefficients. Mathematically, the unitary $\mathcal{U}_i$ can be expressed as follows:

$$\boxed{\boxed{\mathcal{U}_i = \sum_{\nu} w_\nu^i \sigma^\nu}}. \tag{2.4}$$

Combining equations 2.3-2.4, the parameterized unitary operator can be expressed as the sum of linear combination of the basis operators σ spanning over $P-$dimensional space with parameter θ as its coefficient. Mathematically, the unitary operator $\mathcal{U}_\theta$ can be expressed as follows:

$$\boxed{\boxed{\textbf{\textit{Parameterized Unitary :}}\ \mathcal{U}_\theta = \sum_{\nu=1}^{P} \theta_\nu \sigma^\nu}}, \tag{2.5}$$

where $\theta_\nu = g^\nu(w)$ is a function of weights. The parameters θ gets updated by the learning algorithm to optimize the loss function. The unitary operator $\mathcal{U}_\theta$ given by equation 2.5 acts on the initial density matrix given by equation 2.2 to produce the output density matrix ρ_{out}. We measure the output density matrix ρ_{out} using the observer operator B to get an expected value of the observation B given by $\mathrm{Tr}(B\rho_{out})$. The aim of QNN is to optimize this expected observation value to a target observation value $\bar{B}$ as an output. For every input dataset $\{x_i\}$, we consider a corresponding output observation dataset $\{B_i\}$. During training period, the training dataset $(x_i, \bar{B}_i)$ is sampled under the distribution $P(x)$. The parameterized unitary operator $\mathcal{U}_\theta$ maps the input $\{x_i\}$ encoded initial density matrix to output density matrix. The observer maps this output density matrix to the loss function value f. Now using a learning algorithm (here, stochastic gradient descent) the QNN must find a sub-space of the unitary operator $\mathcal{U}_{\bar{\theta}}$ for which the loss function f is at its minimum. But again this sub-space doesn't guarantee generalization over dataset or minimum of testing error.[∥] The framework for Quantum Neural Network can be summarized following equations 2.1-2.2 and 2.3-2.5 as:

$$\boxed{\boxed{\textbf{\textit{Output Density Matrix :}}\ \rho_{\text{out}}(x_i) = \mathcal{U}_\theta^\dagger \rho_{\text{in}}(x_i)\mathcal{U}_\theta}}, \tag{2.6}$$

and

$$\boxed{\boxed{\textbf{\textit{Loss function :}}\ f = \frac{1}{N}\sum_{i=1}^{N}\left(\bar{B}_i - \mathrm{Tr}(B\rho_{\text{out}}(x_i))\right)^2}}, \tag{2.7}$$

where N is the length of the training dataset. We introduce a mapping Ω which the equations 2.6-2.7 maps $\Omega : \theta \mapsto f$ for every epoch. It is important to note that there is a reduction in dimension to form a surjective $K - \text{to} - 1$ mapping. This results in many sub-spaces of optimal parameters for which the loss function f is minimum. Under a learning algorithm (here, stochastic gradient descent) QNN changes the unitary operator every epoch t by applying a mapping $\Theta : \theta_t \mapsto \theta_{t+1}$. To optimize the loss function, the learning algorithm tends to map Θ such that $\mathrm{E}(f(\theta_t)) > \mathrm{E}((\theta_{t+1}))$. SGD tends the corresponding loss function to decreases iteratively at every epoch and finally reach the unitary $\mathcal{U}_{\bar{\theta}}$, characterized by $\bar{\theta}$. In other words, when the parameters optimize $\theta \to \bar{\theta}$, the expected value of observation B will tend towards $\bar{B}$, resulting in the zero training error. The parameters $\bar{\theta}$ are the optimal parameters. So, the matrix $\bar{B}$ can be represented as:

$$\boxed{\boxed{\bar{B}_i = \left(\mathrm{Tr}(B\mathcal{U}_{\bar{\theta}}^\dagger \rho_{\text{in}}^i \mathcal{U}_{\bar{\theta}}) + \eta\right)}}, \tag{2.8}$$

[∥] There may be an overlapping sub-sub space for which the QNN could both optimize and generalize. Note that there can also be a situation when the sub-space of optimization doesn't overlap with the sub-space of generalization. In that case, a trade-off takes place which is not favorable and a different neural architecture or *Encoder Circuit* should be looked after.

where we used a shorthand $\rho^i_{\text{in}} = \rho_{\text{in}}(x_i)$ and η is the Gaussian noise with mean zero and σ^2_Γ variance, similar the classical variant used in [46]. The neural network optimizes by updating its parameters using the stochastic gradient descent (SGD). The learning algorithm or the mapping $\Theta : \theta_t \mapsto \theta_{t+1}$ can be given as follows:

$$\boxed{\textbf{\textit{Stochastic Gradient Descent :}}\ \theta_{t+1} = \left(\theta_t - \frac{\Gamma}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{\partial f_i}{\partial\theta},\right)}, \tag{2.9}$$

where Γ represents the learning rate. Rather than optimizing the whole training dataset, we optimize the dataset in batches $\mathcal{B}$, randomly picked from the whole training data. This not only reduces the computational cost but also adds stochasticity due to random sampling of batches which proves to be very essential in the generalization context, which will be discussed later in Complexity & Stability. The batch size $|\mathcal{B}|$ is much less than the total length of the training dataset N. The stochasticity in the SGD arises when $|\mathcal{B}| << N$, which allows for higher stochasticity in random sampling while training. When both the length $|\mathcal{B}| \sim N$ then the stochasticity loses and SGD becomes simple (not stochastic!) gradient-descent. SGD in the last few decades has been experimentally the most efficient in terms of accuracy and computational cost. This lead to its increasing interests and documentations in the computer-science community. Recently, with the works of [14, 15, 35], there is a huge surge in interest in analyzing the SGD from a dynamical system perspective. We discuss this important aspect of SGD in Diffusion Metric.

The loss function plays an essential part in the QNN framework as it is the objective function and purposely drives the gradient in the learning algorithm. Combining equations 2.1-2.2 and equations 2.5-2.7, the loss function can be expressed as:

$$\boxed{f = \sigma^2_\eta + \sum_{\mu,\nu,\delta,\gamma}^{P}(\theta^*_\mu - \bar{\theta}^*_\mu)(\theta_\nu - \bar{\theta}_\nu)(\theta^*_\delta - \bar{\theta}^*_\delta)(\theta_\gamma - \bar{\theta}_\gamma)\mathrm{Tr}\underbrace{\left(\frac{1}{N}\sum_{i=1}^{N}(B\sigma^\mu\rho^i_{\text{in}}\sigma^\nu \otimes B\sigma^\delta\rho^i_{\text{in}}\sigma^\gamma)\right)}_{\equiv\Delta}}. \tag{2.10}$$

Further simplifying the factor Δ under the condition that $N \to \infty$ (large N) as evaluated in [15, 46] we get:

$$\boxed{\Delta = \sum_{j,k,p,q}^{K} A^{\infty}_{j'kp'q}\left(B\sigma^\mu\alpha^j\alpha^k\sigma^\nu \otimes B\sigma^\delta\alpha^p\alpha^q\sigma^\gamma\right)} \tag{2.11}$$

where the expansion co-efficient $A^{\infty}_{j'kp'q}$ can be expressed as:

$$\boxed{\boxed{\textbf{\textit{Encoder-Dataset Tensor :}}\ A^{\infty}_{j'kp'q} = \lim_{N\to\infty}\left\{\frac{1}{N}\sum_{i=1}^{N}\phi^*_j(x_i)\phi^*_p(x_i)\phi_q(x_i)\phi_k(x_i)\right\}}}. \tag{2.12}$$

The 4−rank tensor A^{∞} exists under the assumption that the equation 2.12 thermalizes or reaches equilibrium. The basis functions of the *Encoder Circuit* unitary operator $\mathcal{U}_{\phi}$ plays a pivotal part in creating the initial density matrix ρ_{in} in equation 2.1-2.2. The basis functions are constant for a framework and thus selected prior to any training. The selection of these basis functions is thus important as the input dataset gets encoded into the basis functions. The tensor A^{∞} signifies the relation between the *Encoder Circuit* and dataset. Hence, we call the tensor A^{∞} as *Encoder-Dataset* tensor. Here, we assumed that the loss function stabilizes when the number of training data is large enough to neglect the fluctuations. Continuing equation 2.10 and replacing Δ using equation 2.12, the loss function takes the following simplified form:

$$\boxed{\boxed{f = \sigma^2_{\eta} + \sum^{P}_{\mu,\nu,\delta,\gamma}\sum^{K}_{j,k,p,q} A^{\infty}_{j'kp'q}(\theta^*_{\mu} - \bar{\theta}^*_{\mu})(\theta_{\nu} - \bar{\theta}_{\nu})(\theta^*_{\delta} - \bar{\theta}^*_{\delta})(\theta_{\gamma} - \bar{\theta}_{\gamma})\mathrm{Tr}\Bigg(B\sigma^{\mu}\alpha^{j}\alpha^{k}\sigma^{\nu}\otimes B\sigma^{\delta}\alpha^{p}\alpha^{q}\sigma^{\gamma}\Bigg)}}. \tag{2.13}$$

The above equation shows how the loss function is governed by the selection of *Encoder-Dataset* tensor A^{∞}, given the observation matrix B. An important observation from equation 2.13 is that: when the parameters optimize i.e. $\theta \to \bar{\theta}$, the loss function minimizes to a non-zero constant σ^2_{η}, the variance of sampling distribution $P(x)$. It is easy to mathematically validate as the loss function f is a mean squared error loss, the sampling the ordered pairs $(x_i, \bar{B}_i)$ also attributed to the loss function. Consider the sampling distribution $P(x)$ as a delta function with $\sigma^2_{\eta} \to 0$, then the loss function $\min f$ also tends to zero. Now, when we increase the variance of the distribution, or in other words, increase the diversification of the training dataset, the $\min f$ also increases. This also shows that the neural network will fail for a uniform sampling distribution with $\sigma^2_{\eta} \to \infty$, there has to be an underlying structure of the dataset for the neural network to optimize.

3 Diffusion Metric

Previously in 2, we discussed about the notion of stochasticity in stochastic gradient descent (SGD) using equation 2.9. The hyper-parameters of SGD i.e. the batch size $|\mathcal{B}|$ and the learning rate Γ are crucial in achieving optimal learning trajectory in the context of computational cost and accuracy [14]. The learning trajectory accounts for the trajectory of parameters in the learning manifold, from the initial condition governed by the dynamical equation given by equation 2.9. A faster learning rate will skip minimal in the learning manifold, thus reducing the probability of achieving better minimal for optimization and generalization. Though a smaller batch size will reduce the computational cost, it also on the other hand increase the stochastic behavior of SGD large enough to skip the minimal, increasing training and testing error. This brings the focus to find a metric to calculate the stochasticity in the SGD, to analyze the effect of the hyperparameters, the *Encoder Circuit* along with the neural architecture of QNN on the behavior of stochasticity in the learning trajectory. In the literature by Foressi *et. al.* [35] showed that the Diffusion matrix D which is essentially the covariance matrix of the gradient of the loss function, provides a great insight into the stochastic nature of SGD. The diffusion matrix D becomes a null matrix when the learning trajectory is governed by a simple (not stochastic!) gradient-descent. This implies that when the matrix D is null, the sampling of the batches is irrelevant to the learning algorithm. At this time, the loss function has reached its critical point or in other words, the model has learned the training dataset. The magnitude of the Diffusion matrix determines the amount of stochasticity of SGD. The work [35] introduced a metric called Diffusion metric $\widetilde{D}$, which is created by perturbing the Minkowski space of parameters with the magnitude of noise in the stochastic gradient descent. Foressi *et. al.* [35] showed that the trajectory of parameters θ governed by SGD follows a geodesic path in the diffusion metric under a potential given by V. Mathematically, the diffusion metric is given by the following expression:

$$\boxed{\textit{\textbf{Diffusion Metric :}}\ \ \widetilde{D}_{\mu\nu} = \Big(\delta_{\mu\nu} + \epsilon D_{\mu\nu}\Big)}, \tag{3.1}$$

where $\epsilon < 1/\max\lambda_D$ where λ_D is the set of eigenvalues of diffusion matrix D and ϵ is the order of perturbation to the Minkowski space in the above expression. The Minkowski space corresponds to the diffusion matrix being $D = 0$ or in other words, the learning trajectory is governed by simple gradient-descent. Perturbing this Minkowski space with weak perturbation will distort the straight line geodesic path of parameters governed by simple gradient descent. This path corresponds to the path of the parameters with no excitation to explore other minima, thus increasing the probability of finding a better minima point resulting in better generalization. As mentioned earlier, the diffusion matrix is a covariance matrix of the gradient of the loss function f, which is mathematically expressed by the following expression:

$$\boxed{\textit{\textbf{Diffusion Matrix :}}\ \ D_{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial f_i}{\partial\theta_\mu}\right)\left(\frac{\partial f_i}{\partial\theta_\nu}\right) - \frac{1}{N^2}\sum_{i,j=1}^{N}\left(\frac{\partial f_i}{\partial\theta_\mu}\right)\left(\frac{\partial f_j}{\partial\theta_\nu}\right)} \tag{3.2}$$

To evaluate the diffusion matrix, we consider evaluating the gradient of the loss function f as follows:

$$\boxed{f_i = \left\{ \sum_{\mu,\nu}^{P} (\theta^*_\mu - \bar{\theta}^*_\mu)(\theta_\nu - \bar{\theta}_\nu)\mathrm{Tr}\Big(B\sigma^\mu \rho^i_{\mathrm{in}} \sigma^\nu\Big) + \eta \right\}^2}, \tag{3.3}$$

using which we compute:

$$\left(\frac{\partial f_i}{\partial \theta_\zeta}\right) = \Big\{ 2 \sum_{\mu,\nu,\delta,\gamma}^{P} (\theta^*_\mu - \bar{\theta}^*_\mu)(\theta_\nu - \bar{\theta}_\nu)\Big(\bar{\mathcal{G}}^\delta_\zeta(\theta_\gamma - \bar{\theta}_\gamma) + \mathcal{G}^\gamma_\zeta(\theta^*_\delta - \bar{\theta}^*_\delta)\Big)\mathrm{Tr}\Big(B\sigma^\mu \rho^i_{\mathrm{in}} \sigma^\nu \otimes B\sigma^\delta \rho^i_{\mathrm{in}} \sigma^\gamma\Big)$$
$$+2\eta \sum_{\delta,\gamma}^{P} (\bar{\mathcal{G}}^\delta_\zeta(\theta_\gamma - \bar{\theta}_\gamma) + \mathcal{G}^\gamma_\zeta(\theta^*_\delta - \bar{\theta}^*_\delta))\mathrm{Tr}\Big(B\sigma^\delta \rho^i_{\mathrm{in}} \sigma^\gamma\Big)\Big\}, \tag{3.4}$$

$$\left(\frac{\partial f}{\partial \theta_\zeta}\right) = \Big\{ \sum_{\mu,\nu,\delta,\gamma}^{P} \sum_{j,k,p,q}^{K} A^\infty_{j'kp'q}(\theta^*_\mu - \bar{\theta}^*_\mu)(\theta_\nu - \bar{\theta}_\nu)\Big(\bar{\mathcal{G}}^\delta_\zeta(\theta_\gamma - \bar{\theta}_\gamma) + \mathcal{G}^\gamma_\zeta(\theta^*_\delta - \bar{\theta}^*_\delta)\Big)$$
$$\mathrm{Tr}\Big(B\sigma^\mu \alpha^j \alpha^k \sigma^\nu \otimes B\sigma^\delta \alpha^p \alpha^q \sigma^\gamma\Big)$$
$$+\eta \sum_{\delta,\gamma}^{P} \sum_{j,k}^{K} A^\infty_{j'k}\Big(\bar{\mathcal{G}}^\delta_\zeta(\theta_\gamma - \bar{\theta}_\gamma) + \mathcal{G}^\gamma_\zeta(\theta^*_\delta - \bar{\theta}^*_\delta)\Big)\mathrm{Tr}\Big(B\sigma^\delta \alpha^j \alpha^k \sigma^\gamma\Big)\Big\}, \tag{3.5}$$

where the dependency of θ^μ with respect to θ^ν can be given by the Jacobian $\mathcal{G}^\mu_\nu$ as follows:

$$\boxed{\textbf{\textit{Jacobian Matrix :}}\ \mathcal{G}^\mu_\nu = \left(\frac{\partial \alpha^\mu}{\partial \alpha^\nu}\right) = \left\{ \begin{array}{ll} \sum\limits_{l \in \{L(\nu)\}} \left(\dfrac{\partial g^\mu(w)}{\partial w_l}\right)\left(\dfrac{\partial w_l}{\partial g^\nu(w)}\right) & \mu \neq \nu \\ 1 & \mu = \nu \end{array} \right\}} \tag{3.6}$$

where $\{L(\nu)\}$ is the collection of indexes l for which $\frac{\partial g^\nu(w)}{\partial w^l} \neq 0$ as shown in [15]. It is important to note that $\mathcal{G}$ changes with epochs as weights evolve with time. The matrix $\mathcal{G}$ measures the dependence between the different parameters represented as coordinates. The matrix $\mathcal{G}$ being Dirac-delta function infers that the parameters are independent and the parameter space corresponds to the Minkowski space. From the dynamical system perspective, the matrix $\mathcal{G}$ governs the dependence of a parameter with other parameters, which in turn changes that parameter itself. One can correlate this scenario with many-body interactions with long-range hopping where the hopping energy from lattice site i to j corresponds to the magnitude of the matrix $\mathcal{G}^j_i$. When the magnitude of each element of the matrix $\mathcal{G}$ is large enough, the hopping energy is large, making the disorder strength to decrease- ergodicity arises. On the other hand, when the magnitude of each element of the matrix $\mathcal{G}$ is small enough, the hopping energy is small making the disorder strength to increase- localization arises and ergodicity is lost. In the work by [21], the inverse temperature β is defined by the hyper-parameters of SGD. This motivated us to correlate the Jacobian matrix $\mathcal{G}$ with the hopping energy. The correlation provides a holistic phase diagram between temperature T and disorder strength W similar to any Ising-like models as shown in [49, 50]. The phase diagram will provide a deeper understanding between the equilibrium systems or non-equilibrium systems in an artificial neural network context. The study of equilibrium and non-equilibrium aspects in artificial neural networks has been discussed in [21].

Similar to the the assumption in the equation 2.12 that the 4−rank tensor A^{∞} thermalizes or reaches equilibrium, the diffusion matrix given in 3.2 also thermalizes. Using the similar treatment as [15, 46], the approximated diffusion matrix can be given by

Approximated Diffusion Matrix :

$$\begin{aligned} D^{\infty}_{\zeta\eta} = \lim_{N\to\infty} D_{\zeta\eta} = 4\Bigg[\sum_{\mu,\nu,\delta,\gamma,\omega,\kappa,\xi,\pi}^{P} \sum_{j,k,p,q,r,s,a,b}^{K} \Big(A^{\infty}_{j'kp'qr'sa'b} - A^{\infty}_{j'kp'q}A^{\infty}_{r'sa'b}\Big)(\theta^*_\mu - \bar{\theta}^*_\mu)(\theta_\nu - \bar{\theta}_\nu) \\ \Big(\bar{\mathcal{G}}^{\delta}_{\zeta}(\theta_\gamma - \bar{\theta}_\gamma) + \mathcal{G}^{\gamma}_{\zeta}(\theta^*_\delta - \bar{\theta}^*_\delta)\Big)(\theta^*_\omega - \bar{\theta}^*_\omega)(\theta_\kappa - \bar{\theta}_\kappa)\Big(\bar{\mathcal{G}}^{\pi}_{\eta}(\theta_\xi - \bar{\theta}_\xi) + \mathcal{G}^{\xi}_{\eta}(\theta^*_\pi - \bar{\theta}^*_\pi)\Big) \\ \mathrm{Tr}\Big(B\sigma^{\mu}\alpha^{j}\alpha^{k}\sigma^{\nu} \otimes B\sigma^{\delta}\alpha^{p}\alpha^{q}\sigma^{\gamma} \otimes B\sigma^{\omega}\alpha^{r}\alpha^{s}\sigma^{\kappa} \otimes B\sigma^{\pi}\alpha^{a}\alpha^{b}\sigma^{\xi}\Big) \\ +\sigma^2_\eta \sum_{\delta,\gamma,\pi,\xi}^{P} \sum_{p,q,a,b}^{K} A^{\infty}_{p'qa'b}\Big(\bar{\mathcal{G}}^{\delta}_{\zeta}(\theta_\gamma - \bar{\theta}_\gamma) + \mathcal{G}^{\gamma}_{\zeta}(\theta^*_\delta - \bar{\theta}^*_\delta)\Big)\Big(\bar{\mathcal{G}}^{\pi}_{\eta}(\theta_\xi - \bar{\theta}_\xi) + \mathcal{G}^{\xi}_{\eta}(\theta^*_\pi - \bar{\theta}^*_\pi)\Big) \\ \mathrm{Tr}\Big(B\sigma^{\delta}\alpha^{p}\alpha^{q}\sigma^{\gamma} \otimes B\sigma^{\pi}\alpha^{a}\alpha^{b}\sigma^{\xi}\Big)\Bigg], \end{aligned} \tag{3.7}$$

where the matrix index $\bar{\mathcal{G}}^{\delta}_{\zeta}$ denotes the dependency of the complex conjugate of θ_δ on the parameter θ_ζ. Mathematically, the matrix index is given by $\bar{\mathcal{G}}^{\delta}_{\zeta} = \left(\frac{\partial\theta^*_\delta}{\partial\theta_\zeta}\right)$. The 8−rank *Encoder-Dataset* matrix $A^{\infty}_{j'kp'qr'sa'b}$ can be expressed as:

$$\boxed{A^{\infty}_{j'kp'qr'sa'b} = \lim_{N\to\infty}\left\{\frac{1}{N}\sum_{i=1}^{N}\phi^*_j(x_i)\phi^*_p(x_i)\phi^*_r(x_i)\phi^*_a(x_i)\phi_q(x_i)\phi_k(x_i)\phi_b(x_i)\phi_s(x_i)\right\}}. \tag{3.8}$$

The stochasticity of SGD changes with time which provides a temporal variation of the magnitude of perturbation in the Minkowski space. This perturbation in the approximated Diffusion metric can be correlated with the movement of masses in a Riemannian manifold where the parameters form the space-time coordinates. We now shift the problem from the approximated Diffusion metric with parameters to the trajectory of particles in the Riemannian manifold with the presence of small random masses. The magnitude of these masses is given by the magnitude of the noise in SGD, thus changes with time. The mass distribution on the Riemannian manifold also changes. Now, imagine you are told to control the trajectory of a particle from an initial point to its final point, by changing the mass distribution. The reward or aim of the particle is to visit more number of intermediate points while also reaching the target within a considerable time. The number of intermediate points corresponds to the generalization capability of the neural network and time here is the training time the QNN takes to reach the optimal points. In zero-mass distribution configuration, the parameters' trajectory would've been a straight line, reaching in less training time but also with less generalization capability. Changing the mass distribution increases the probability of the particle to more number of intermediate points, thus increasing the generalization capability. Analyzing the temporal distribution of mass thus becomes important in controlling the particle trajectory in maximizing its rewards. Realizing the correspondence, it thus becomes important to understand the stochasticity flow of SGD to control the parameter trajectory. The equation 3.7 shows the dependence of neural architecture and *Encoder-Dataset* tensor on the stochasticity. Thus the neural architecture and *Encoder-Dataset* tensor plays an important role in controlling the trajectory of parameters in the Diffusion metric, in the context of generalization capability and convergence rate. A detailed study is performed in Complexity & Stability.

Approximated diffusion matrix D^∞ being a covariance matrix, is a positive semi-definite matrix with all positive eigenvalues. This property of the matrix D^∞ restricts the domain of parameters θ. The QNN in their whole learning trajectory should always have a parameter set θ for which $\lambda_D(\theta) \geq 0$. To visualize the limiting condition on the parameters, let us consider the parameters are equally optimized in all the directions $\theta_\mu - \bar{\theta}_\mu = \Delta\theta$ for all index $\mu \leq P$. The jacobian matrix $\mathcal{G}$ and the observation matrix B are considered as an identity matrix. The dimensions $P, K = 4$ where all the basis operators are Pauli operators including the identity matrix. Using equation 3.7, the restrictions on the difference of parameters $\Delta\theta$ can be evaluated as:

$$\begin{aligned} D^\infty_{\zeta\eta} = \; & 8\mathrm{Re}^2(\Delta\theta)\Bigg[\sum_{\mu,\nu,\omega,\kappa}^{4}\sum_{j,k,p,q,r,s,a,b}^{4}\underbrace{\Big(A^\infty_{j'kp'qr'sa'b} - A^\infty_{j'kp'q}A^\infty_{r'sa'b}\Big)}_{\equiv\Psi}\Big|\Delta\theta\Big|^2 \\ & \times\underbrace{\mathrm{Tr}\Big(\sigma^\mu\sigma^j\sigma^k\sigma^\nu \otimes \sigma^\zeta\sigma^p\sigma^q\sigma^\zeta \otimes \sigma^\omega\sigma^r\alpha^s\sigma^\kappa \otimes \sigma^\eta\sigma^a\sigma^b\sigma^\eta\Big)}_{\equiv\Phi_2} \\ & +\sigma^2_\eta\sum_{p,q,a,b}^{4} A^\infty_{p'qa'b}\underbrace{\mathrm{Tr}\Big(\sigma^\zeta\sigma^p\sigma^q\sigma^\zeta \otimes \sigma^\eta\sigma^a\sigma^b\sigma^\eta\Big)}_{\equiv\Phi_1}\Bigg]. \end{aligned} \quad (3.9)$$

Further simplifying the factor Ψ using the definitions of *Encoder-Dataset* tensor in equation 2.12 and 3.8, we get

$$\boxed{\boxed{\Psi = \mathrm{cov}\Big(\phi_r^*(x)\phi_a^*(x)\phi_s(x)\phi_b(x), \phi_j^*(x)\phi_p^*(x)\phi_q(x)\phi_k(x)\Big)}} \tag{3.10}$$

where $\mathrm{cov}(a, b)$ is the covariance between two vectors $\vec{a}$ and $\vec{b}$. Using the properties of Pauli matrices, the quantity Φ_1 can be simplified to

$$\Phi_1 = \ 4\delta_{pq}\delta_{ab} \tag{3.11}$$

It is important to observe than Φ_1 is independent of the index (ζ, η). Similarly, the quantity Φ_2 is also independent of the index (ζ, η). Thus the approximated diffusion matrix in 3.9 is a constant matrix with all elements as a constant number c, where the quantity $c = D^{\infty}_{\eta\zeta}$ for all indexes (η, ζ) as shown in 3.9. The eigenvalues of matrix D^{∞} are $\lambda_D = \{0, 0, 0, 4c\}$, which has to be a positive quantity:

$$\begin{aligned} 0 \leq \ & \Bigg[\sum_{\mu,\nu,\omega,\kappa}^{4} \sum_{j,k,p,q,r,s,a,b}^{4} \mathrm{cov}\Big(\phi_r^*(x)\phi_a^*(x)\phi_s(x)\phi_b(x), \phi_j^*(x)\phi_p^*(x)\phi_q(x)\phi_k(x)\Big)\Big|\Delta\theta\Big|^2 \\ & \mathrm{Tr}\Big(\sigma^{\mu}\sigma^{j}\sigma^{k}\sigma^{\nu} \otimes \sigma^{\zeta}\sigma^{p}\sigma^{q}\sigma^{\zeta} \otimes \sigma^{\omega}\sigma^{r}\alpha^{s}\sigma^{\kappa} \otimes \sigma^{\eta}\sigma^{a}\sigma^{b}\sigma^{\eta}\Big) + 4\sigma_\eta^2 \sum_{p,q,a,b}^{4} A^{\infty}_{p'qa'b}\delta_{pq}\delta_{ab}\Bigg]. \end{aligned} \tag{3.12}$$

The above equation may be not always be true as it completely depends on the selected *Encoder-Dataset* tensor A^{∞}. Notice that when $4-$rank *Encoder-Dataset* tensor A^{∞} has all negative values at $A^{\infty}_{p'pa'a}$ for all $a, p \leq 4$, then the inequality 3.12 *cannot* reach $|\Delta\theta| = 0$. If the parameters optimizes completely i.e. $|\Delta\theta| = 0$ then the inequality doesn't hold true, thus a contradiction. In these cases, a stricter inequality can be evaluated according to the elemental value of the matrix A^{∞} and thus a limit to optimization can be evaluated. One can again correlate the inequality 3.12 with particles in the Riemannian manifold where the $K = 4$ parameters are the 4 space-time coordinates. In the space-time context, the inequality 3.12 shows that in certain metric (or *Encoder-Dataset* tensor), the space-time coordinates can be restricted in reaching to its final coordinate i.e. $\bar{\theta}$, making the difference $|\Delta\theta|$ a non-zero quantity. It is certainly no surprising that these types of space-time restrictions are quite common to see, reflecting an interesting correlation with QNN.

4 Complexity & Stability

At every epoch of training of QNN, a particular unitary operator $\mathcal{U}_\theta$ is prepared by the *Parameterized Quantum Circuit*. Brown and Susskind [37, 44, 51, 52] viewed the preparation of $\mathcal{U}_\theta$ as a time-series of discrete motions of an auxiliary particle on the Special Unitary SU group space. The particle starts at the identity operator I and ends at a target unitary operator $\mathcal{U}$. The complexity of the unitary operator $\mathcal{U}_\theta$ is the number of minimum operators required to create $\mathcal{U}_\theta$ by the given circuit. Mathematically, it is given by the geodesic on the SU group space. QNN employs a unitary operator $\mathcal{U}_\theta$ at every epoch, thus the complexity of the QNN changes with epochs. In the QNN context, the final or target unitary operator is given by $U_{\bar{\theta}}$ to produce minimum training error. The particle in group space travels from initial unitary operator $\mathcal{U}_{\theta,}$ to the unitary operator $\mathcal{U}_{\bar{\theta}}$. On the other hand, this can be corresponded with a particle in Diffusion metric traveling from initial parameter configuration θ_0 to the optimal parameter set $\bar{\theta}$ as discussed in Diffusion Metric. A consequence of this correspondence is that the geodesic path traveled by the particle in the Diffusion metric can be correlated with the complexity as the geodesic traveled in the group space. Reflecting from the parameterized version of the approximated diffusion metric in equation 3.7, the complexity as a function of parameters can be correlated. Based on the parameterized complexity, one can further study the quantum chaos and complexity in QNN.

4.1 Parameterized Complexity

The main objective of this section is to establish the complexity [36, 37, 52–58] as a function of parameters. This is motivated by the parameterized version of the approximated diffusion metric presented in equation 3.7. The parameterized complexity is evaluated by corresponding the diffusion metric introduced in [35]. The stochastic gradient descent follows a geodesic path on this diffusion metric, which has been discussed in [35] is given by the following equation:

$$\boxed{\textit{Geodesic on Diffusion Metric :}\ \left(\frac{\partial\theta}{\partial t}\right) = -(I - \epsilon D^{\infty})\left(\frac{\partial f}{\partial\theta}\right)} \tag{4.1}$$

where D^∞ is the approximated diffusion matrix and measures the degree of stochasticity. When the matrix D^∞ becomes a null matrix, then the equation 4.1 represents the simple gradient descent as a learning algorithm. The optimal parameter $\bar{\theta}$ by integrating equation 4.1 can be shown as:

$$\boxed{\bar{\theta}_\nu = \theta_0 - \int_0^T \sum_{\mu=1}^{4} (\delta_{\mu\nu} - \epsilon D^{\infty}_{\mu\nu}) \left(\frac{\partial f}{\partial\theta_\mu}\right) dt} \tag{4.2}$$

where T is a hypothetical total training time to reach the optimal parameters from the initial parameter set θ_0. The equation 2.5 in Mathematical Background correlates the parameters set θ with the unitary $\mathcal{U}_\theta$. The trajectory of a particle in group space from initial unitary operator $\mathcal{U}_{\theta,}$ exactly correspondence to the trajectory of parameters in Diffusion metric from the initial parameter set θ_0 to $\bar{\theta}$ due to the linearity in 2.5. Using this correspondence, the evolution of unitaries in the unitary space as follows:

$$\boxed{\textbf{\textit{Evolution of Unitary :}} \quad \left(\frac{\partial \mathcal{U}_\theta}{\partial t}\right) = -\sum_{\nu=1}^{4}\sum_{\mu=1}^{4}(\delta_{\mu\nu} - \epsilon D^{\infty}_{\mu\nu})\left(\frac{\partial f}{\partial \theta_\mu}\right)\sigma^{\nu}} \tag{4.3}$$

Initiating with an initial unitary operator $\mathcal{U}_{\theta,}$, the unitary $\mathcal{U}_\theta$ evolves with epochs tending towards the target unitary operator $\mathcal{U}_{\bar{\theta}}$. Susskind [37] discretized the special unitary group space in ϵ_0-balls, where the auxiliary particle takes discretized steps into these balls to corresponds with the evolution of Unitaries. We assume a parameter set to belong to the optimal parameter set $\theta \in \bar{\theta}$, when the unitaries corresponding to the parameter fall in the ϵ_0 ball or, in other words, $|\mathcal{U}_\theta - \mathcal{U}_{\bar{\theta}}| < \epsilon_0$. The work [35] showed that SGD follows a geodesic path in diffusion metric at every epoch, which also corresponds to the complexity path on the group space. Based on the *Complexity-Action conjecture* in [36, 37, 44, 51, 54–56, 58], we correspond the complexity of the unitaries in the group space with the action on the diffusion metric. The *Complexity-Action conjecture* as shown in [36, 37, 44, 51] is given by:

$$\boxed{\textbf{\textit{Complexity-Action Conjecture :}} \ \mathcal{C} = \frac{\mathcal{A}}{\pi}}. \tag{4.4}$$

The above equation shows that a change of complexity in group space will reflect a change in action in the diffusion metric and vice versa. The definition of Action defined on the diffusion metric [35] is given by

$$\boxed{\textbf{\textit{Action :}} \ \mathcal{A} = \int \left[\sqrt{\sum_{\mu,\nu} \widetilde{D}^{\infty}_{\mu\nu}\dot{\theta}^{\mu}\dot{\theta}^{\nu}} - V\right] dt \qquad \text{where, } V = -\int_{\theta}^{\bar{\theta}} \sum_{\mu}\left[\frac{\partial}{\partial t}\left(\frac{\partial f}{\partial \theta}\right)\right] d\theta} \tag{4.5}$$

From the *Parameterized Quantum Circuit* perspective, a change in the complexity will ensure a change in the unitary operator. This change in the unitary operator will cause a change in the parameter configuration in the diffusion metric, thus changing the action in the metric. Using 4.5 equation 4.3, the unitary as a function of complexity can be written as:

$$\boxed{\left(\frac{\partial \mathcal{U}}{\partial \mathcal{C}}\right) = \frac{1}{\pi\left[\sqrt{\sum_{\mu,\nu} \widetilde{D}^{\infty}_{\mu\nu}\dot{\theta}^{\mu}\dot{\theta}^{\nu}} - V\right]} \sum_{\nu,\mu=1}^{4} \left(\delta_{\mu\nu} - \epsilon D^{\infty}_{\mu\nu}\right)\left(\frac{\partial f}{\partial \theta_{\mu}}\right)\sigma^{\nu}}. \tag{4.6}$$

where V is the potential under which the parameters evolve. On the other hand, using equation 2.5 the gradient of unitary with respect to change in complexity is given by:

$$\boxed{\left(\frac{\partial \mathcal{U}}{\partial \mathcal{C}}\right) = \sum_{\mu}^{4}\sum_{\nu}^{4} \frac{\mathcal{G}^{\nu}_{\mu}\sigma^{\nu}}{\mathcal{C}^{\mu}_{\theta}}} \tag{4.7}$$

where we have introduced the following quantity:

$$\mathcal{C}^{\mu}_{\theta} = \left(\frac{\partial \mathcal{C}}{\partial \theta_{\mu}}\right) \tag{4.8}$$

Therefore, equating the equations 4.6 and 4.7, one can conclude:

$$\boxed{\left(\frac{\partial \mathcal{C}}{\partial \theta_{\mu}}\right) = \frac{\pi\left[\sqrt{\sum_{\zeta,\eta} \widetilde{D}^{\infty}_{\zeta\eta}\dot{\theta}^{\zeta}\dot{\theta}^{\eta}} - V\right]}{\sum_{\nu}\left[\mathcal{G}^{-1}(I - \epsilon D^{\infty})\right]_{\mu\nu}\left(\frac{\partial f}{\partial \theta_{\nu}}\right)}} \tag{4.9}$$

The above equation establishes the distribution of complexity as a function of parameter and $\left[\mathcal{G}^{-1}(I - \epsilon D^{\infty})\right]_{\mu\nu}$ represents the elemental value of the matrix $\mathcal{G}^{-1}(I - \epsilon D^{\infty})$. The complexity is thus given by the following expression:

$$\boxed{\textbf{\textit{Parameterized Complexity :}}\ \mathcal{C}(\theta) = \pi\sum_{\mu}\int_{\theta_0}^{\theta}\left[\frac{\sqrt{\sum_{\zeta,\eta} \widetilde{D}^{\infty}_{\zeta\eta}\dot{\theta}^{\zeta}\dot{\theta}^{\eta}} - V}{\sum_{\nu}\left[\mathcal{G}^{-1}(I - \epsilon D^{\infty})\right]_{\mu\nu}\left(\frac{\partial f}{\partial \theta_{\nu}}\right)}\right] d\theta_{\mu}} \tag{4.10}$$

The above expression of Complexity $\mathcal{C}$ is difficult to evaluate exactly, analytically. We evaluate the complexity at specific epochs of the learning trajectory, as established in the next sub-section 4.2.

4.2 Stability analysis using Quantum Lyapunov exponents

Using [42, 43, 45, 59–61], one can write down the following relation between OTOC and complexity:

$$\boxed{\boxed{\mathcal{C} = -\log(\text{OTOC}) \qquad \text{and} \quad \text{OTOC} = \exp\left(-\exp(\lambda\theta)\right)}}, \tag{4.11}$$

which further implies the following combined universal relation which particularly holds good in the context of quantum description of chaotic phenomena:

$$\boxed{\boxed{\textit{Complexity :}\ \mathcal{C} = -\log(\text{OTOC}) = \exp(\lambda\theta)}}, \tag{4.12}$$

where λ is identified to be the Quantum Lyapunov exponent for quantum chaos and satisfy the well known Maldacena Shenker Stanford (MSS) bound [39], which is given by:

$$\lambda \leq \frac{2\pi}{\beta}, \tag{4.13}$$

where the temperature late time equilibrium saturation temperature for the maximal chaos is given by the present context is defined as:

$$\boxed{\boxed{\textit{Equilibrium Temperature :}\ \beta^{-1} = T = \frac{\Gamma}{2|\mathcal{B}|}}}, \tag{4.14}$$

where, Γ being the learning rate and $|\mathcal{B}|$ is the batch size [21]. The equality sign or the upper bound corresponds to the maximal chaos where late time saturating behaviour can be observed.

Here one can compute the Lyapunov exponent in terms of the Complexity as follows:

$$\boxed{\boxed{\textit{Quantum Lyapunov Exponent :}\ \lambda = \left(\frac{\partial \log(\mathcal{C})}{\partial\theta}\right) \leq \frac{2\pi}{\beta} = \frac{\pi\Gamma}{|\mathcal{B}|}}}, \tag{4.15}$$

which is basically can be measured from the slope of the $\log(\mathcal{C})$ vs θ plot, which we have plotted in the later half of this paper. This further implies the following bound on the complexity:

$$\boxed{\boxed{\textit{Complexity Bound :}\ \mathcal{C} \leq \exp\left(\frac{2\pi}{\beta}\theta\right) = \exp\left(\frac{\pi\Gamma}{|\mathcal{B}|}\theta\right)}}. \tag{4.16}$$

In equation 4.10, the complexity expression is difficult to evaluate. We analyze complexity in certain critical learning epochs when the system is in a steady state i.e. the velocity of parameters $\dot{\theta} = 0$. We introduce a steady-state parameter set θ_{ss} during which the system is stationary. Thus, in the framework of QNN the Lyapunov exponent is evaluated from the complexity for these epochs by the following simplified expression:

$$\boxed{\vec{\lambda} = \frac{1}{\displaystyle\int_{\theta_0}^{\theta_{ss}} \sum_{\mu} \frac{d\theta_\mu}{\displaystyle\sum_{\nu} \Big[\mathcal{G}^{-1}(I - \epsilon D^{\infty})\Big]_{\mu\nu} \left(\frac{\partial f}{\partial \theta_\nu}\right)}} \times \left[\frac{1}{\displaystyle\sum_{\nu} \Big[\mathcal{G}^{-1}(I - \epsilon D^{\infty})\Big]_{\mu\nu} \left(\frac{\partial f}{\partial \theta_\nu}\right)\bigg|_{\theta=\theta_{ss}}}\right]_{\vec{\mu}}}, \tag{4.17}$$

where $[.]_{\vec{\mu}}$ represents vector running through the parameter μ. Using equation 4.17, it is evident that $\text{sign}\left(\displaystyle\sum_{\nu} \Big[\mathcal{G}^{-1}(I - \epsilon D^{\infty})\Big]_{\mu\nu} \left(\frac{\partial f}{\partial \theta_\nu}\right)\bigg|_{\theta=\theta_{ss}}\right)$ determines the nature of the system. It is important to note that there can be a case of inflection in terms of the system stability when the Lyapunov exponent changes its sign. The Lyapunov exponent takes an intermediate form when the matrix $\mathcal{G}^{-1}(I - \epsilon D^{\infty}) \left(\dfrac{\partial f}{\partial \theta}\right)\bigg|_{\theta=\theta_{ss}}$ becomes singular. For the further computational simplification purpose we introduce a function, $p(\theta)$ which is defined as:

$$\boxed{p(\theta) :\equiv \mathcal{G}^{-1}(I - \epsilon D^{\infty}) \left(\frac{\partial f}{\partial \theta}\right)} \tag{4.18}$$

We analyse the equation 4.17 by tending the quantity p at its extremal values i.e. initially we analyse when $p \to 0$ and then we analyse when $p \to \infty$. Using this mentioned identification we get the following simplified expression for the Lyapunov exponent:

$$\boxed{\lambda = \left(\frac{1}{p(\theta_{ss}) \displaystyle\int_{\theta_0}^{\theta_{ss}} \frac{d\theta}{p(\theta)}}\right) = \left(\frac{1}{\theta_{ss} - \theta_0 + \displaystyle\int_{\theta_0}^{\theta_{ss}} \underbrace{\left(p'(\theta) \int \frac{d\theta}{p(\theta)}\right)}_{\equiv \mathcal{K}(\theta)} d\theta}\right)}, \tag{4.19}$$

where we introduce a new function $\mathcal{K}(\theta)$, which is given by the following expression:

$$\boxed{\mathcal{K}(\theta) = p'(\theta) \int \frac{d\theta}{p(\theta)}}. \tag{4.20}$$

During this simplification we have used the integration by parts in the above mentioned second step.

This paper considers two extremal situations i.e. when the quantity $p(\theta) \to 0$ and $p(\theta) \to \infty$. Initially, a conditioned analysis on the stability of the system is performed here using $p(\theta) \to 0$:

$$\boxed{\boxed{\lambda = \lim_{p\to 0}\left(\frac{1}{p(\theta_{ss})\displaystyle\int_{\theta_0}^{\theta_{ss}}\frac{d\theta}{p(\theta)}}\right) = \lim_{p\to 0}\left(\frac{1}{\theta_{ss}-\theta_0+\displaystyle\int_{\theta_0}^{\theta_{ss}}\underbrace{\left(p'(\theta)\int\frac{d\theta}{p(\theta)}\right)d\theta}_{\equiv\mathcal{K}(\theta)}}\right)}}, \tag{4.21}$$

Using the equation, the following observations can be made when $p \to 0$:

1. If $p'(\theta)$ is a negative finite quantity at the critical parameter set θ^* when $p(\theta^*) = 0$, then Lyapunov exponent tends towards 0^-, thus the system stabilises with oscillations or limit cyclic behaviour in the phase space.

2. While if $p'(\theta)$ is a positive finite quantity at the critical parameter set θ^* when $p(\theta^*) = 0$, then Lyapunov exponent tends towards 0^+, where the choatic nature of the system arises with unstable limit cycles.

3. If $p'(\theta) = 0$, then $\mathcal{K}(\theta)$ can be further simplified to be following form:

$$\boxed{\boxed{\mathcal{K}(\theta) = \log(p(\theta)) + \int\left(p''(\theta)\int\frac{d\theta}{p(\theta)}\right)d\theta}}. \tag{4.22}$$

   When $p''(\theta) = 0$, the value of $\mathcal{K}(\theta) \to -\infty$ thus the Lyapunov exponent $\lambda \to 0^-$, stabilising the system with limit cycles.

So, the condition with the quantity $p(\theta) \to 0$, shows the system inherently can execute stable or unstable limit cycles in the phase space.

Now, let us consider another limiting condition with $p(\theta) \to \infty$, the Lyapunov exponent is given by the following simplified expression:

$$\boxed{\lambda = \lim_{p\to\infty}\left(\frac{1}{p(\theta_{ss})\displaystyle\int_{\theta_0}^{\theta_{ss}}\frac{d\theta}{p(\theta)}}\right) = \lim_{p\to\infty}\left(\frac{1}{\theta_{ss}-\theta_0+\displaystyle\int_{\theta_0}^{\theta_{ss}}\underbrace{\left(p'(\theta)\int\frac{d\theta}{p(\theta)}\right)}_{\equiv\mathcal{K}(\theta)}d\theta}\right) \approx \frac{1}{\theta_{ss}-\theta_0}} \tag{4.23}$$

which can be further generalized to the set of Quantum Lyapunnov exponents $\vec{\lambda} = \left[\frac{1}{\theta_{ss}-\theta_0}\right]_{\vec{\mu}}$ from the obtained result. Now, here it is important to note that the result here we have obtained is actually irrespective of $p'(\theta)$.

Using the equation 4.23, it is evident that when $\theta \to \bar{\theta}$, the Lyapunov exponent $\lambda \to \infty$. But the expression for the max Lyapunov exponent for maximal chaotic phenomena can be achieved from the MSS bound [39, 45] (see previous discussion). Thus the relative difference, $\widetilde{\delta\theta} = \theta_{ss} - \theta_0$ is constrained using equation 4.23 and equation 4.13 as:

$$\boxed{\textbf{\textit{Optimization Rate Bound :}}\ \ \widetilde{\delta\theta} \geq \frac{|\mathcal{B}|}{\pi\Gamma}} \tag{4.24}$$

The above inequality shows that the rate of optimization of QNN is restricted and the maximum optimization rate corresponds to the maximum chaotic system. An optimization limit is claimed in inequality 3.12, where the conditions are applied to neural architecture and observation matrix, but it is applicable for the whole learning trajectory. Here, in inequality 4.24, no conditions are imposed on neural architecture and observation matrix but we are analyzing in the different learning epochs of the training phase.

Optimization for the artificial neural network(ANN) with a high number of trainable parameters is empirically easy and can fit any training dataset resulting in zero training data. But the zero training error doesn't assure of the generalization capability of neural network [62–64]. When a neural network generalizes over a dataset, it understands the underlying structure of the dataset and thus reduces the difference between training error and testing error, called generalization error. Not much is discussed in the context of generalization or optimization in QNN as compared to ANN. Recently, [65] showed that QNN with the same structure as the corresponding ANN will have a better generalization property. In this paper, we mainly discussed or focused on the optimization property of QNN. We focused on the trajectory of parameters to its optimal parameter set in the learning manifold. But this is focusing on the one-half of the portion- training error. We introduce the generalization capability of QNN inspired by the notion of generalization in

ANN as discussed in [14, 17, 18, 62, 63, 65]. As the QNN framework used in the paper is a quantum-classical hybrid, where the parameters are optimized in classical SGD, we can use the concepts of generalization in ANN. Thereby, we focused on the fact that the generalization capability of a neural network is associated with the variance of the parameters [14, 21], higher generalization capability has high variance. Intuitively, it is taking into account the fact that a higher variance of parameters will increase the probability of finding a better optimal point which would result in better generalization capability. Thus the variance of parameters is a measure of the generalization in neural networks. The variance of parameters when the system is in a steady-state condition is given by the following expression:

$$\begin{aligned}\sigma_\theta^2\Big|_{\theta=\theta_{ss}} &= \frac{1}{K}\sum_{\mu}^{K}(\theta_\mu-\widehat{\theta})^2\Big|_{\theta=\theta_{ss}} \\ &\geq \frac{K}{\left(\displaystyle\sum_{\mu}^{K}\frac{1}{(\theta_\mu-\widehat{\theta}_0)^2}\right)\Bigg|_{\theta=\theta_{ss}}}+|\widetilde{\delta\theta}|^2+\frac{2K|\widetilde{\delta\theta}|}{\left(\displaystyle\sum_{\mu}^{K}\frac{1}{(\theta_\mu-\widehat{\theta}_0)}\right)\Bigg|_{\theta=\theta_{ss}}}\end{aligned} \tag{4.25}$$

where in the last step we have used the well known *Cauchy-Schwarz Inequality*. After working out a bit we derive the following bound on the variance:

$$\boxed{\textit{Generalization Capability Bound :}\ \sigma_\theta^2\Big|_{\theta=\theta_{ss}} \geq \frac{K}{\mathrm{Tr}(\lambda\lambda^T)}+\left(\frac{|\mathcal{B}|}{\pi\Gamma}\right)^2+\frac{2K|\mathcal{B}|}{\pi\Gamma\mathrm{Tr}(\sqrt{\lambda\lambda^T})}}. \tag{4.26}$$

where $\widehat{\theta}$ is the averaged parameters over its indexes. The inequality 4.25 shows that when the Lyapunov exponent is minimum or $\lambda \to 0$, then the generalization capability is at its maximum. This is shown as when the system shows limit cycles with $\lambda \to 0$ in phase space, the generalization capability reaches maximum. Interestingly, the work by [24] also argued that these oscillations in phases space are a crucial part of the stability of continuous memories in the human brain. The inequality 4.25 gives a theoretical perspective to this argument. Moreover, inequality 4.25 also shows that with an increase in inverse temperature β, the generalization capability increases. Correlating with the phase diagram [49, 50], this corresponds to many-body localization or non-equilibrium states as the work [21] showed for an artificial neural network. Along with oscillations, [24] argued coherent phases like many-body localization also plays a pivotal role in the stability of continuous memories. But on the other hand, increasing the inverse temperature also corresponds to a slower convergence rate as shown in the equation 2.9. Thus there is a trade-off between convergence rate and generalization capability as previously intuitively mentioned.

The change in the nature of the Lyapunov exponent is due to the matrix $\mathcal{G}^{-1}$. For $p \to 0$, the matrix $\mathcal{G} \to \infty$ and the system can be unstable or stable limit cycles depending on gradient p' with no significant chaos with maximum generalization capability. But for $p \to \infty$, using the Lyapunov exponent, the complexity $\mathcal{C}$ [59] can be given as:

$$\boxed{\boxed{\frac{\partial \log(\mathcal{C})}{\partial \theta} = \frac{1}{\theta_{ss} - \theta_0} \qquad \Longrightarrow \qquad \mathcal{C} = k \prod_{\mu} \exp\left(\frac{\theta_\mu}{\theta_{ss\mu} - \theta_{0\mu}}\right)}} \tag{4.27}$$

where $\theta_{ss,\mu}$ is the $\mu-$index of the steady state parameter θ_{ss}. and k is the constant to integration.

The out-of-order correlator OTOC [42, 43, 45, 60, 61] is given by OTOC $= \exp(-\mathcal{C})$ which shows that out-of-order correlator can be represented as:

$$\boxed{\boxed{\text{OTOC} = \exp\left[-k \prod_{\mu} \exp\left(\frac{\theta_\mu}{\theta_{ss,\mu} - \theta_{0,\mu}}\right)\right]}}. \tag{4.28}$$

The scrambling time t_* as shown in [37, 41–43, 61, 66] is given by:

$$\boxed{\boxed{t_* = \left(\theta_{ss} - \theta_0\right) \log(N)}}. \tag{4.29}$$

The entropy S as shown in [37, 44, 59, 61] can be expressed as:

$$\boxed{\boxed{S = \beta\left(\frac{\partial \mathcal{C}}{\partial t}\right) = \frac{2k|\mathcal{B}|}{\Gamma} \sum_{\mu} \left(\frac{\dot{\theta}_\mu}{\theta_{ss,\mu} - \theta_{0,\mu}}\right) \exp\left(\frac{\theta_\mu}{\theta_{ss,\mu} - \theta_{0,\mu}}\right) \prod_{\nu \neq \mu} \exp\left(\frac{\theta_\nu}{\theta_{ss,\nu} - \theta_{0,\nu}}\right)}}. \tag{4.30}$$

A 2-dimensional independent parameter system has been studied numerically for a dataset of $N = 10^3$. The evolution of parameters is governed by the equation 4.1 while ignoring the order $\mathcal{O}(\epsilon)$ terms. The independent parameter system has the jacobian matrix $\mathcal{G} = I$. The observation matrix B and $4-$rank tensor A^∞ is chosen. The $4-$rank tensor A^∞ is symmetric under permutation. The observation matrix is chosen to be identity $B = I$ and the eigenvalues of the reduced matrix A^∞ are taken as $\{0, 0, \hat{\epsilon}, \hat{\epsilon}\}$, which is varied in the results in Figures 4.1, 4.2 and 4.3. The evolution of parameters are evaluated using these inputs and then the Complexity is evaluated using equation 4.26 with $k = 2$. The evolution of Complexity is shown in Figure 4.1. The Lyapunov exponent is calculated by considering the change of y-axis value over the range of the x-axis value i.e. between the point of rising and point of saturation of Figure 4.1. The time of point of rising is shown as $t_1 = 0$ and the time of saturation is shown as $t_2 = 5$. So, mathematically, the Lyapunov

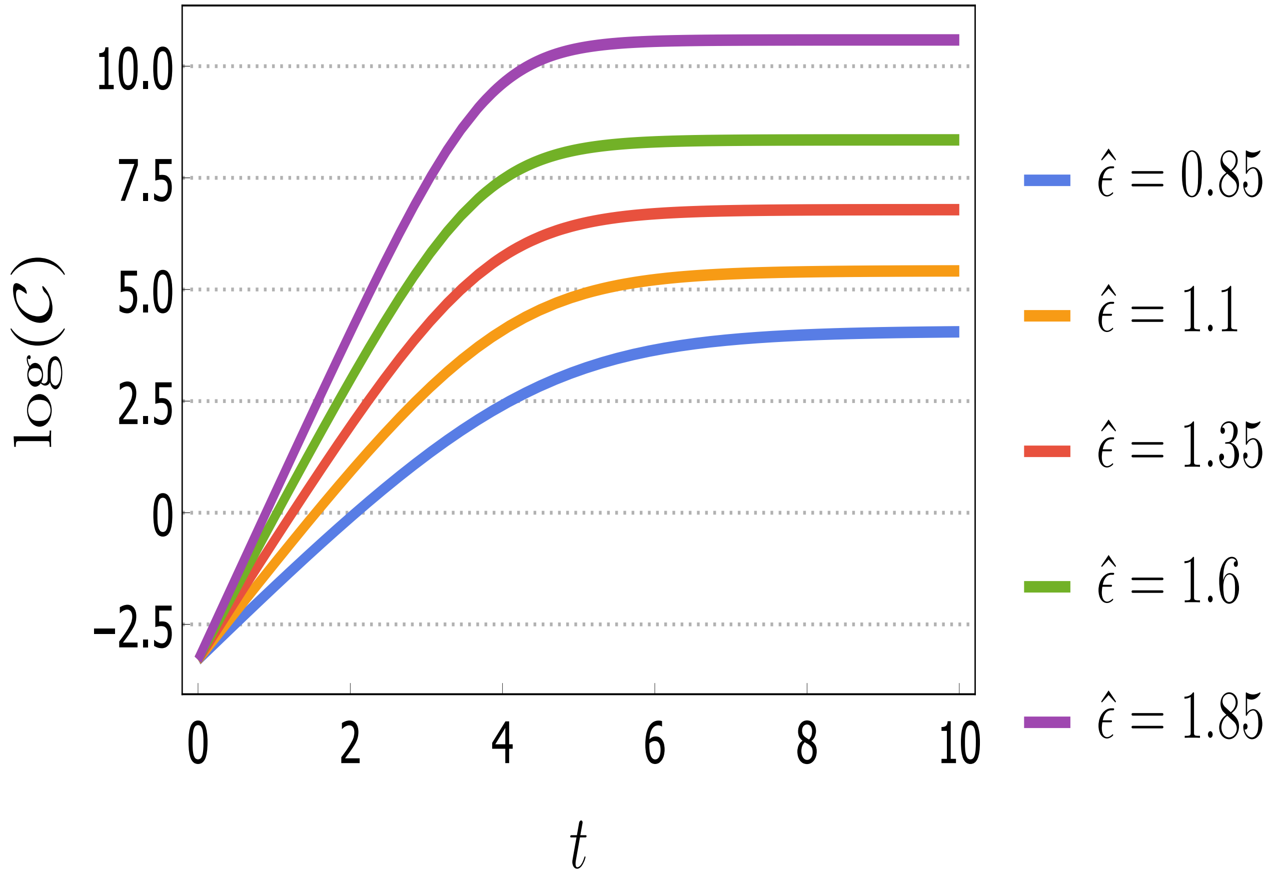

Figure 4.1: Evolution of logarithm of Complexity for a $P = 2$ parameter system with a variation of eigenvalue of *Encoder-Dataset* tensor A^∞

exponent is given by:

$$\boxed{\lambda = \frac{\log \mathcal{C}(t)\Big|_{t=t_2} - \log \mathcal{C}(t)\Big|_{t=t_1}}{t_2 - t_1}}. \tag{4.31}$$

Further the quantities OTOC and Entropy are plotted versus time t using equation 4.27 and 4.29 in Figures 4.2 and 4.3 respectively. The scrambling time t_* has been evaluated using equation 4.28. From Figure 4.1, it is important to observe that as the eigenvalues of *Encoder-Dataset* tensor increases, the maximum complexity of the system also increases. Thus, the role of *Encoder-Dataset* matrix A^∞ can be interpreted.

5 Conclusion

The paper uses Parameterized Quantum Circuits (PQCs) in the hybrid quantum-classical framework to perform optimization of quantum data with a classical gradient-based learn-

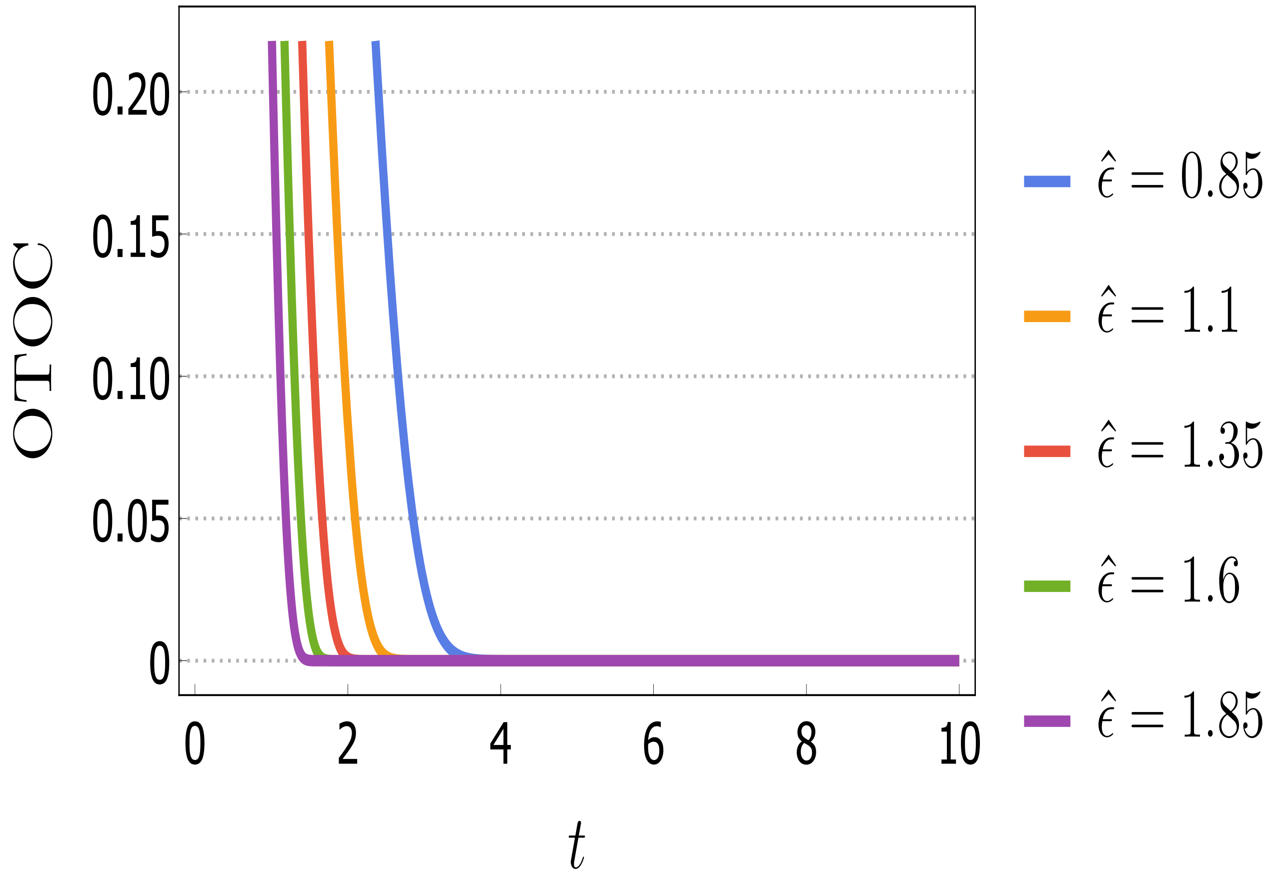

Figure 4.2: Evolution of out-of-order correlator (OTOC) for a $P = 2$ parameter system with a variation of eigenvalue of *Encoder-Dataset* tensor A^{∞}

Choas parameters Details across Eigenvalues			
$\hat{\epsilon}$	λ	t_*	T_{eq}
0.85	1.296	2.3148	0.2063
1.1	1.632	1.8382	0.2597
1.35	1.950	1.5385	0.3104
1.6	2.286	1.3123	0.3638
1.85	2.738	1.0957	0.4358

Table 4.1: Table showing Hyper parameter Details used in Neural Network for three different number of sampling

ing algorithm like stochastic gradient descent. The optimization is executed by updates the parameters in the unitary operators of quantum circuits. The trajectory of unitaries in the unitary space is correlated with the trajectory of parameters in a Riemannian manifold called Diffusion metric [35]. A statistical learning theory framework is introduced as

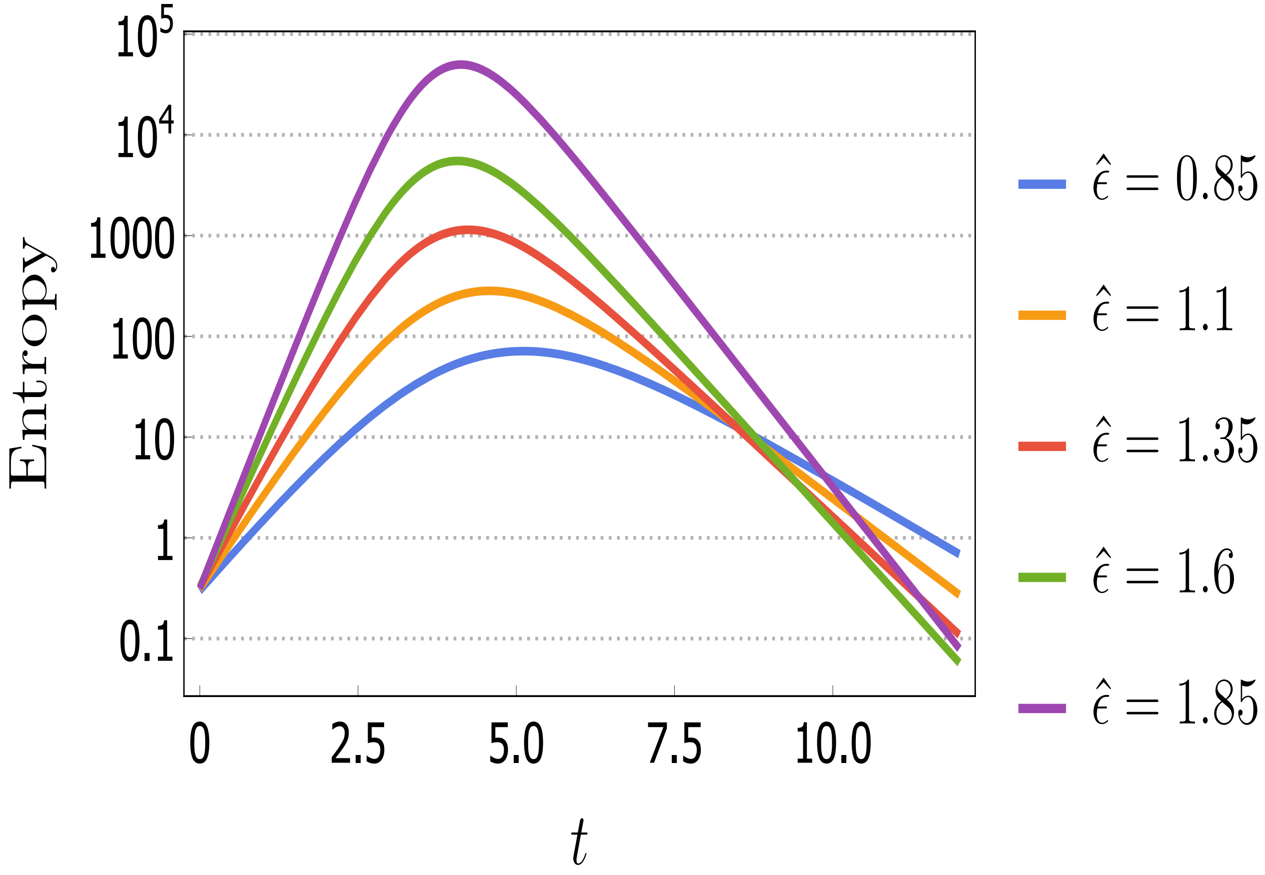

Figure 4.3: Evolution of entropy for a $P = 2$ parameter system with a variation of eigenvalue of *Encoder-Dataset* tensor A^{∞}

a quantum analog of [46]. In doing so, the relation between the learning dynamics and the neural architecture of QNN is established. The relation is used to also establish the dependency of the noise in SGD on the neural architecture of QNN using the Diffusion metric. Using the definition of complexity [36, 37, 52–58], the paper established dependency of the parameters on complexity. The parameterized Lyapunov exponent has been derived which estimates the stability of the system. The MSS bound [39] on the maximum Lyapunov exponents establishes a lower bound on the degree of optimization rate in parameters. The paper also proves that when the system executes limit cycles or oscillations in the phase space, the generalization capability of QNN is maximized. This is consistent with the biological notion argued by [24] that oscillations in phase space are important in the stability of the formation of continuous memories. The important contributions or results of the paper can be listed as follows:

- Correlation between the evolution of unitary operators in the unitary space with the trajectory of parameters in the Diffusion metric.

- Establishing Complexity, Lyapunov exponent, OTOC, and Entropy as a function of parameters of QNN.
- Estimating the stability of QNN using Lyapunov exponent.
- Proving that QNN with limit cycles or oscillations in phase space will have maximum generalization capability.
- A lower bound on optimization rate has been determined using the MSS bound.

Moreover, as neuroscience holds the fundamental architecture of neural networks, despite the proposal of quantum processing in neurons by Fisher [30] not much progress has been made to understand learning systems like human cognition from the perspective of quantum chaos and learning manifolds. Thus it not only becomes important to appreciate the application capability of QNN but also to analyze the quantum learning systems through the lens of statistical learning of QNN. A possible way of connecting the human brain with the models of neuroscience is correlating the famous Hodgkin-Huxley model [25] with the parameters' trajectory. Reverse engineering the QNN model that would correspond to the Hodgkin-Huxley model, can give much insight into the mechanism of the human brain.

Acknowledgements

The research fellowship of SC is supported by the J. C. Bose National Fellowship of Sudhakar Panda. Also SC take this opportunity to thank sincerely to Sudhakar Panda for his constant support and providing huge inspiration. SC also would line to thank School of Physical Sciences, National Institute for Science Education and Research (NISER), Bhubaneswar for providing the work friendly environment. Particularly SC want to give a separate credit to all the members of the EINSTEIN KAFFEE Berlin Alexanderplatz for providing work friendly environment, good espresso shots, delicious chocolate and caramel cakes and cookies, which helped to write the most of the part of the paper in that coffee shop in the last few months. SC also thank all the members of our newly formed virtual international non-profit consortium "Quantum Structures of the Space-Time & Matter" (QASTM) for elaborative discussions. SC also would like to thank all the speakers of QASTM zoominar series from different parts of the world (For the uploaded YouTube link look at: https://www.youtube.com/playlist?list=PLzW8AJcryManrTsG-4U4z9ip1J1dWoNgd) for supporting my research forum by giving outstanding lectures and their valuable time during this COVID pandemic time. Last but not the least, we would like to acknowledge our debt to the people belonging to the various part of the world for their generous and steady support for research in natural sciences.

References

[1] Arute *et al.* Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, Oct 2019.

[2] Zhao-Yun Chen, Qi Zhou, Cheng Xue, Xia Yang, Guang-Can Guo, and Guo-Ping Guo. 64-qubit quantum circuit simulation. *Science Bulletin*, 63(15):964–971, Aug 2018.

[3] M. Mohseni, Peter Read, Hartmut Neven, Sergio Boixo, Vasil Denchev, Ryan Babbush, Austin Fowler, Vadim Smelyanskiy, and John Martinis. Commercialize early quantum technologies. *Nature*, 543:171–174, 03 2017.

[4] John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, Aug 2018.

[5] Joseph K Iverson and John Preskill. Coherence in logical quantum channels. *New Journal of Physics*, 22(7):073066, Aug 2020.

[6] Xiao Yuan. A quantum-computing advantage for chemistry. *Science*, 369(6507):1054–1055, 2020.

[7] Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J. Osborne, Robert Salzmann, Daniel Scheiermann, and Ramona Wolf. Training deep quantum neural networks. *Nature Communications*, 11(1):808, Feb 2020.

[8] Ashish Kapoor, Nathan Wiebe, and Krysta Svore. Quantum perceptron models. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3999–4007. Curran Associates, Inc., 2016.

[9] R. C. Wiersema and H. J. Kappen. Implementing perceptron models with qubits. *Phys. Rev. A*, 100:020301, Aug 2019.

[10] Patrick Rebentrost, Maria Schuld, Leonard Wossnig, Francesco Petruccione, and Seth Lloyd. Quantum gradient descent and newton's method for constrained polynomial optimization. *New Journal of Physics*, 21(7):073023, jul 2019.

[11] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Phys. Rev. A*, 98:032309, Sep 2018.

[12] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, nov 2019.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* The MIT Press, 2016.

[14] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys, 2016.

[15] Ankan Dutta and Arnab Rakshit. Geometry perspective of estimating learning capability of neural networks, 2020.

[16] Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen,

Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2299–2308, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[17] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information, 2017.

[18] Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes, 2019.

[19] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1019–1028. JMLR.org, 2017.

[20] Andrew K. Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks, 2018.

[21] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks, 2017.

[22] Huitao Shen, Pengfei Zhang, Yi-Zhuang You, and Hui Zhai. Information scrambling in quantum neural networks. *Phys. Rev. Lett.*, 124:200504, May 2020.

[23] David Deutsch and Patrick Hayden. Information flow in entangled quantum systems. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 456(1999):1759–1774, Jul 2000.

[24] Han Yan, Lei Zhao, Liang Hu, Xidi Wang, Erkang Wang, and Jin Wang. Nonequilibrium landscape theory of neural networks. *Proceedings of the National Academy of Sciences*, 110(45):E4185–E4194, 2013.

[25] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952.

[26] Henri Korn and Philippe Faure. Is there chaos in the brain? ii. experimental evidence and related models. *Comptes Rendus Biologies*, 326(9):787 – 840, 2003.

[27] L. P. Wang, E. E. Pichler, and J. Ross. Oscillations and chaos in neural networks: an exactly solvable model. *Proceedings of the National Academy of Sciences of the United States of America*, 87(23):9467–9471, Dec 1990. 2251287[pmid].

[28] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos, 2016.

[29] Alex Potapov and M. Ali. Robust chaos in neural networks. *Physics Letters A*, 277:310–322, 12 2000.

[30] Matthew P.A. Fisher. Quantum cognition: The possibility of processing with nuclear spins in the brain. *Annals of Physics*, 362:593 – 602, 2015.

[31] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. 2017.

[32] Renqian Luo, Fei Tian, Tao Qin, Enhong Chen, and Tie-Yan Liu. Neural architecture optimization, 2018.

[33] Liam Li and Ameet Talwalkar. Random search and reproducibility for neural architecture search, 2019.

[34] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient neural architecture search via parameter sharing, 2018.

[35] Soatto Stefano Fioresi Rita, Chaudhari Pratik. A geometric interpretation of stochastic gradient descent using diffusion metrics. *Entropy 22, no. 1: 101*, 2020.

[36] Adam R. Brown, Daniel A. Roberts, Leonard Susskind, Brian Swingle, and Ying Zhao. Holographic complexity equals bulk action? *Physical Review Letters*, 116(19), May 2016.

[37] Leonard Susskind. Three lectures on complexity and black holes, 2018.

[38] Adam R. Brown, Daniel A. Roberts, Leonard Susskind, Brian Swingle, and Ying Zhao. Complexity, action, and black holes. *Physical Review D*, 93(8), Apr 2016.

[39] Juan Maldacena, Stephen H. Shenker, and Douglas Stanford. A bound on chaos. *Journal of High Energy Physics*, 2016(8):106, Aug 2016.

[40] Stephen H. Shenker and Douglas Stanford. Black holes and the butterfly effect. *Journal of High Energy Physics*, 2014(3):67, Mar 2014.

[41] Brian Swingle and Debanjan Chowdhury. Slow scrambling in disordered quantum systems. *Physical Review B*, 95(6), Feb 2017.

[42] Brian Swingle, Gregory Bentsen, Monika Schleier-Smith, and Patrick Hayden. Measuring the scrambling of quantum information. *Physical Review A*, 94(4), Oct 2016.

[43] Brian Swingle. Unscrambling the physics of out-of-time-order correlators. *Nature Physics*, 14(10):988–990, Oct 2018.

[44] Adam R. Brown and Leonard Susskind. Second law of quantum complexity. *Physical Review D*, 97(8), Apr 2018.

[45] Sayantan Choudhury. The cosmological otoc: Formulating new cosmological micro-canonical correlation functions for random chaotic fluctuations in out-of-equilibrium quantum statistical field theory. *Symmetry*, 12(9):1527, Sep 2020.

[46] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

[47] Elizabeth Crosson, Tomas Jochym-O'Connor, and John Preskill. Universal quantum computation in thermal equilibrium. In *APS March Meeting Abstracts*, volume 2018 of *APS Meeting Abstracts*, page S28.001, January 2018.

[48] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991.

[49] Ehud Altman. Many-body localization and quantum thermalization. *Nature Physics*, 14(10):979–983, Oct 2018.

[50] Matthias Vojta. Quantum phase transitions. *Reports on Progress in Physics*, 66(12):2069–2110, Nov 2003.

[51] Cesar A. Agón, Matthew Headrick, and Brian Swingle. Subsystem complexity and holography. *Journal of High Energy Physics*, 2019(2), Feb 2019.

[52] Douglas Stanford and Leonard Susskind. Complexity and shock wave geometries. *Physical Review D*, 90(12), Dec 2014.

[53] Fernando G. S. L. Brandão, Wissam Chemissany, Nicholas Hunter-Jones, Richard Kueng, and John Preskill. Models of quantum complexity growth, 2019.

[54] Kanato Goto, Hugo Marrochio, Robert C. Myers, Leonel Queimada, and Beni Yoshida. Holographic complexity equals which action? *Journal of High Energy Physics*, 2019(2), Feb 2019.

[55] Alice Bernamonti, Federico Galli, Juan Hernandez, Robert C Myers, Shan-Ming Ruan, and Joan Simón. Aspects of the first law of complexity. *Journal of Physics A: Mathematical and Theoretical*, 53(29):294002, Jul 2020.

[56] Dean Carmi, Robert C. Myers, and Pratik Rath. Comments on holographic complexity. *Journal of High Energy Physics*, 2017(3), Mar 2017.

[57] Minyong Guo, Juan Hernandez, Robert C. Myers, and Shan-Ming Ruan. Circuit complexity for coherent states. *Journal of High Energy Physics*, 2018(10), Oct 2018.

[58] Robert A. Jefferson and Robert C. Myers. Circuit complexity in quantum field theory. *Journal of High Energy Physics*, 2017(10), Oct 2017.

[59] Kaushik Y. Bhagat, Baibhab Bose, Sayantan Choudhury, Satyaki Chowdhury, Rathindra N. Das, Saptarshhi G. Dastider, Nitin Gupta, Archana Maji, Gabriel D. Pasquino, and Swaraj Paul. The generalized otoc from supersymmetric quantum mechanics: Study of random fluctuations from eigenstate representation of correlation functions, 2020.

[60] Gregory Bentsen, Brian Swingle, Monika Schleier-Smith, and Patrick Hayden. Measuring signatures of quantum chaos in strongly-interacting systems. In *APS Division of Atomic, Molecular and Optical Physics Meeting Abstracts*, volume 2017 of *APS Meeting Abstracts*, page T7.009, April 2017.

[61] Parth Bhargava, Sayantan Choudhury, Satyaki Chowdhury, Anurag Mishara, Sachin Panneer Selvam, Sudhakar Panda, and Gabriel D. Pasquino. Quantum aspects of chaos and complexity from bouncing cosmology: A study with two-mode single field squeezed state formalism, 2020.

[62] Jingling Li, Yanchao Sun, Jiahao Su, Taiji Suzuki, and Furong Huang. Understanding generalization in deep learning via tensor methods, 2020.

[63] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals.

Understanding deep learning requires rethinking generalization, 2016.

[64] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5949–5958, Red Hook, NY, USA, 2017. Curran Associates Inc.

[65] JinZhe Jiang, Xin Zhang, Chen Li, YaQian Zhao, and RenGang Li. Generalization study of quantum neural network, 2020.

[66] Nima Lashkari, Douglas Stanford, Matthew Hastings, Tobias Osborne, and Patrick Hayden. Towards the fast scrambling conjecture. *Journal of High Energy Physics*, 2013(4), Apr 2013.