Exploring Neural Network-Quantum Field Theory Correspondence

Master Thesis
Department of Information & Computing Sciences

Student
Julian A.W. Markus
6134165

Supervisor
Dr. Ro Jefferson

October 22, 2023

Contents

Abstract

Signal propagation is an important factor impacting the trainability of neural networks where better signal propagation leads to better training performance. Signal propagation can be ordered, where signals decay, chaotic, where signals decorrelate or critical, where a signal neither decays nor decorrelates. Mean Field Theory (MFT) based formulations exist to predict signal propagation in neural networks, which have shown empirical success in achieving better training performance. MFT formulations, however, operate under the assumption of infinitely wide network layers, which do not exist in practice. Quantum field theory-based formulations aim to correct this assumption by formulating a new signal propagation prediction called the NN-QFT correspondence. In this thesis, the accuracy of predictions from the NN-QFT correspondence have been empirically explored. The NN-QFT correspondence appears to be data-invariant for feedforward neural networks and accurate in predictions for networks sufficiently distanced from the critical point. For linear networks, the theory appears to predict the critical point correctly. The theoretical critical point for nonlinear networks does not match the empirically found critical point. Critical behaviour observed for these networks appears to agree with theory. Additionally, observations appear to correlate the periodic behaviour of individual neurons and the critical to chaotic transition.

1 Introduction

According to LeCun et al. [1], automatic learning by training a neural network (NN) to a global performance criterion outperforms more classical machine learning methods that feature handcrafted heuristics. Classical machine learning models used for classification typically feature two components: A feature extraction module and a trainable classifier module. The goal of the feature extraction module is to filter and transform inputs from a given dataset into a feature vector from which the classifier module predicts class scores. The difficulty of such a classification technique exists in the fact that the feature extractor is often entirely handcrafted, which means that the ability to classify data accurately is determined by the ability of the designer of the feature extractor to come up with an appropriate set of features [1]. A multilayer NN, on the contrary, unifies the process of feature extraction and classification. Here, each layer transforms the input data in such a way that the eventual network output is a prediction of class scores. In practice, training occurs by optimising weights to minimise a loss function through backpropagation.

Besides the ease of optimisation, an additional benefit is the high expressivity of neural networks. A network of higher expressivity can perform more complex approximation tasks, and the easiest way to create a network of higher expressivity is by adding more layers to the network [2]. The dense connections between layers of a neural network mean that adding extra layers to the network leads to a theoretically exponential increase in expressivity. This increase in expressivity is also not necessarily a result of the increase in neuron count, in fact in the Imagenet competition Szegedy et al. [3] present a network that uses $12\times$ fewer parameters than Krizhevsky et al. [4], while being significantly more accurate. This higher classification accuracy while using fewer neurons has been achieved by using a deeper network (along with several other more clever architectural decisions). Szegedy et al. [3] note, however, that deepening a network to benefit its performance comes at the cost of a (disproportional) increase in computational effort required to train the network.

This disproportional increase in computational effort is, at least partially, due to the fact that deeper networks are increasingly susceptible to the Exploding/Vanishing Gradients Problem (EVGP) [5], making it so that any increase in network performance due to higher expressivity is potentially nullified by a decreased ability to train the network efficiently. This makes for a delicate balancing act of balancing expressivity and trainability when designing NN architectures [6].

Therefore, it is desirable to find ways to boost the trainability of networks without compromising the expressivity. Attempts to mitigate the EVGP and, therefore, increase trainability can be made by either developing more clever network architectures specific to the machine learning task at hand, i.e. the work by Szegedy et al. [3], more generally applicable architectural decisions, such as residual networks, or by manipulating fundamental parameters of the network to ensure we do not encounter the EVGP. The latter can be achieved by stabilising certain metrics across increasing network depth. Examples of such metrics include normalising Layer Sequential Unit Variance as in [7] or ensuring dynamical isometry is maintained as in [8], with both showing improved training performance compared to a baseline.

At the core of the EVGP lies the issue that signals fail to propagate through all layers of the network. Mean Field Theory (MFT) creates a tractable way of predicting signal propagation through a neural network.

The advantage of modelling signal propagation using MFT is that the parametrically complex NN (i.e. each individual neuron) can be approximated by a significantly smaller number of parameters. Theoretical relations formulated using Mean Field Theory (MFT) provide useful predictions of the signal propagation and, therefore, trainability of feedforward- and recurrent neural networks, examples of this include Poole et al. [9], who predict the propagation of input signals using MFT, Schoenholz et al. [10], based on [9], present the existence of "depth scales" that can predict the depth limit to which Deep Neural Networks (DNNs) remain trainable and Chen et al. [11], who use MFT predictions for Recurrent Neural Networks (RNNs) to enforce dynamical isometry, aiding in trainability.

The MFT predictions make inconsistencies possible due to the assumption of infinite-width networks and their finite real-world counterparts as is experienced empirically by [10, 11]. A new prediction is proposed by Grosvenor and Jefferson [12] modelling signal propagation using Quantum Field Theory (QFT). QFT differs from MFT by discarding the infinite-width assumption; this QFT formulation aims to provide a more accurate prediction of signal propagation in NNs compared to MFT. This thesis aims to explore if the predictions made by this NN-QFT correspondence are empirically observed in neural networks.

The main question we seek to answer is: *Does the NN-QFT correspondence provide an accurate prediction of empirically observed NN behaviour?*

With the potential follow-up question: *What network dynamics causes deviations between predicted and observed NN behaviour?*

2 Theory of signal propagation

In 1990 Langton [13] presented a paper that studied the behaviour of cellular automata. This paper is (to the author's knowledge) one of the first papers considering chaos in dynamical systems, specifically as a way to support computation. Langton [13] mentions specifically that optimal conditions for information transmission, storage and modification occur at and near the phase transition from *ordered* to *chaotic* behaviour. To ease into the subject, we shall start with citing Langton's [13] observations on cellular automata as a toy example: A dynamic system (here, cellular automata) consists of an initial state (1D array of cells) and a tunable transition function. The dynamics of this system can be changed by tuning parameters of the transition function, ranging from ordered to chaotic. Figure 1 shows the behaviour of the cellular automata over a certain number of timesteps. The tunable parameter λ is used by Langton [13] to control the probability of a cell transitioning to the next timestep, which gives rise to the behaviour described as follows: Figure 1(a) shows an example of the ordered regime, in this regime the state of a system decays to a common state within a finite amount of time (here, all-white cells). The chaotic regime, as figure 1(b) exemplifies, contrasts this, where states transition to chaotic and unstructured subsequent states. A point of balance lies between the ordered and chaotic regimes, named the *critical point*. At the critical point, a system shows neither purely ordered nor chaotic behaviour; instead, the system may show structural behaviour in its phase transitions, leading to cell activity propagating much deeper w.r.t. the number of state transitions before decaying to an ordered or chaotic state. This critical behaviour can be seen in figure 1(c) which shows how the system is neither fully chaotic nor ordered until finally decaying to an ordered state after more than 12.000 timesteps.

We can draw parallels from Langton [13] (or rather, a dynamic system in general) to that of a vanilla Deep Neural Network (DNN). Where, in DNN terminology, the initial state is the input vector, and a tunable[1] transition function exists in the form of the network weights.

Given that the EVGP arises from either a lack of signals propagating leading to the backpropagation algorithm to perform a smaller update of network weights than desired, or when input signals propagate in a chaotic manner, the backdropagation algorithm may overcorrect weights leading to an exploding gradient if these overcorrections accumulate. For a chaotic signal propagation, the additional risk exists to saturate the activation function, hindering trainability further [14]. From this it should be clear that to aid the trainability of the DNN we desire a structured signal propagation within a reasonable order of magnitude[2]. Langton [13] has shown these properties to exist around the critical point in the dynamical system of cellular

[1] "Tunable" here refers to the ability to initialise the network with different weights, and not necessarily the act of tuning weights through network training (although the latter is not completely unrelated as will be mentioned later).

[2] Shuch that the signal does not saturate any activation function.

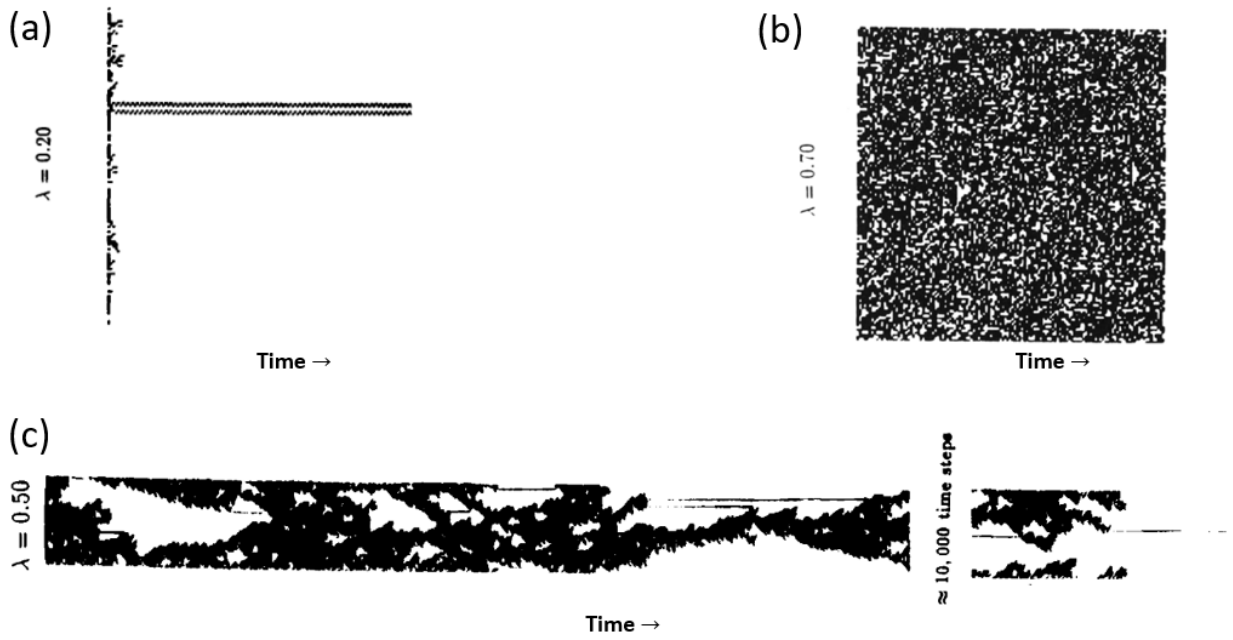


Figure 1: 3 different evolutions of cellular automata showing (a) A system of cellular automata in the ordered regime, (b) cellular automata falling in the chaotic regime, and (c) cellular automata at the edge between ordered and chaotic regimes. Source: [13]

automata, Sompolinsky et al. [15] show similar properties to indeed exist in neural networks.

2.1 Mean Field Theory

As we are now aware that the initialisation of network weights influences the regime (ordered, critical, or chaotic) of the network w.r.t. signal propagation. The question arises how to initialise these weights such that we get the desired, critical signal propagation?

Several works have used Mean Field Theory (MFT) to approximate the behaviour of neural networks [9–11, 15, 16]. Mean Field Theory allows for the study of a high-dimensional model's behaviour by studying a simpler model, where the simpler model averages over degrees of freedom (here, neurons) [17].

Constructing MFT formulations describing neural networks, therefore, involves averaging over neurons per network layer. As will be formalised in subsequent sections, MFT formulations typically assume the network's layer weights to be zero-mean, normally distributed. This assumption gives us a tunable transition function parameter in the form of the standard deviations, σ, used to initialise these weight distributions.

2.2 MFT Network Notation

Poole et al. [9] consider a deep feedforward network consisting of D weights $W^1, ..., W^D$, with $D+1$ 'neural activity vectors' $x^0, ..., x^{D+1}$ consisting of N_l neurons each s.t. $x^l \in \mathbb{R}^{N_l}$. With network weights and biases drawn from distributions as follows: $W^l_{ij} \sim \mathcal{N}\left(0, \sigma_w^2/N_{l-1}\right)$, and $b_i \sim \mathcal{N}\left(0, \sigma_b^2\right)$, where σ_w^2/N_{l-1} is chosen to ensure neuron activity remains $\mathcal{O}(1)$ and thus preventing saturation of the activation function. The network is then formulated as follows:

$$x^l = \phi(h^l) \quad h^l = W^l x^{l-1} + b^l \quad \text{for } l = 1, ..., D. \tag{1}$$

2.3 MFT Signal propagation

As a measure of signal propagation, one can observe the convergence and/or divergence of a pair of inputs as they propagate, per layer, through the network. Observing the similarity between two sets of data (the

inputs, in this case) can be done with the (Pearson) correlation coefficient. For the theoretical prediction of the correlation coefficient Poole et al. [9] present the following formulation, given two inputs $x^{0,1}$ and $x^{0,2}$:

$$q^l_{ab} = \frac{1}{N_l} \sum_{i=1}^{N_l} h^l_i(x^{0,a}) h^l_i(x^{0,b}) \quad a, b \in 1, 2. \tag{2}$$

The diagonal terms q^l_{11} and q^l_{22} are predicted by the length map/variance:

$$q^l = \sigma^2_w \int \mathcal{D}z \phi\left(\sqrt{q^{l-1}} z\right) + \sigma^2_b \quad \text{for} l = 2, ..., D. \tag{3}$$

Where $\mathcal{D}z = \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ is a standard Gaussian measure describing the normal distribution, $q^1 = \sigma^2_w q^0 + \sigma^2_b$, and $q^0 = \frac{1}{N_0} x^0 \cdot x^0$.

The covariance q_{12} is then predicted by the following relation:

$$q^l_{12} = \sigma^2_w \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1) \phi(u_2) + \sigma^2_b, \tag{4}$$

With:

$$u_1 = \sqrt{q^{l-1}_{11}} z_1 \quad u_2 = \sqrt{q^{l-1}_{22}} \left[c^{l-1}_{12} z_1 + \sqrt{1 - (c^{l-1}_{12})^2} z_2 \right]$$

Here, u_1 and u_2 are the Gaussian measure z, scaled by the respective parameters from the previous layer, which combine to formulate q^l_{12}. This way, the signal propagation, expressed as the correlation coefficient, can recursively be computed/predicted for each layer based on the weight distributions of the network. Given these formulations, the correlation coefficient is given as $c^l_{12} = q^l_{12}(q^l_{11} q^l_{22})$.

The reason for using the correlation length of a pair of inputs is that given the initial correlation between two inputs as a reference, if the correlation between inputs increases over time (or rather, over layers), this indicates that the two signals of these inputs are converging, which is characteristic of an ordered network. On the contrary, a correlation that decays (exponentially) over time indicates a divergence of signals, characteristic of a chaotic network. The strength with which the correlation increases/decays indicates how far a network lies into the regime, i.e. a decay to 0 correlation in 5 network layers is a more chaotic network than a network that decays in 100 layers. Around the critical point, we then expect the correlation to remain roughly constant, or at least a significantly slower decay to complete (de-)correlation.

To this end, to further aid in interpreting the correlations over time, Poole et al. [9] also present an "iterative correlation coefficient map", or $\mathcal{C}$-map, using the measure χ_1 to interpret the evolution of states under the dynamics of the aforementioned equations. χ_1 is defined as:

$$\chi_1 \equiv \left. \frac{\partial C^l}{\partial C^{l-1}} \right|_{c=1} = \sigma^2_w \int \mathcal{D}z \left[\phi'(\sqrt{q^*} z) \right]^2. \tag{5}$$

χ_1 can be understood as the multiplicative stretch factor [9]. They note that the regime of a network can be identified through the slope where $\chi_1 < 1$, and a fixed point $c^* = 1$ is characteristic of an ordered network, i.e. as indicated by χ_1, the correlation length is compressed due to signals converging on a common state, $c^* = 1$ is representative of signals having converged to this state. A chaotic network is indicated by $\chi_1 > 1$ and $c^* < 1$, i.e. $\chi_1 > 1$ causes the correlation length between signals to increase and de-correlate, where maximum de-correlation has been reached at the fixed point $c^* = 0$. A network at criticality will have $\chi_1 \approx 1$, i.e. the slope of correlations over time lies on, or very close to, the diagonal of the $\mathcal{C}$-map, $\chi_1 \approx 1$ here indicates that the propagating signals do not have rapidly converging/diverging correlation lengths, the fixed point will then ideally lie somewhere close to the value of the initial correlation: c^0_{12}.

2.4 Depth Scales

Building upon the results of Poole et al. [9], Schoenholz et al. [10] present two depth scales by which information of a single or pair of inputs may propagate the network. These depth scales are expected to

follow the asymptotic relations $|q^l_{aa} - q^*| \sim e^{-l/\xi_q}$ and $|c^l_{ab} - c^*| \sim e^{-l/\xi_c}$ for single and paired inputs respectively, where q^* and c^* are the fixed points. Schoenholz et al. [10] fitted the empirical results of $|q^l_{aa} - q^*|$ and $|q^l_{aa} - q^*|$ to exponential functions in order to obtain the *depth scales* ξ_q and ξ_c respectively. By repeatedly obtaining depth scales for varying hyperparameter configurations, they have found that the depth scale ξ_c has an asymptote at the critical point. Initialising networks within the depth scale means that signals are expected to propagate the full depth of the network and would, therefore, offer a universal constraint for hyperparameter selection [10]. For networks that lie at the critical point, the divergence of the depth scale suggests that such a network may be arbitrarily deep and propagate a signal along its entire depth.

Besides the theoretical results, empirical results show that network configurations lying within $6\xi_c$ are trainable to high accuracy. It remains an open question why networks within specifically a factor 6 of the depth scale ξ_c remain trainable.

2.5 Quantum Field Theory

Quantum field theory (QFT) potentially allows for more accurate correlation predictions by assuming a network is of finite width, as it is in practice. A consequence of finite-width networks is that the weight distributions previously used to formulate correlation predictions are no longer perfectly normally distributed. A perfect normal distribution of infinite width has two cumulants that can be non-zero: the first cumulant, the mean, and the second cumulant, the variance. A normal distribution, as we encounter them in neural networks, therefore, has non-zero cumulants beyond the second cumulant. E.g. a distribution may have a non-zero third cumulant, i.e. there exists a certain asymmetry in the distribution. These higher-order cumulants may cause interactions between neurons -within a layer- to occur, which affect signal propagation and correlation, that are not accounted for in MFT, which could lead to a significant discrepancy between theoretical and empirical results. The goal of the QFT formulation is then to account for these interactions in an attempt to create a more accurate prediction w.r.t. the signal correlation. This could, in turn, explain the mismatch between theoretical and empirical trainability observed by Schoenholz et al. [10].

2.6 QFT Network Notation

Grosvenor and Jefferson [12] expand the MFT predictions of Poole et al. [9] by applying finite-width corrections to the formulation. Additionally, constraints are made to the weight distributions of the network to aid in the formulation. The formulation Grosvenor and Jefferson [12] present considers neural networks with identical weights across all layers. As a side effect, this constrains networks to equally sized layers. A network abiding by these constraints shares many similarities with the 'vanilla' implementation of a recurrent neural network. To provide as general of a formulation as possible, Grosvenor and Jefferson [12] therefore expand the network formulation to include recurrent neural networks. The theoretical equation for the network is as follows:

$$f(x, h) = (A - \gamma)h + Bx + W\phi(h) + U\varphi(x) + b. \tag{6}$$

Parameters A and B are included for generality and can be ignored in our experimental setup by setting them to [0]. Taking equation 6 and transforming it into a concrete neural network is done as follows: Given weights initialised by taking: $W_{ij} \sim \mathcal{N}\left(0, \sigma_w^2/L\right)$, $V_{ij} \sim \mathcal{N}\left(0, \sigma_v^2/N\right)$ and $b_i \sim \mathcal{N}\left(0, \sigma_b^2\right)$.

$$h^{l+1} = W\phi\left(h^l\right) + V\varphi\left(x^l\right) + b. \tag{7}$$

With x^l as the input vector for layer l, h^l is the previous hidden state, and h^{l+1} is the hidden state for the next layer of the network. ϕ and φ are arbitrary activation functions s.t. $\phi : \mathbb{R} \to \mathbb{R}$, $\varphi : \mathbb{R} \to \mathbb{R}$. For simplicity and unless otherwise specified, we will assume $\varphi = \phi$.

It should be emphasised that the network formulation of equation 1 is identical to 7 if the former uses identical weights for all its layers and the latter receives an input vector that is *only non-zero for* x^0.

2.7 QFT Signal propagation

The primary interest of the work by Grosvenor and Jefferson [12] is the propagator $G_{hh}(\tau) = \langle h_i(t_1)h_i(t_2)\rangle$. Where, contrary to Poole et al. [9], the correlation length is determined on the input of a single signal between

two states separated in time, $h_i(t_1)$ and $h_i(t_2)$ where $\tau = t_1 - t_2$. Given the difference in notation, we shall repeat the formulation for the correlation between two states as presented by Grosvenor and Jefferson [12] here:

Two states, $h_i(t_1)$ and $h_i(t_2)$, are random Gaussian variables (due to the nature of the network weights) with covariance matrix $G_{hh}(\tau)$. h_1 and h_2, that is shorthand for $h_i(t_1)$ and $h_i(t_2)$ respectively, can be drawn from the bivariate normal distribution as follows:

$$(h_1, h_2) \sim \mathcal{N}\left(0, \begin{pmatrix} G_{hh}(0) & G_{hh}(\tau) \\ G_{hh}(\tau) & G_{hh}(0) \end{pmatrix}\right) = \mathcal{N}\left(0, \begin{pmatrix} c_0 & c_\tau \\ c_\tau & c_0 \end{pmatrix}\right) \tag{8}$$

Where $c_0 = G_{hh}(0)$ for $t_1 \neq t_2$ and $c_\tau = G_{hh}(\tau)$ otherwise.

$$C_{\phi\phi}(t_1, t_2) = \langle \phi(h_1)\phi(h_2) \rangle = \int \mathcal{D}h_1 \mathcal{D}h_2 \phi(h_1)\phi(h_2) \tag{9}$$

With the Pearson correlation coefficient noted as $\rho = \frac{c_\tau}{c_0}$. Filling in for the correlated Gaussian variables h_1, h_2 then gives:

$$h_1 = \sqrt{c_0} h_a, \qquad h_2 = \sqrt{c_0}\left(\rho h_a + \sqrt{1-\rho^2} h_b\right) \tag{10}$$

Which accordingly provides an estimation of the distribution at any given depth. Now the equation:

$$C_{\phi\phi} = \int \mathcal{D}h_a \mathcal{D}h_b \phi\left(\sqrt{c_0} h_a\right) \phi\left(\sqrt{c_0}\left(\rho h_a + \sqrt{1-\rho^2} h_b\right)\right) \tag{11}$$

provides a substitute of

When constructing the models for signal propagation, it is beneficial to define so-called auxiliary fields which, describe correlations of individual components of the network, affecting the propagator $G_{hh}(\tau)$. These auxiliary fields are calculated by use of equation 11. In reference to the individual components present in equation 6 Grosvenor and Jefferson [12] produce:

$$\begin{aligned}
\mathfrak{A}_0^{\alpha\beta}(t_1, t_2) &= \sigma_A^2 \langle h_i^\alpha(t_1) h_i^\beta(t_2) \rangle_{X_0} = \sigma_A^2 C_{hh}^{\alpha\beta}(t_1, t_2), \\
\mathfrak{B}_0^{\alpha\beta}(t_1, t_2) &= \sigma_B^2 \langle x_i^\alpha(t_1) x_i^\beta(t_2) \rangle_{X_0} = \sigma_B^2 C_{xx}^{\alpha\beta}(t_1, t_2), \\
\mathfrak{W}_0^{\alpha\beta}(t_1, t_2) &= \sigma_w^2 \langle \phi(h_i^\alpha(t_1)) \phi(h_i^\beta(t_2)) \rangle_{X_0} = \sigma_w^2 C_{\phi\phi}^{\alpha\beta}(t_1, t_2), \\
\mathfrak{U}_0^{\alpha\beta}(t_1, t_2) &= \sigma_u^2 \langle \varphi(x_i^\alpha(t_1) \varphi(x_i^\beta(t_2)) \rangle_{X_0} = \sigma_u^2 C_{\varphi\varphi}^{\alpha\beta}(t_1, t_2), \\
\tilde{A}_0^{\alpha\beta}(t_1, t_2) &\simeq \sigma_A^2 \langle \tilde{z}_i^\alpha(t_1) \tilde{z}_i^\beta(t_2) \rangle_{X_0} = 0, \\
\tilde{B}_0^{\alpha\beta}(t_1, t_2) &\simeq \sigma_B^2 \langle \tilde{z}_i^\alpha(t_1) \tilde{z}_i^\beta(t_2) \rangle_{X_0} = 0, \\
\tilde{W}_0^{\alpha\beta}(t_1, t_2) &\simeq \sigma_w^2 \langle \tilde{z}_i^\alpha(t_1) \tilde{z}_i^\beta(t_2) \rangle_{X_0} = 0, \\
\tilde{U}_0^{\alpha\beta}(t_1, t_2) &\simeq \sigma_u^2 \langle \tilde{z}_i^\alpha(t_1) \tilde{z}_i^\beta(t_2) \rangle_{X_0} = 0
\end{aligned} \tag{12}$$

Where α and β are labels of the 'state copies' such that $c^{\alpha\beta}(t_1, t_2) = \langle h_i^\alpha(t_1) h_i^\beta(t_2) \rangle$, for the propagator $G_{hh}(\tau)$ we are interested in the propagation when $\alpha = \beta$. Additionally, $\tilde{z}$ is the "response field" derived from solving the Stochastic Differential Equation of the RNN. As the name implies, this response field is used to interpret the response to perturbations of the system [12]. Specifically note that as A and B are included for generality, in practice, they are set such that $\sigma_A^2 = 0$ and $\sigma_B^2 = 0$. As will be shown in subsequent sections, these auxiliary fields play an important role in accounting for the corrections of neuron interactions.

The data invariant notation, that is, assuming a certain uniformity over time for the input sequence x^l, of the propagator $G_{hh}(\tau)$ is then given as:

$$G_{hh}(\tau) = \frac{\kappa}{2} \xi_0 e^{-1|\tau|/\xi_0} + \xi_0^2 \sigma_{b,\text{eff}}^2. \tag{13}$$

where:

$$\sigma_{w,\text{eff}}^2 = \sigma_w^2 (1 - G_{hh}(0)). \tag{14}$$

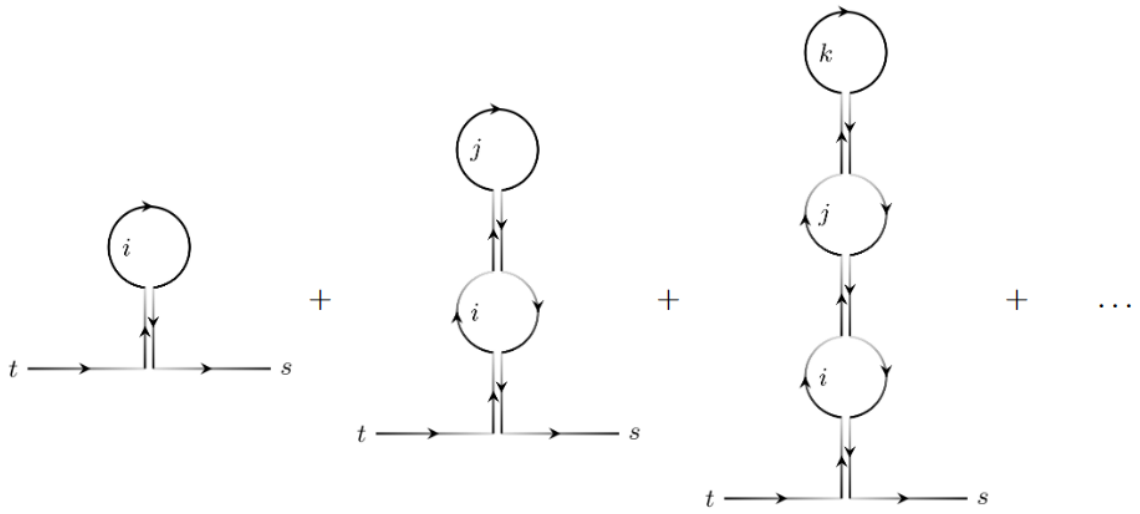


Figure 2: "The leading correction to the MFT result is due to an infinite series of $\mathcal{O}(1)$ cactus diagrams; here, we have illustrated those arising to first order in the expansion of tanh, i.e., $\phi(h) \approx h$. Each pair of vertices contributes a combined factor of 1/N, which is cancelled by a complementary factor of N from the sum over an internal neuron index, labelled by i, j, k." Solid edges represent the propagation of h to h, while shaded vertices represent propagation from h to $\tilde{z}$ for dark to light, $\tilde{z}$ to h vice versa. Source: [12]

$$\sigma^2_{b,\text{eff}} = \sigma^2_b + \sqrt{\frac{\sigma^2_u}{N}}\mathfrak{U}_0, \tag{15}$$

$$\xi_0 = \frac{1}{\sqrt{\gamma^2 - \sigma^2_{b,\text{eff}}}} \tag{16}$$

Equation 13 is the primary, so-called tree-level propagator. Sections 2.7.1 and 2.7.2 will present several corrections to this propagator. Presenting and discussing the derivations behind these corrections are outside of the scope of this paper, and shall therefore only mention the type of corrections formulated as context to the presented corrected correlator.

2.7.1 Linear Models

Grosvenor and Jefferson [12] first consider corrections for networks featuring a linear activation i.e. $\phi(h) = h$, as this allows for a simpler expansion of the propagator terms.

The first correction is presented in the form of 'infinite $\mathcal{O}(1)$ cacti', these cacti represent interactions between individual neurons in the network in between two layers of the network, specifically so that the path between interacting neurons is a simple cycle. These cacti, as shown in figure 2, are represented with a Feynman diagram, following the Feynman rules as presented in [12]. Using these Feynman rules a cactus diagram consisting of n loops can be expressed as:

$$\text{cactus}_n(\tau) = \sigma_w^{2n} c * (f * \bar{f})^{*n}(\tau). \tag{17}$$

Where $f = G_{h\tilde{z}}(\tau)$, $\bar{f} = G_{\tilde{z}h}(\tau)$ and $c = G_{hh}(\tau)$. This equation is then rewritten to use the Fourier transform for convenience, substituting τ for ω. Then, by taking the sum over n, all $\mathcal{O}(1)$ cacti are accounted for,

giving the total correction caused by all $\mathcal{O}(1)$ cacti in the network:

$$\sum_{n=0}^{\infty} \text{cactus}(\omega) = c(\omega)\frac{\omega^2+\gamma^2}{\omega^2+\gamma^2-\sigma_w^2} = X^{(0)}. \tag{18}$$

Following from the $\mathcal{O}(1)$ cacti, [12] formulate the contribution of $\mathcal{O}(T/N)$ mushrooms. Mushrooms differ from cacti in that their propagator may form closed loops along the neuron interactions, so-called "caps" [12]. The correction from the $\mathcal{O}(T/N)$ mushrooms is calculated with two contributions: The contribution from n-loop 1-Particle Irreducible (1PI) mushroom diagrams (mushrooms which have their cap somewhere along the vertices shown in figure 2, as well as diagrams that feature an arbitrary n-loop cactus preceded or followed by an n=1 mushroom (a mushroom type where the cap is directly between t or s and the stem). With $X^{(0)}$and $X^{(1)}$ denoting $\mathcal{O}(1)$ and $\mathcal{O}(T/N)$ contributions

$$X = X^{(0)} + \frac{\gamma T}{N} X^{(1)} \tag{19}$$

The $\mathcal{O}(1)$ contribution has previously been presented in equation 18, Grosvenor and Jefferson [12] present the $x^{(1)}$ contributions as:

$$X^{(1)}(\omega) = \frac{\sigma_w^2}{2\left[1-\sigma_w^2 f(\omega)\bar{f}(\omega)\right]} X^{(0)} = \frac{\sigma_w^2}{2} c(\omega)\frac{\omega^2+\gamma^2}{(\omega^2+\gamma^2-\sigma_w^2)^2}. \tag{20}$$

Performing the inverse Fourier transform on X then yields the corrections for positional space (i.e. neurons) as follows:

$$\begin{aligned} X(\tau) = \frac{\kappa}{4}\xi_0 e^{-|\tau|/\xi_0}\Bigg\{ & 1+\gamma^2\xi_0^2 + \xi_0\sigma_w^2|\tau| \\ & + \frac{T}{N}\frac{\gamma\sigma_w^2}{8}\xi_0\left[\xi_0\left(1+3\gamma^2\xi_0^2+\sigma_w^2\tau^2\right)+\left(1+3\gamma^2\xi_0^2\right)|\tau|\right]\Bigg\} \\ & + \frac{\gamma^2}{2}\xi_0^4\left(2+\gamma\frac{T}{N}\xi_0^2\sigma_w^2\right)\sigma_{b,\text{eff}}^2 \end{aligned} \tag{21}$$

There exists an issue in that τ is part of polynomial and exponential expressions, which complicates reasoning about the correlation length as the distance τ changes. In an effort to further simplify reasoning about the propagator, further derivations such that the parameter τ is only present in the exponential term $e^{-|\tau|/\xi}$, the loop-corrected propagator then is presented as:

$$X(\tau) \approx \frac{\kappa}{2}\xi e^{-|\tau|/\xi} + \frac{\gamma^2}{2}\xi_0^4\left(2+\gamma\frac{T}{N}\xi_0^2\sigma_w^2\right)\sigma_{b,\text{eff}}^2. \tag{22}$$

with the loop corrected correlation length:

$$\xi = \frac{\xi_0}{2}\left[1+\gamma^2\xi_0^2+\frac{\gamma T}{8N}\left(1+3\gamma^2\xi_0^2\right)\xi_0^2\sigma_w^2\right]. \tag{23}$$

And ξ_0 remains unchanged from equation 16. Any corrections attributed to the effects of $\sigma_{b,\text{eff}}^2$ have been ignored, as the contribution of this effective bias variance is expected to vanish as $\tau \to \infty$ [12]. This is a reasonable compromise as we are mostly interested in the correlations across large τ, as this is where issues with trainability arise.

2.7.2 Non-Linear models

In order to account for nonlinear models, an additional term is considered in the perturbative expansion of G_{hh}. Due to this, an infinite number of Feynman diagrams can contribute at all orders T^m/N^n with $0 < m < n$ [12]. However, Grosvenor and Jefferson [12] limit the corrections to diagrams that have $m = n = 1$. This introduces a new type of interaction that previously did not occur in the linear model: 5-point interactions, which feature 3, instead of 1, connections to the auxiliary field $\tilde{z}$.

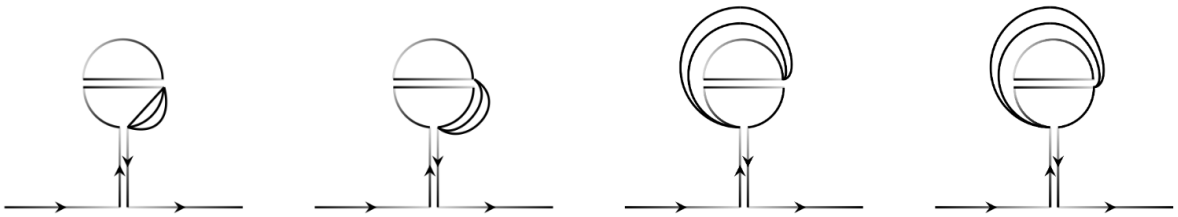


Figure 3: Possible mushroom cap diagrams due to the inclusion of 5-pt interactions. Source: [12]

The important topological addition this causes is the introduction of "petals" (closed hh propagators) and "branches" (self-similar cacti with $n > 1$ loops) [12]. This combination of closed hh propagators and self-similar cacti allows the same recursive formulation to be used as in section 2.7.1:

$$Y = Y^{(0)} + \frac{\gamma T}{N} Y^{(1)}. \tag{24}$$

Grosvenor and Jefferson [12] derive the following relations for this correction:

$$Y^{(0)}(\omega) = c(\omega) + \hat{\sigma}_w^2 f(\omega)\bar{f}(\omega)Y^{(0)}(\omega). \tag{25}$$

Solving for $Y^{(0)}$ and simplifying to:

$$Y^{(0)}(\omega) = c(\omega)\frac{\omega^2 + \gamma^2}{\omega^2 + \gamma^2 - \hat{\sigma}_w^2}. \tag{26}$$

where:

$$\hat{\sigma}_w^2 = \sigma_w^2(1 - 2Y^{(0)}). \tag{27}$$

Branching mushrooms remain similar to the mushrooms in the linear model in how the cap forms a closed loop along the cactus, the 5-pt interactions allow for several new caps to exist, pictured in figure 3. These additional caps each have correction relations attached to them in the 1PI form, alongside the permutations for non-1PI this produces finally the loop-corrected propagator:

$$\begin{aligned} Y(\omega) = \frac{f\bar{f}}{1 - \hat{\sigma}_w^2 f\bar{f}} \Bigg\{ \frac{c}{f\bar{f}} + \frac{\gamma T}{N} Y^{(0)} \left(\frac{\hat{\sigma}^2}{2} - 2\sigma_w^2 Y_0^{(1)} - \frac{\sigma_w^2}{2\gamma f\bar{f}} Y^{(0)} \right) \\ + 4\frac{\gamma T}{N}\sigma_w^4 \frac{f}{\bar{f}} \left[\frac{1}{3} Y^{(0)*3} + 2b_0^2 Y^{(0)} + \frac{1}{f\bar{f}} Y^{(0)} \left(Y^{(0)} * Y^{(0)} * (f\bar{f}) \right) \right] \Bigg\} \end{aligned} \tag{28}$$

Where ω dependence has been omitted to the rhs of the equation for compactness. Grosvenor and Jefferson [12] make several simplifications beyond this for compactness; however, the observant reader may notice the remaining presence of the Fourier transforms present in equation 28 compared to equation 22. This has been done as inverting the Fourier transform yields an expression that is too big to discuss, instead, the curious reader is encouraged to take a look at the work of Grosvenor and Jefferson [12] for the expanded notation.

3 Experimental setup

The theoretical formulations presented in section 2.7 encompass both Recurrent Neural Networks (RNNs) and Feedforward Neural Networks (FFNNs) with identical weights across hidden layers; therefore, as a matter of convenience, the overall setup involves performing all experiments on an RNN architecture. The RNN architecture offers advantages in ensuring network weights remain identical across each layer, as required by the theoretical predictions for an FFNN.

3.1 Observations of signal propagation

Section 2 has presented specifically two ways to observe signal propagation through the constructed networks. We will observe the networks through two correlation measures:

First, the correlation coefficient from Poole et al. [9], empirically calculated as follows: Given a pair of inputs x^a, x^b we look at the Pearson correlation coefficient between states $h^{l,a}$, $h^{l,b}$ for each l, as a shorthand notation, we shall refer to this correlation as the "paired-correlation."

The propagator $G_{hh}(\tau)$ has been formulated as a measure of the correlation over time. As we are interested in the correlation of the value h^l for a range of l, specifically betweenh^0 and a time-lagged h^τ, we observe the autocorrelation on the hidden states h^l over time. We will subsequently refer to the autocorrelation on a single input as the "single-correlation."

The observations on the single correlations of our setup are complicated by the existence of transient chaos in our network. As Lai and Tél [18] explain, the difference between sustained chaos and transient chaos is the average lifetime of the chaotic signal. For sustained chaos, the average lifetime is infinite, whereas transient chaos has a finite lifetime, which means that, during an observation, a crossover point may be reached where the chaotic signal (suddenly) dies out. It is mentioned in [18] how this issue can be avoided by observing a signal in its asymptotic properties of the signal, but they also debate that richer dynamics may be observed by following the signal from the beginning. As we are computationally constrained to finite hidden state sequences in our network, as well as our desire to observe the decay of non-chaotic signals over time, we want to identify the point in our hidden state sequences where this transient chaos ceases to exist. In order to capture as much of the remaining correlation dynamics as possible. Verzelli et al. [19] run into a similar issue of transient isolation where they observe measures of hamming distance, energy, activity, and entropy of the state of their network to observe when the effects of their transient chaos have died out.

Two fairly reliable indicators in detecting the dropoff in transient chaos for our network appear to be the hidden state average and variance over time. Both show asymptotic (or, rather, predictable) behaviour within the bounds of the network depths explored.

3.2 Datasets and input vectors

The datasets chosen to observe signal propagation on are MNIST [1] and a grayscale version of CIFAR-10 [20], the justification for these datasets being that both datasets are presented in a significant amount of previous works, allowing for comparison to these results and a stepping stone for experiments conducted. Furthermore, MNIST and CIFAR-10 differ in the expression of their images, i.e. MNIST has a relatively higher variance and a lower average of pixel values compared to the CIFAR-10 dataset. This difference should allow for a rudimentary observation of whether dataset invariance exists w.r.t. signal propagation.

As we are interested in exploring signal propagation in both RNN and FFNN settings, these datasets must be transformed accordingly. As mentioned in section **??**, the network architecture uses an RNN architecture, and FFNN behaviour is emulated by transforming the dataset accordingly. To emulate FFNN behaviour, the entire image is fed as a 1D vector, x_l, where $l = 0$, and subsequent input vectors at $l = 1, ..., L$ as all-zero vectors. This should allow the hidden state to represent the evolution of the input signal without interference from subsequent inputs, therefore strictly behaving as an FFNN.

For the RNN setup, images are split along the network depth. For a 28×28 image and network depth $L = 196$ this means the image will be flattened and split into 196 vectors of 4 pixels, with each vector being the input at a new timestep/layer. For network depths that do not yield a whole number of pixels, a number of pixels equal to the remainder is first removed from the end of the flattened image vector before splitting into timesteps.

3.3 Initial hyperparameters

Initial hyperparameters differ based on the activation function used in the networks:

The initial choice of hyperparameters for the linear networks is based on derivations from equation 22, which predicts that $X(\tau)$ predicts an asymptotic decrease in decay as $\sigma_w \to 1$, this will therefore be the main point of analysis. To maintain unity with the non-linear setups, the remaining parameters have been set as $\sigma_v = 0.025$, $\sigma_b = 10^{-10}$, and $N = 1024$. The choice for these latter parameters will be explained through the scope of the non-linear networks in reference to work by Chen et al. [11].

The initial choice of hyperparameters for the non-linear networks is derived from one of the setups presented by Chen et al. [11] and forms the stepping stone for further network parameter permutations. Chen et al. [11] observe high trainability of their network across depth for $\sigma_w = 1.1$, implying that a network initialised using this weight deviation should show critical behaviour. Experiments aiming to verify this setup have found that this is indeed the case for our networks. However, an ad hoc neighbourhood search of surrounding σ_w has shown that our networks show behaviour closer to criticality around $\sigma_w = 1.05$.

Therefore, the initial non-linear setup in question consists of a network with weight initialization with $\sigma_w \in [0.5, 1.05, 1.5]$, fixed $\sigma_v = 0.025$ and fixed $\sigma_b = 1 \times 10^{-10}$. The σ_b value deviates from the setup by Chen et al. [11] has been chosen in such a way as to minimize the effects of bias in observations of chaotic/ordered behaviour but allow the network's paired correlations to converge[3] due to the high bias regime, as presented by Poole et al. [9]. We set the bias at such a low value in order to minimize the effects of σ_b as much as possible, such that it will be apparent that the single-correlation observations are purely related to the dynamics of the input and hidden state transitions, and not from convergence to an ordered state.

Preliminary experiments that trained the network have yielded accuracies relatively similar to Chen et al.'s results (Appendix A) for a network of width $N = 1024$ neurons (due to computational constraints), and depth of $L = 200$. For further experiments, the network depth has been set to $L = 1000$, as this will allow for higher fidelity in observing layer-to-layer network behaviour. Experiments featuring other network depths for FFNNs have been omitted as early experiments have shown that the first-to-last layer network dynamics scale uniformly between shallower and deeper networks due to the normalisation of hidden state weights over network depths.

3.4 Additional considerations

Firstly, as presented in Zhang et al. [21], a network is able to transition between regimes as a result of training, in fact, they mention that the optimal epoch occurs at the critical border. This transitioning of regimes makes intuitive sense as the entire goal of training is to make the network weights conform to the training data in pursuit of a certain performance criterion. In order to exert more control over the network regime we observe, we forego network training entirely. Hereby eliminating any issues that may arise from non-determinism introduced by the network optimizer.

Secondly, due to the nature of the critical regime, it can be expected that slight permutations in starting conditions (e.g. the network weight distributions) can lead to large differences in observed behaviour. Preliminary experiments show that repeated trials tend to produce outliers that show different network dynamics, primarily related to a longer lifetime of transient chaos in the network, compared to the dynamics of the other trials. In an effort to reduce the influence of such outliers, the experiment of a certain network parameter has been repeated in 5 trials unless otherwise specified. No outlier detection has been attempted to exclude such trials from the analysis. The reasoning for this is that the outlier behaviour originates primarily from the phase of transient chaos in the network, causing the observed hidden state average and variance to deviate from other observed trials significantly. Once this transient has died down, the outlier network starts to exhibit similar behaviour as other trials w.r.t. the evolution of hidden state averages, variances, and correlation decay.

4 Experimental Results

The experimental results are presented in subsections to differentiate between the specific setup and goal of the experiments, section 4.1 considers all experiments related to the predictions of linear networks, while sections 4.2 trough 4.4 consider experiments related to nonlinear networks. Additionally, section 4.5 presents additional observations not directly related to the goals of the individual experiments.

[3] On a more practical note: this is due to networks in the ordered regime having their hidden state decay to an all-zero vector. Such a vector has a variance of 0, leading to undefined results in calculating the Pearson correlation. This is circumvented by having the hidden state decay to the bias vector, which has a nonzero variance.

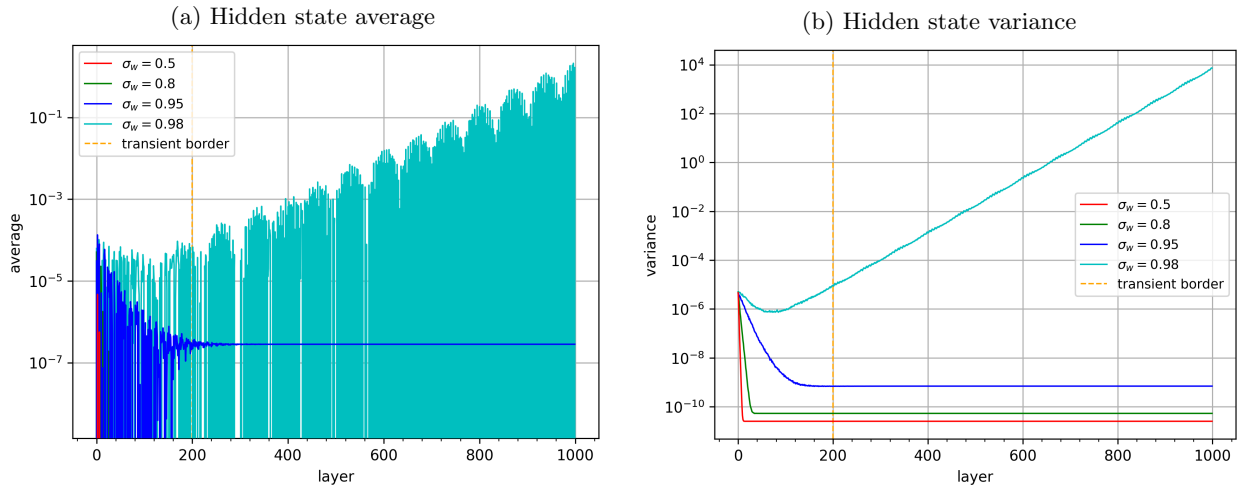


Figure 4: Hidden state average (a) and variance (b), for a linear FFNN propagating MNIST data, both plots show the border beyond which no transient chaos is assumed to persist.

4.1 Linear Networks

The first corrections from the NN-QFT correspondence relate to linear networks, and these corrections are understood to be theoretically solved. The goal of this experiment is to observe the empirical behaviour of the signal propagation and relate it to the theoretical predictions. This section presents the two architectures and three networks: an identical-weight FFNN propagating MNIST data and two RNNs propagating MNIST and CIFAR10 data. We have the following linear activation functions for these linear networks: $\phi(x) = \varphi(x) = x$.

Figure 4 shows the hidden state averages and variances across layer depth, plotted with a logarithmic y scale as the values for $\sigma_w = 0.98$ continue to grow exponentially, the transient border has been set at $l = 200$, under the considerations presented in section 3.1. Figure 5a shows the resulting single correlations. Performing Detrended Fluctuation Analysis (DFA) [22] on individual networks shows us that the higher-weight networks show strong anti-correlated behaviour ($\alpha = 0.113$ for $\sigma_w = 0.95$ and $\alpha = 0.008$ for $\sigma_w = 0.98$), this is desirable as lower anti-correlated values, according to Hardstone et al. [22] mean a low displacement of the correlation sequence, which is in turn linked to stronger propagation of the signal. We are specifically interested in the absolute autocorrelation values, as strong anticorrelated hidden states are an indication that the network is still propagating the input signal through the lagged states. For this reason, and to ease the interpretation of these correlations, we plot the envelope of the absolute values of the single correlations in figure 5b, it is important to note that this is particularly coarse envelope as we are mainly interested in observing the decay of the highest correlation values across a wide timeframe. Additionally, figure 5b displays the predicted correlation decay associated with each network. As can be seen from these plots, the lower weight networks of $\sigma_w = 0.5$ and $\sigma_w = 0.8$ show agreement with theoretical predictions. The $\sigma_w = 0.95$ network shows interesting behaviour, where the empirical decay curve decays significantly later than that of the theoretical prediction. Two theories exist for what causes this: The curvature of this decay seems to visually match the isolated contributions of loop corrections, obtained by taking $X(\tau) - G_{hh}(\tau)$. Specifically, the $\mathcal{O}(T/N)$ correction appears to exert too little contribution to the predicted correlation. Due to this similarity, it is hypothesised that the loop corrections do not contribute a significant enough magnitude to offset the exponential decay curve of the tree-level propagator. Alternatively, as the network initialisation lies close to the critical point, signals will at first propagate the network without de-correlating until an inflexion point is reached where the network transitions turn ordered, and the correlation decreases at a similar rate that the theory predicts. This latter behaviour is more likely as it is typically observed for systems that are on the edge of turning critical, such as previously observed in e.g. [13]. The lack of significant decay in correlation in the $\sigma_w = 0.98$ network indicates that we are approaching the point of criticality even closer. This point of criticality is predicted in theory to lie at $\sigma_w = 1.0$, in approaching this point empirically, we indeed observe an increased length of signal propagation with the correlation envelope no longer showing

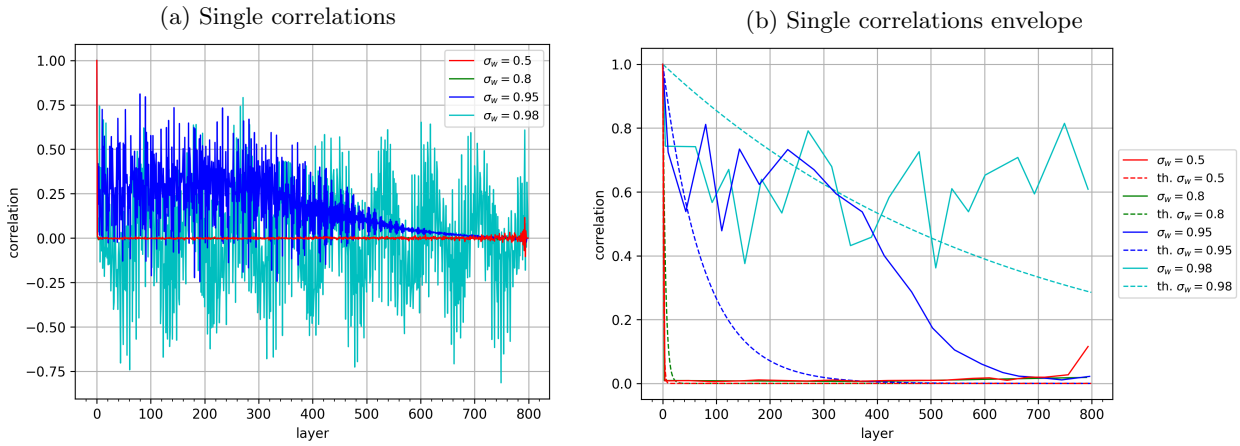


Figure 5: Single correlations (a) and the derived envelope of these correlations (b), for a linear FFNN propagating MNIST data, the envelope plots are complimented by the predicted decay curves of $G_{hh}(\tau)$ (dashed lines).

exponential decay but instead appearing to more closely resemble a power-law decay.

Moving to the linear RNN setup, we have a slightly shallower network depth due to the MNIST dataset shape. Figure 6 shows the identified transient border based on the hidden state average and variance, noteworthy is the earlier decay of transient chaos compared to the FFNN experiment. The observed periodicity among the state averages and variances is directly caused by an equal periodicity in the values of the dataset, which has been confirmed by identifying significant spikes of equal frequency when observing spectrograms of each sequence (dataset and single correlations).

Figure 7 shows the correlations associated with a linear-activation RNN, the predicted decay curves under the assumption that $\mathfrak{U}_0 = 0$ (i.e. an identical curve to the FFNN prediction, for a slightly shallower network). In comparison to the FFNN setup, we can see a slower decay of the correlation envelope for the lowest weight networks $\sigma_w = 0.5$ and $\sigma_w = 0.8$, a relatively unchanged decay for $\sigma_w = 0.95$, with the additional observation that its decay starts without the lag observed in 5 and a significantly stronger decay for $\sigma_w = 0.98$, all of these decay curves appear to contract around the theoretical decay curve associated with $\sigma_w = 0.95$. This contraction of empirical decay curves is caused by the interaction of the network's hidden state and the sequentially fed data: The two lowest-weight networks have their decay slowed, as the continuous addition of additional data causes correlation to rise above what it would be if it were to decay in the FFNN setting. While the higher-weight network of $\sigma_w = 0.98$ retains a marginally higher correlation, likely due to the relatively smaller influence on h^{l+1} by the input weights W compared to the hidden state weights U, but the correlation with the initial state is still significantly affected by the sequential inputs. The $\sigma_w = 0.95$ network appears to follow a relatively unchanged decay, which may be for similar reasons as those of the $\sigma_w = 0.98$ network.

CIFAR10 offers a more uniform dataset in its average input sequence, which may explain the mismatch in correlation decay rate for the MNIST RNN between prediction and observation. The plots showing the hidden state average and variance have been omitted as they were deemed not to be able to provide any meaningful information and have been omitted. The transients appear to have died out around $l = 100$, similar to the MNIST RNN.

Single correlations shown in figure 8a behave significantly differently from the MNIST dataset, with surprisingly a more pronounced periodicity of single correlations for all networks. This periodicity appears to align with the unflattened width of a CIFAR10 image, suggesting that a common element of these images (e.g. image border) 'boosts' the correlations. Figure 8b shows us that the three lower-weight networks converge to a similar correlation roughly along the predicted decay curve of the $\sigma_w = 0.95$ network, and appear to converge on a correlation of ≈ 0.08, which would be in accordance with the expected increase in effective bias present due to the non-zero x^l inputs. The $\sigma_w = 0.98$ network appears to show similar

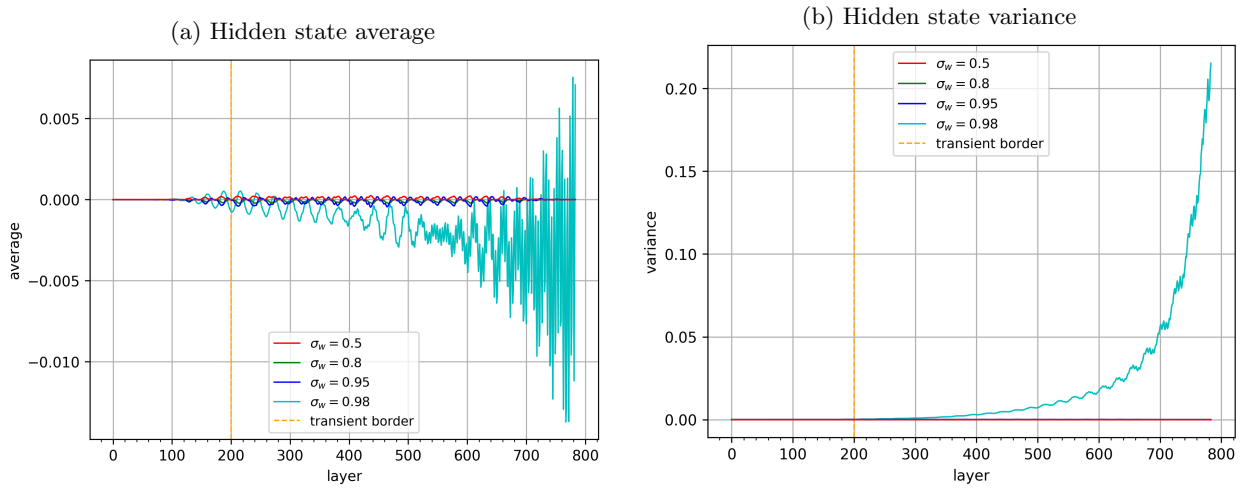


Figure 6: Hidden state average (a) and variance (b), for linear RNNs propagating MNIST data, both showing the border beyond which no transient chaos is assumed to persist.

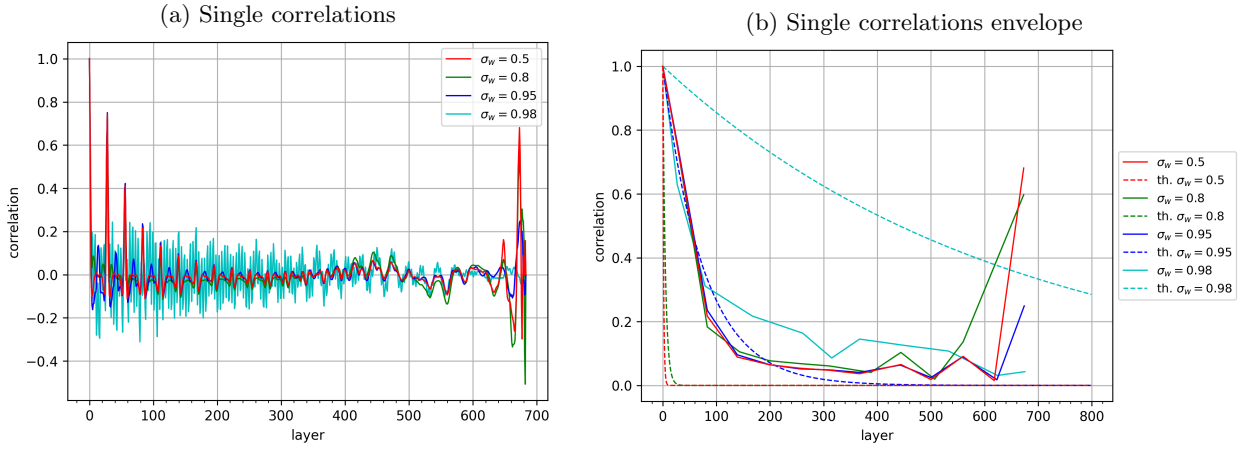


Figure 7: Single correlations (a) and the derived envelope of these correlations (b), for linear RNNs propagating MNIST data, the envelope plots are complimented by two plots of the predicted decay curve.

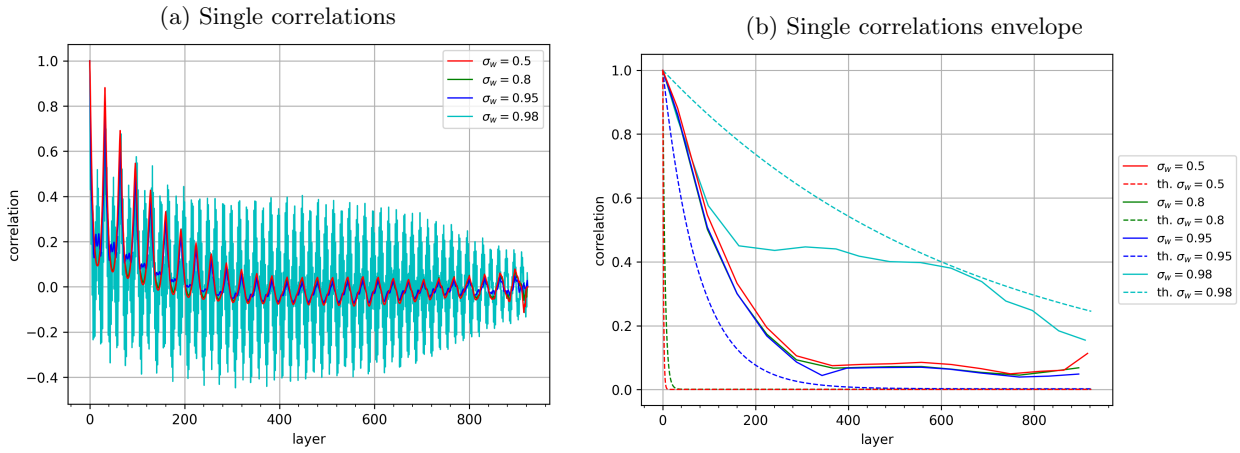


Figure 8: Single correlations (a) and the derived envelope of these correlations (b), for a linear RNN propagating CIFAR10 data, the envelope plots are complimented by two plots of the predicted decay curve.

behaviour to the $\sigma_w = 0.95$ FFNN network, again be a sign that these networks are nearing criticality but are still too subcritical to sustain the signal propagation for depths $\geq L$. As we are seeing this behaviour for a higher value of σ_w in the RNN compared to the FFNN, this may suggest that the critical point lies further from the RNN configuration using $\sigma_w = 0.98$ than it is for the FFNN using a similar weight. This would imply that the critical point is not data invariant for RNN setups, and may also provide an alternative theory for reduced signal propagation in the MNIST RNN for the higher-weight configurations.

Overall observations w.r.t. the change in correlation decay for RNN configurations is a possible indication that $\sigma^2_{b,\text{eff}}$ doesn't only affect the correlation value towards which the $G_{hh}(\tau)$ asymptotically converges, but the rate of decay as well, as seen in figures 7 and 8.

4.2 Identical-weight FFNN

The first nonlinear setup presented features a set of networks in the ordered, chaotic and critical regimes. Theoretical predictions of $Y(\tau)$ (which are computationally derived from $Y(\omega)$) suggest asymptotic behaviour of the propagator to occur at $\sigma_w = 0.55$. Empirical observations show us, however, that this is not the critical point for the considered network setups. We see single-correlations continue to decay for values $\sigma_w < 1.0$, as figure9 shows. Additionally, what can be seen in the plotted envelope is the transition of exponential decay of the correlation for the $\sigma_w = 0.9$ and $\sigma_w = 0.93$ networks to decay resembling a power-law in $\sigma_w = 0.96$ and $\sigma_w = 1.0$.

As the exact dynamics of $Y(\tau)$ are not fully understood, we shift away from relating observed correlation behaviour to individual components of the theoretical (nonlinear) prediction. Instead, we focus on verifying if the general qualities of the NN-QFT correspondence hold in empirical observations.

4.2.1 MNIST Exploratory Setup

The exploratory setup forms the baseline of comparison against other setups. For the nonlinear networks, we additionally observe the paired correlations as explained in section 3.1. Figure 10a shows the evolution of the paired correlation over time, as can be seen, the ordered network lies far into the ordered regime, with hidden states decaying to a common state within ~ 20 layers. This decay is represented in the $\mathcal{C} - map$ of figure 10bby corresponding dots representing the discrete values of χ_1, which can be traced to form an arc that lies above the critical line (the grey diagonal), implying $\chi_1 < 1$, with most dots concentrating around $c = 1$, i.e. the network's fixed exists as $c^* = 1$. From these observations and their parallels to Poole et al.'s [9] theory, we can indeed conclude that the network is in the ordered phase. The chaotic network reaches maximal de-correlation of hidden states around 200 deep into the network. It is observed that $\chi_1 > 1$ and $c^* \approx 0$. The

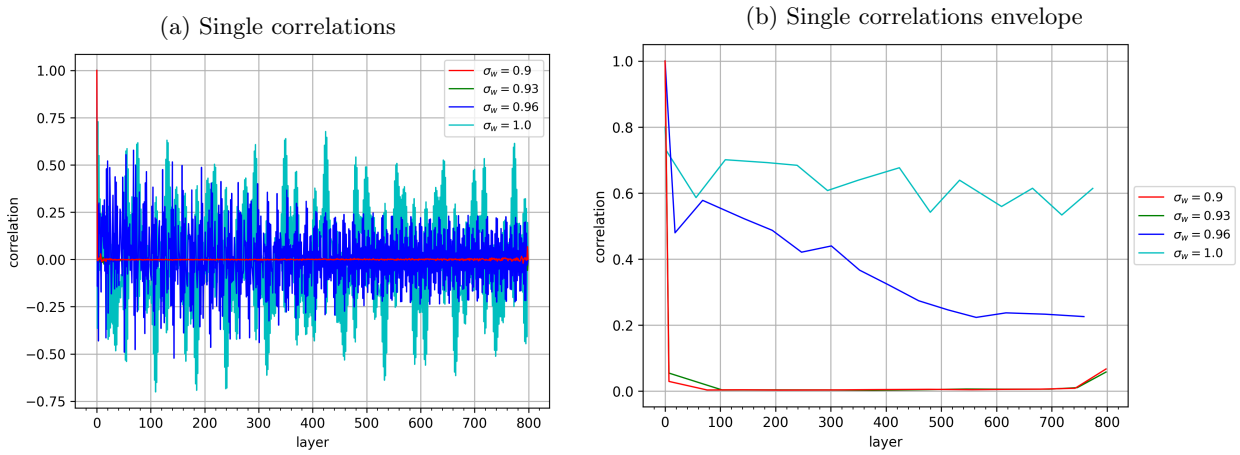


Figure 9: Decay curves for networks approaching the critical point. Single correlations (a) and the derived envelope of these correlations (b) for a nonlinear FFNN propagating MNIST data.

critical network also behaves, generally[4], in agreement with the theoretical predictions of a critical network, it has a slope of $\chi_1 \approx 1$, and its fixed point appears as $c^* \approx 0.39$. Having confirmed that our own empirical results based on [11] agree with the theoretical predictions presented by Poole et al. [9], we will now use the paired correlation plots as shown as a point of reference when discussing the regime of other network configurations, and as a cross-reference for the observed behaviour in the single-correlation plots.

As mentioned in section 3.1 and previously done for the linear networks, we again identify an interval of the network's hidden states where signals stemming from transient chaos are no longer detected. Most transient behaviour of the linear networks' hidden state average and variance had subsided around layer 100. Based on the shown hidden state average and variance plots it may appear that the transient has already died out around ≈ 100 layers, consistent with the linear networks. However, the observed transients for individual nonlinear networks appear to persist to a greater depth, up to ≈ 200 layers. For setting the transient border, we simply adhere to setting the border at the end of the longest persisting transient chaos.

Taking the autocorrelation over the hidden states to produce the single correlation plot, as shown in figure 12a, we see some interesting behaviour. One thing this figure shows (and which figure 11 also showed signs of) is the presence of clear periodic behaviour in the propagating signal along the layers of the network expressed in the periodic nature of the correlation. DFA of individual networks shows that the critical ($\sigma_w = 1.05$) networks' correlations over time are anti-correlated with $\alpha \approx 0.02$. For the ordered and chaotic correlations, these values are $\alpha \approx 0.58$ and $\alpha \approx 0.19$, respectively. The high α for the ordered networks, indicating the presence of white noise, is almost certainly caused by machine precision errors occurring during calculations related to the hidden state values. The hidden state values converge towards an average order of magnitude of 10^{-7}, which is the same order of magnitude of the machine epsilon (2^{-23}) of the 32-bit floating point numbers used and would therefore be highly susceptible to rounding errors in calculation. The chaotic network features a weaker anti-correlated relation compared to the critical network, which could be due to an increased contribution of white noise due to the chaotic nature of the network's state transitions.

Ignoring the anti-correlated oscillations, the critical network appears to have an additional periodicity of around ≈ 75 layers. The chaotic network appears to show oscillations in correlation across a period of ≈ 40 layers, based on spectral analysis.

In figure 12b, the corresponding envelope for this correlation sequence is shown, with additionally the predicted correlation of $Y(\tau)$. For the ordered network, the prediction appears relatively accurate, as we essentially have a negligible $\sigma^2_{b,\text{eff}}$ due to the network configuration, the theoretical prediction for chaotic networks is approximately equal to the prediction of the critical network, suggesting that the expected de-correlation over time for higher σ_w networks is underrepresented in the $Y(\tau)$. Additionally, there is a mismatch between the observed and predicted decay of the correlation for the critical network (and by proxy,

[4]To be discussed later.

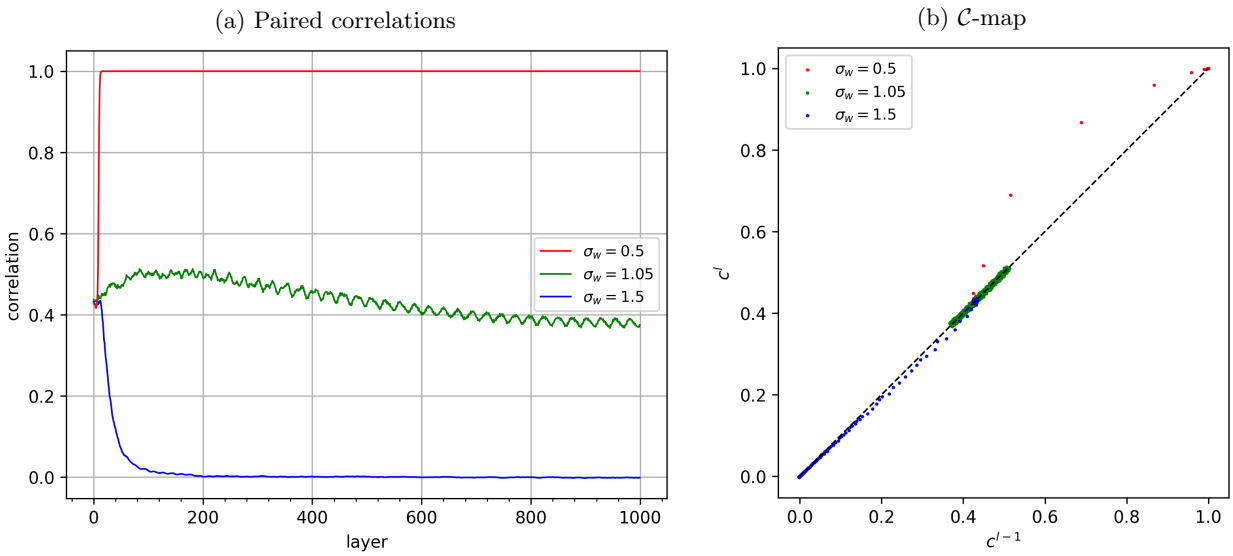


Figure 10: Paired correlation evolution across network layers (a), and the derived $\mathcal{C}$-map (b), of FFNNs propagating MNIST data, for 3 network setups in the ordered ($\sigma_w = 0.5$), critical ($\sigma_w = 1.05$), and chaotic ($\sigma_w = 1.5$) regimes

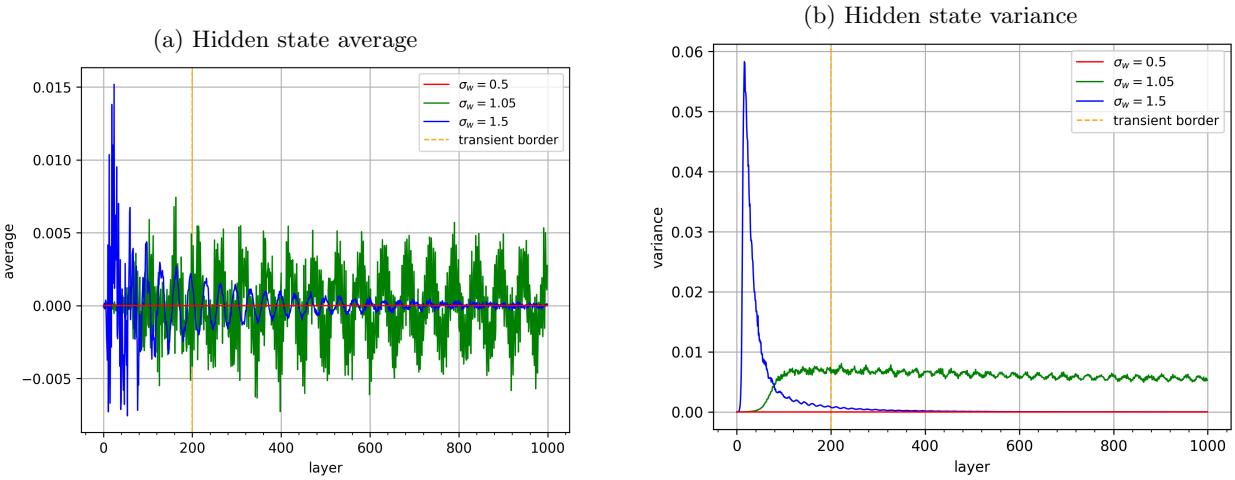


Figure 11: Hidden state average (a) and Hidden state variance (b), of FFNNs propagating MNIST data, both showing the border beyond which no transient chaos is assumed to persist.

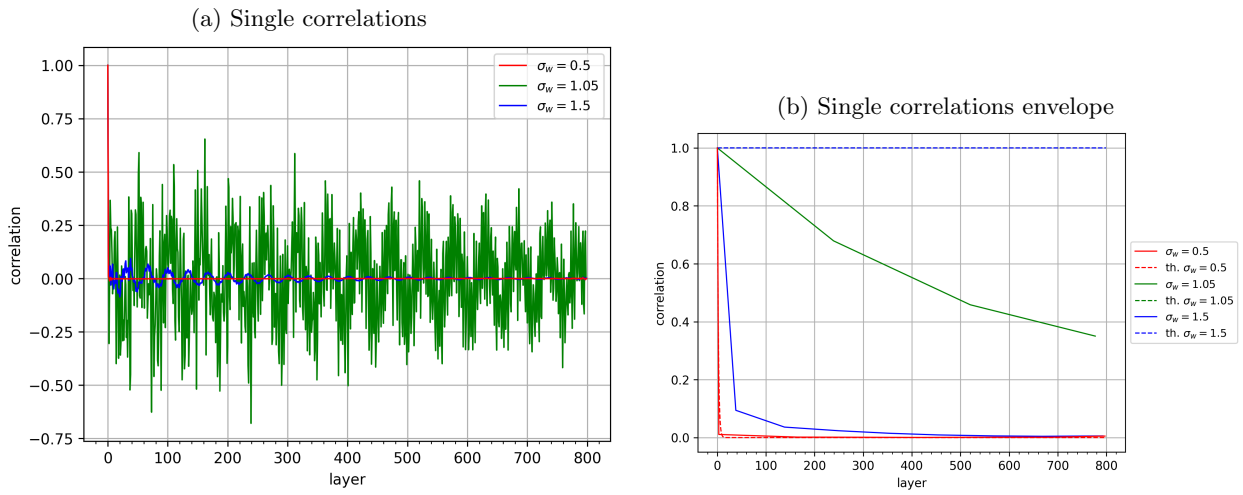


Figure 12: Single correlations (a) and the derived envelope of these correlations (b), of FFNNs propagating MNIST data, the envelope plots are followed by two plots of the fitted decay curve.

the chaotic network) this is due to both networks lying beyond the point where the theoretical prediction diverges. As these predictions remain similar for all values $\sigma_w \geq 0.55$ this would mean we have to limit observations of nonlinear networks to only the very ordered regime. For this reason, and the motivation to additionally explore signal propagation in the critical and ordered regimes, we distance ourselves from any further direct comparison to the theoretical predictions. Instead, we focus on several factors that have been identified in theory that affect signal propagation and observe their empirical behaviour.

4.2.2 CIFAR-10 Comparison

The CIFAR10 dataset is introduced to observe if the behaviour observed on the CIFAR10 dataset remains consistent with the previously observed MNIST dataset. Figures 13 and 15 feature equivalent plots as presented in section 4.2, and we will therefore primarily focus on observed similarities and/or differences compared to the plots from this previous section.

We start by observing figure 13, as the figure shows, we have networks within a similar regime, the ordered hidden states rapidly converge, and chaotic states rapidly de-correlate, maximal de-correlation appears to occur sooner than in the previous setup using the MNIST dataset, whereas previously maximal de-correlation was reached around 200 layers deep, it now occurs around a depth of 100 layers and settles on a value of $c^* \approx 0.1$ instead of 0. The critical network is also observed to have a higher fixed point of $c^* \approx 0.81$. The $\mathcal{C}$-map shows that values of χ_1 remain relatively unchanged w.r.t. individual regimes.

Figure 14a shows relatively unchanged behaviour w.r.t. the duration of suspected transient chaos. However, it appears that the longer period oscillations in the average of the critical network states, as shown in figure 11a, behave more slowly in the CIFAR10 networks, with an approximately two times longer period. A similar increase in periodicity is observed in the hidden state variance over time, as can be seen in comparing figure 11b and figure 14b. The overall behaviour w.r.t. the network transients appear to remain similar.

This longer period of the hidden state averages and variances observed in figure 14 carries over to the single correlations sequence, as can be seen in figure 15a. Here, we observe an increase in the wavelength of these oscillations, compared to the MNIST network, of ≈ 75 layers to ≈ 200 layers. The derived envelope shown in figure 15b remains largely unchanged The sudden shortening of periodicity and amplitude after layer 700 is rather unexpected, as such behaviour hasn't been observed in other FFNN networks in such a fashion. The state average and variance graphs in figure 14 show no indication of the decrease in correlation in comparison to previous network layers. Instead, it is possible that this is an effect of the calculation of the (auto) correlation breaking down near the end of the sequence.

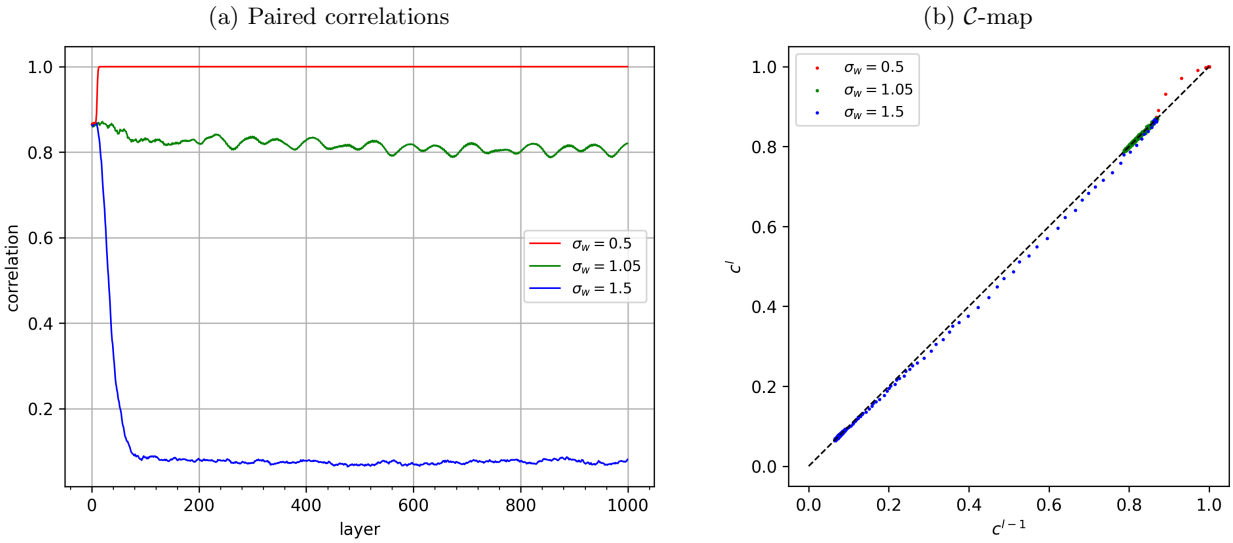


Figure 13: Paired correlation evolution across network layers (a), and the derived $\mathcal{C}$-map (b), of FFNNs propagating CIFAR10 data, for 3 network setups in the ordered ($\sigma_w = 0.5$), critical ($\sigma_w = 1.05$), and chaotic ($\sigma_w = 1.5$) regimes.

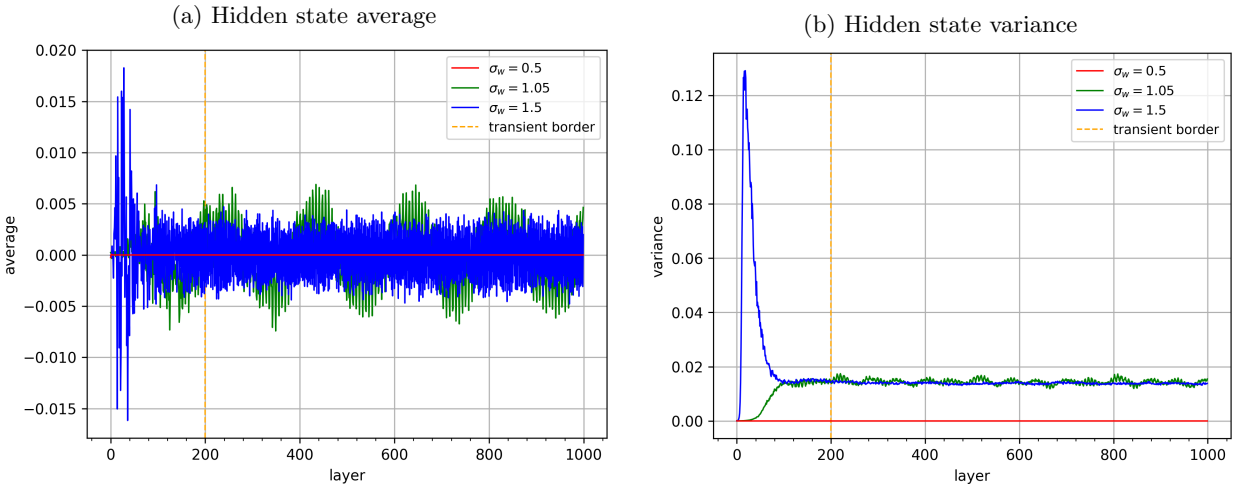


Figure 14: Hidden state average (a) and Hidden state variance (b), of FFNNs propagating CIFAR10 data, both showing the border beyond which no transient chaos is assumed to persist.

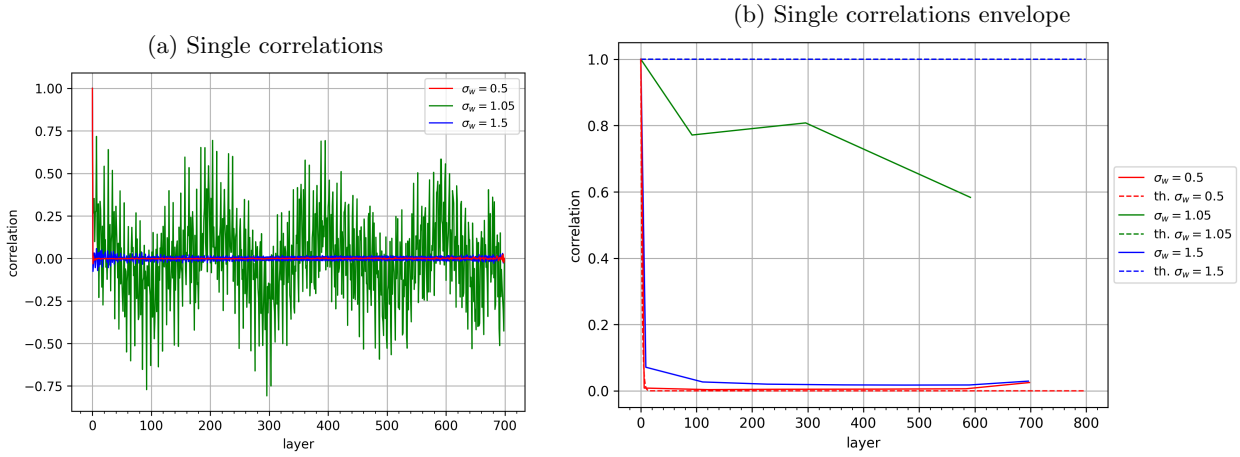


Figure 15: Single correlations (a) and the derived envelope of these correlations (b), of FFNNs propagating CIFAR10 data, the envelope plots are followed by two plots of the fitted decay curve.

4.2.3 Random Inputs

To further explore the discrepancies between the periodicity of the MNIST and CIFAR-10 networks, we construct two identical critical networks ($\sigma_w = 1.05$) and feed it input vectors drawn from a normal distribution, i.e. $x^0 \sim \mathcal{N}(\mu, \sigma)$. The first network is observed with varying values of μ with fixed $\sigma = 0.25$, while the second network features varying σ with fixed $\mu = 0.5$. Figures 16a and 16b show the resulting correlations associated with each setup. Not only do the individual experiments show nearly identical correlations across the network layers, cross referencing figures 16a and 16b show that all correlations across layers are nearly identical no matter the distribution of the input vector. This makes for a strong argument that the difference in single-correlation behaviour, as shown in figures 12a and 15a, is due to the individual differences between weight matrices giving rise to different propagation behaviour. Additionally, this would suggest that in reference to the NN-QFT correspondence, the assumption that signal propagation is dataset invariant appears to hold for an FFNN architecture.

4.3 Identical-weight FFNN with trailing noise

To check the robustness of the behaviour observed in the previous sections 4.2.1-4.2.3, we can replace the trailing all-zero vectors, x^l for $1 \leq l \leq L$, with noisy vectors in an effort to observe the network behaviour when the initial (unaltered) input vector x^0 is followed by noise, where x^0 are the flattened MNIST images.

In this setup, we re-use the network presented in section 4.2.1 and change the input data to be unaltered for x^0 and for the noise trailing the input vector we take $x^l = \mathcal{N}(0, 1.0)$ for $1 \leq l \leq L$ Due to the low value $\sigma_v = 0.025$ of the original network from section 4.2.1, the impact of the trailing noise may be relatively negligible in attempts to perturb the hidden state of the network, especially for higher values of σ_w. To observe this, an additional network is constructed, identical to our original setup, but with increased value for input weight deviation: $\sigma_v = 0.1$, in an effort to observe the effects of higher-impact trailing noise. We will refer to these setups as the 'low-impact noise' and 'high-impact noise' setups respectively.

Figure 17 shows the paired correlations for the (regular) noisy setup. As can be seen, the addition of noise causes the strong bias regime to no longer be present, as the contribution of the trailing noise is several orders of magnitude greater than the bias w.r.t. their respective weight distributions. This leads to nearly instant maximum de-correlation between input pairs for the ordered network, with the critical and chaotic networks de-correlating relatively quickly after.

Figure 19 compares the effect of noise on the network. Contrasting the paired correlations, the single correlations remain relatively unaffected by the added low-impact noise; both the single correlations and respective envelope show only a decrease in amplitude caused by the de-correlation of the hidden state by

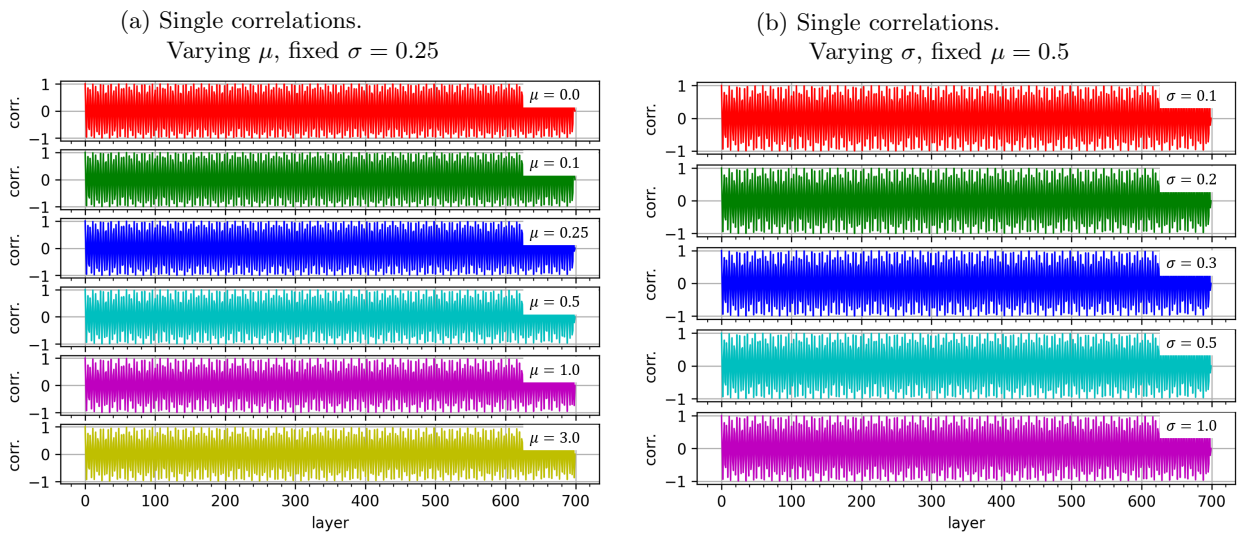


Figure 16: Single correlations for randomly drawn inputs with varying μ (a), and single correlations for randomly drawn inputs with varying σ(b).

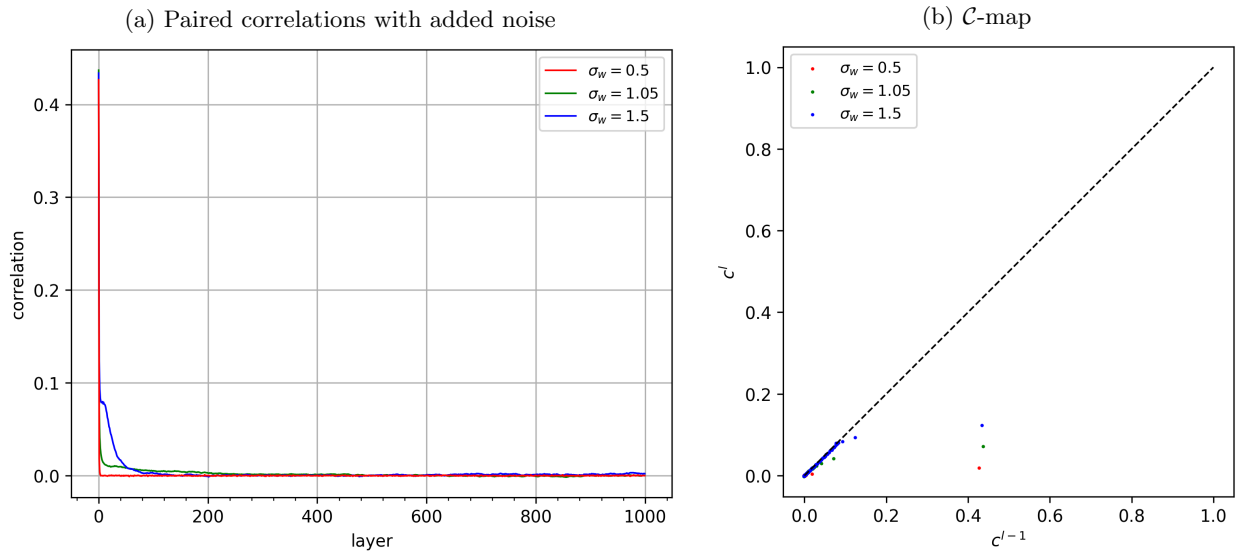


Figure 17: Paired correlations(a), and the derived $\mathcal{C}$-map (b), for a FFNN propagating MNIST data with trailing noise

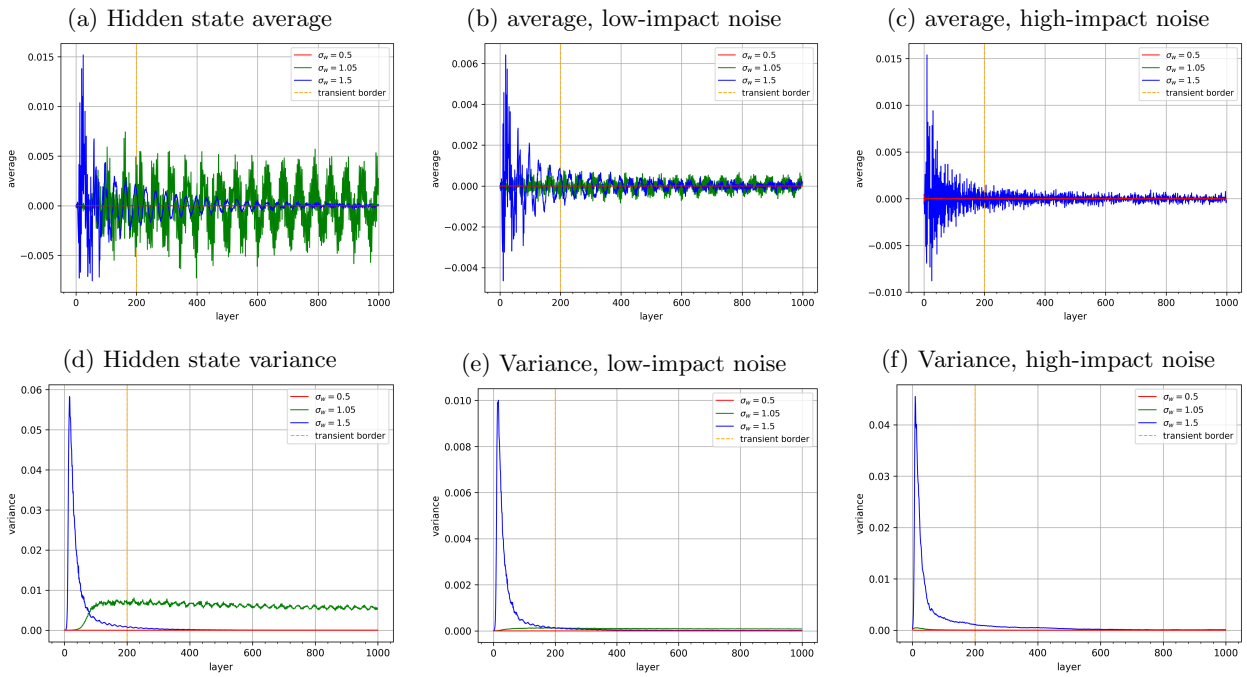


Figure 18: Hidden state averages (a-c) and Hidden state variances (d-f), including the transient borders for each, for 3 FFNNs propagating MNIST data with varying levels of trailing noise.

the added noise. The high-impact noise setup shows a stronger decrease in single-correlations across layers due to the greater effectiveness of noise perturbing the hidden state values in subsequent layers. This leads to maximum de-correlation of hidden states across all regimes within the observed layer depth. In reference to the NN-QFT correspondence, the addition of noise leads to an increase of $\sigma^2_{b,\text{eff}}$, as it is dependent on $\mathfrak{U}_0 = \sigma_u^2\langle\varphi(x_i(t_1))\varphi(x_i(t_2))\rangle$, which is predicted to lead to an increased asymptote for $G_{hh}(\tau)$. This is in agreement with the observations regarding the chaotic network, which appears to feature greater correlation values across the sequence in the high-impact noise network compared to the low- and no-noise networks, where the correlations appear to either converge towards ≈ 0.01, or 0 with a significantly longer tail. The deeply ordered network remains unaffected, suggesting that the noise magnitude is insufficient to overcome the decay caused by the hidden state weights.

4.4 Recurrent Neural Networks

Having observed that an increase in effective bias due to trailing non-zero data leads to a decrease in signal propagation, we now consider the following experiment Where the trailing noise setup of section 4.3 offered a statistically uniform input beyond the first initial input vector x^0, we now explore the setup where the original input vector from our dataset is fed to the network in a many-to-one style. That is, we split the flattened images into chunks that are fed at each layer to the network (not applicable for our setup, where eventually the final hidden state is read and used for classification of the input). In order to maintain consistency with our previous observations, we shall continue referring to the signals as propagating through *layers* instead of the academic distinction made for RNNs of propagation through *time*.

In the RNN setting, the Paired correlations become an inaccurate measure for identifying a network's chaos regime, as similar to the setup featuring trailing noise, the RNN setting causes the high bias regime required for the correlation plot to accurately show correlation and de-correlation of states, to be unsatisfied. Despite this, the paired correlations remain a useful observation for network behaviour. I.e. in figure 20, we can see how a change in the input vector average at a certain layer affects its paired correlation. The ordered network (that is, in reference to the ordered behaviour from previous sections: $\sigma_w = 0.5$) shows how hidden state de-correlation aligns strongly with the periodicity of the input data, which is due to states

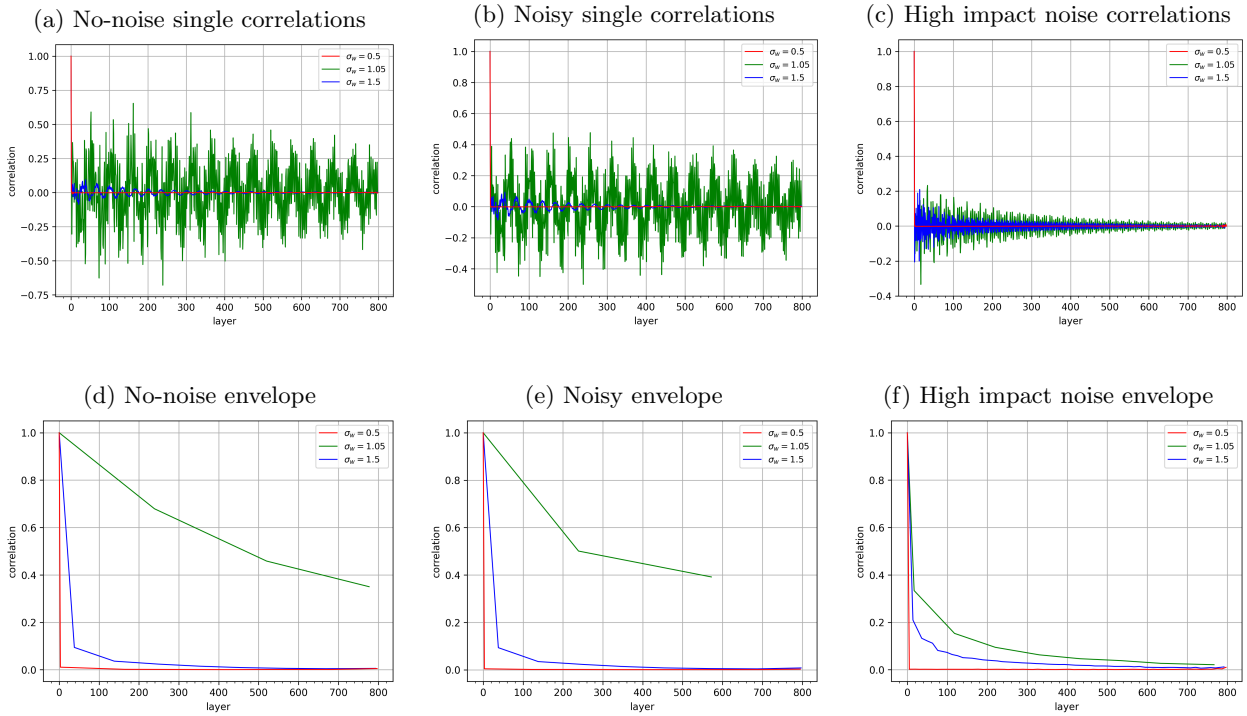


Figure 19: Single correlation evolution across varying levels of noise (a-c), and their respective envelopes (d-f), for 3 FFNNs propagating MNIST data with varying levels of trailing noise.

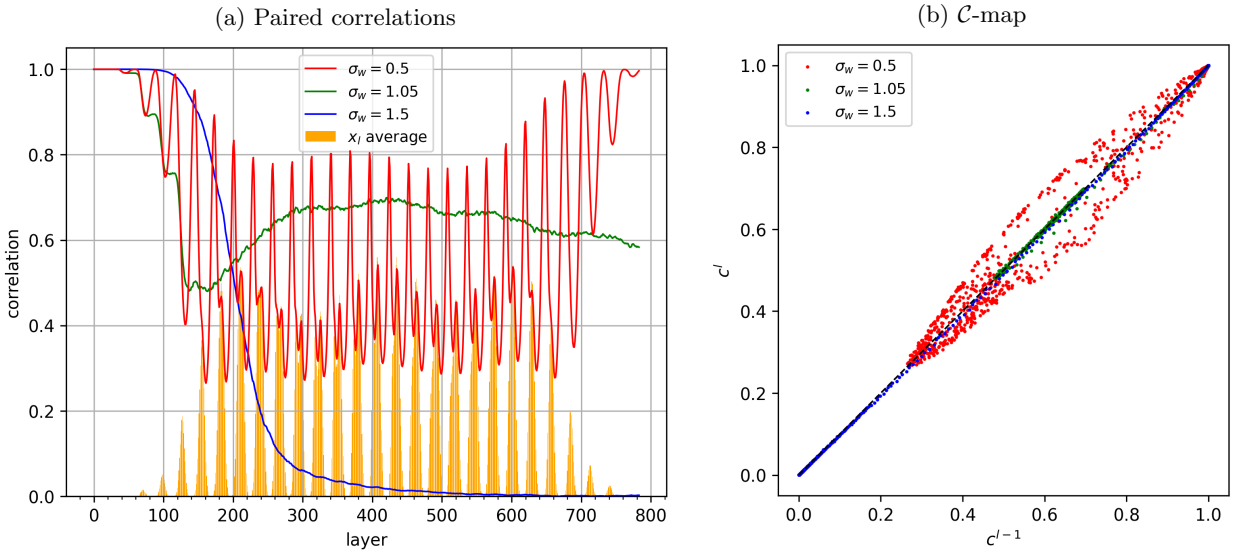
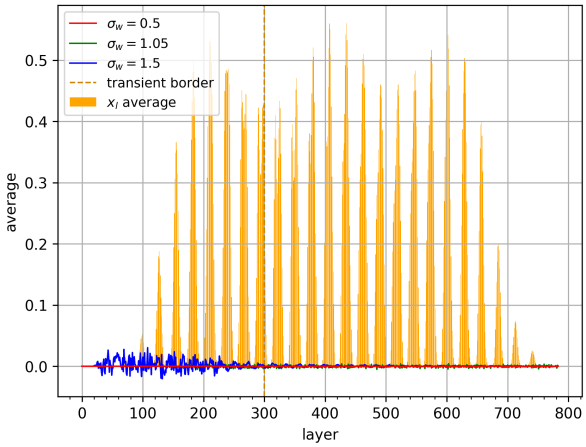


Figure 20: Paired correlation evolution across network layers, for 3 RNN setups propagating MNIST data in 1-pixel iterations.

(a) Hidden state average



(b) Hidden state variance

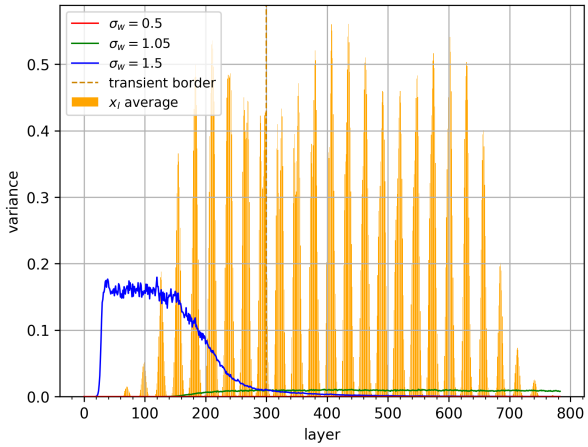


Figure 21: Hidden state average (a) and Hidden state variance (b), both showing the border beyond which no transient chaos is assumed to persist. Additionally, the average value of the input vector at each layer is plotted for RNNs propagating MNIST data in 1-pixel iterations.

rapidly converging to a common state, to subsequently be perturbed by an input vector that is non-zero for one of the input pairs, leading to a drop in correlation. The critical network appears to show an increase in paired correlation associated with an increase in input data, this correlation does not decay in a similair way as occurs in the ordered network, suggesting that the signal fed at a timestep is retained for longer.

Figure 21 shows the averages and variances over layers for an RNN that is fed the MNIST dataset. Observations of this figure show that the presence of transient chaos extends deeper into the network layers compared to the FFNN setting due to the slower convergence to equilibrium. Due to this, the new transient border is placed at $L = 300$. Interestingly, in this regard, the linear and non-linear networks show opposing behaviour, where the transient chaos in an RNN appears to persist for a longer time w.r.t. the FFNN setup for a nonlinear network, compared to the linear setting where chaos in an RNN persists for a shorter period compared to the FFNN.

Figure 22 shows the single correlations over time, where it can be seen that the input vector has a large influence on the hidden state of the ordered network in a similar fashion as observed in figure 20. This observation also applies to the correlation of the critical network, where the periodicity of the single correlations appears to match that of the Pearson correlation, accounting for the axis offset due to the transient border.

The temporally (w.r.t. the input vector ordering) more uniform CIFAR10 dataset shows that although the ordered network correlation remains equally sensitive to perturbations from the input data, the correlation appears to decay towards a higher value than for the MNIST dataset, possibly in agreement with the theoretical prediction of a higher $\sigma^2_{b,\text{eff}}$.

4.5 Additional observations

The cause of periodic oscillations observed in measurements on non-linear networks was of particular interest. To observe and reason about these oscillations, heat maps of the average hidden state neuron values across layers have been constructed for various network configurations. Despite the author's best efforts, no previous works related to the dynamics of signal propagation and a (visual) interpretation of hidden state dynamics have been found, the works that inspired these observations, and were closest related, are [19, 23, 24]. Although these observations and theories are their own, the author does not claim that they are particularly novel.

Figure 24 shows a handful of examples of hidden states of an FFNN propagating MNIST data, the $\sigma_w = 0.9$ network is in the ordered phase and has converged to its maximally correlated state. The $\sigma_w = 1.0$ shows an interesting phenomenon: the variance plot is a flat line (within the considered heatmap interval).

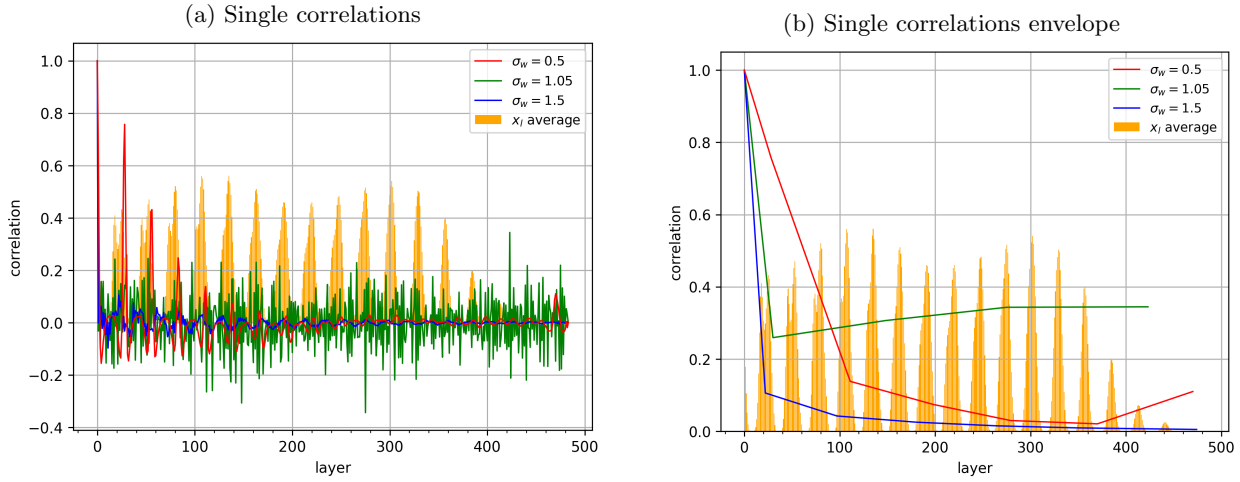


Figure 22: Single correlations (a) and the derived envelope of these correlations (b), included are the average values of the input vectors per layer, for RNNs propagating MNIST data in 1-pixel iterations.

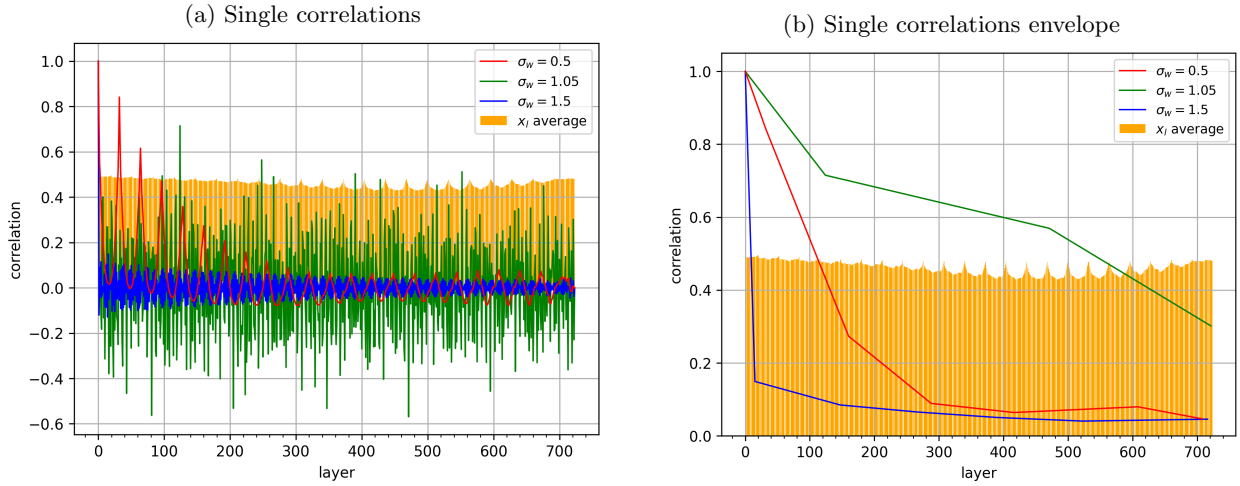


Figure 23: Single correlations (a) and the derived envelope of these correlations (b), included are the average values of the input vectors per layer, for RNN setups propagating CIFAR10 data in 1-pixel steps.

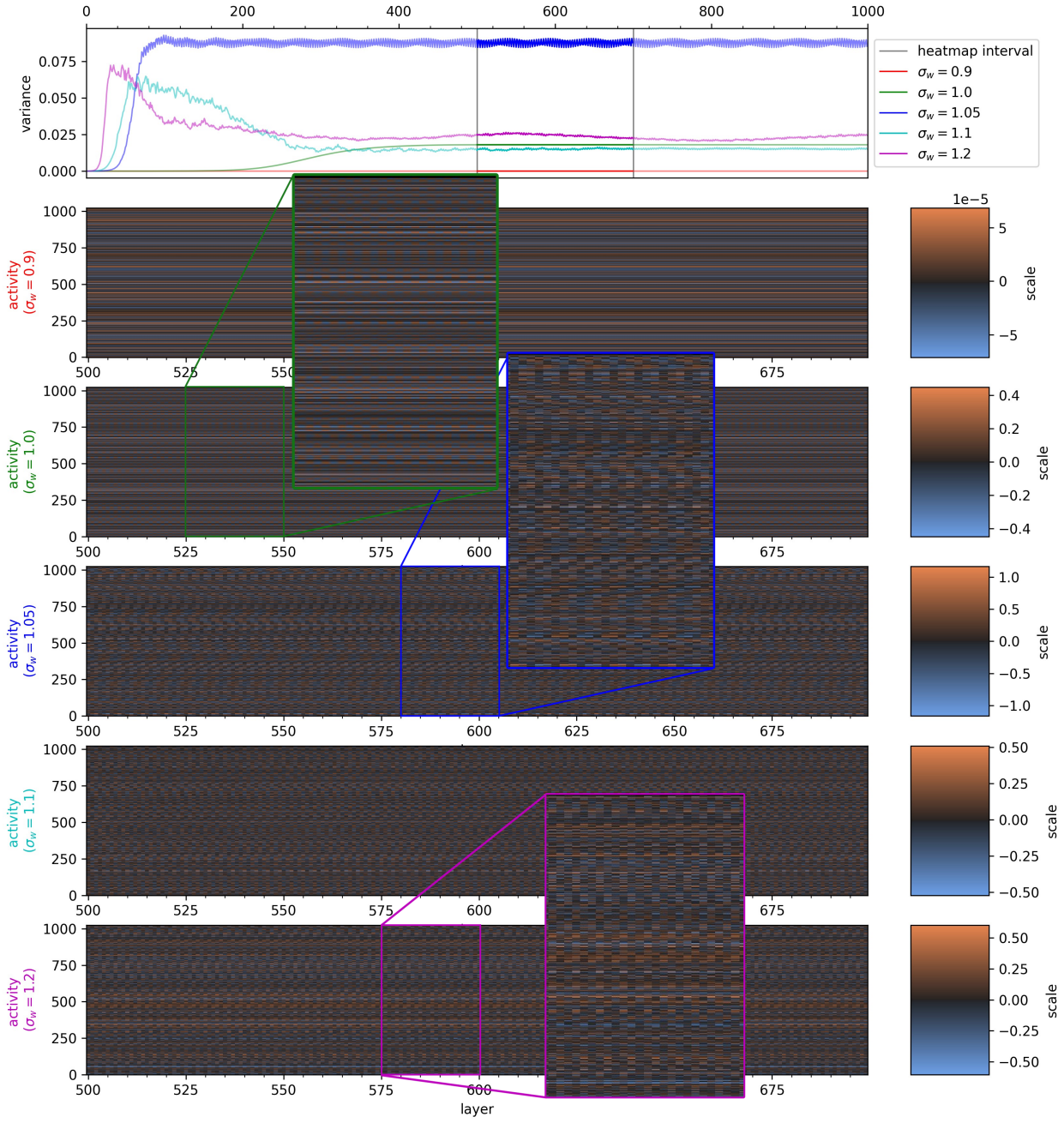


Figure 24: Hidden state heatmap of neuron activity, enlarged sections show a checkerboard pattern (green), offset oscillations (blue), and increasing signs of chaotic behaviour(purple)

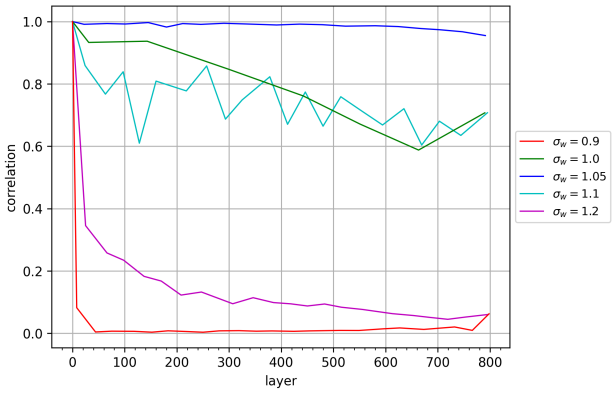


Figure 25: Single-correlation envelopes associated with figure 24

Inspecting the hidden state heatmap, we see that hidden state neurons appear to oscillate rapidly between positive and negative values, forming a rough checkerboard pattern. Such a pattern is likely the cause of the anti-correlated behaviour previously observed. Across all observations, this pattern, as displayed here with a perfectly stable variance, is rare, but the pattern has seen more frequent occurrences in combination with more chaotic/noisy patterns. In the $\sigma_w = 1.05$ network, we observe periodic oscillations in neurons, for which neurons are not aligned in phase. The offset of the oscillation appears to increase with the wavelength of the oscillation. This type of pattern is the most complex type observed. Across various network configurations, this pattern appears most probable to occur around the critical regime and might, therefore, be closely related. Lastly, as highlighted in the $\sigma_w = 1.2$ network, the higher σ_w causes the network to express more chaotic behaviour. A rough periodicity of around 3 layers can still be seen for individual neurons, but the values of individual neurons at a given time appear to show less predictable and more chaotic behaviour. Figure 25 supports the claims w.r.t. the regime of the observed states.

The only two mechanisms affecting the hidden state over time, for the FFNN in this example, are the network weights and the bias. As the bias remains unchanged at $\sigma_b = 10^{-10}$, the only other significant influence is the network weights. The patterns we then see in the network are a product of the probability of incoming neuron connections forming constructive or destructive signals.

Possibly linked to this behaviour is the observation that the periodicity of observed metrics, primarily hidden state average and single correlations, but to a lesser extent also hidden state variance. Appear to shift in relation to how close to the critical point a network lies, the relation appears to be that, as the network becomes critical, periodic oscillations can start to be detected, and the periodicity of these oscillations then continue to increase as the network transitions towards the critical regime.

5 Discussion

The goal of this thesis was to empirically explore if the NN-QFT correspondence is able to predict neural network behaviour accurately. To this extent, it has been partially successful. The main research question: *Does the NN-QFT correspondence provide an accurate prediction of empirically observed NN behaviour?* is difficult to answer conclusively. Theoretical predictions agree with empirical observations in certain regards, while predictions and observations deviate substantially in others. The prediction of the critical point for linear FFNNs appears to be correct. The inaccuracy in the prediction close to the critical point can be attributed to the theory breaking down around it. It has been mentioned that this can be accounted for by a more significant contribution of loop corrections in predicting the correlation length of signals. Additionally, inaccuracies related to correlation decay in RNNs suggest either a lack of influence of $\sigma_{b,\text{eff}}$, or σ_u and $\mathfrak{u}_0$ specifically, on the decay rate, or a shift in the critical point due to the specific input data. Depending on the ability of the QFT framework to account for these interactions, predictions for (linear) RNNs may become more accurate.

The follow-up question: *What network dynamics causes deviations between predicted and observed NN*

behaviour? has also been explored. The primary focus for this was the nonlinear networks, as these showed the greatest mismatch in theoretical and empirical critical points. Through various experiments, it has been shown that behaviour appears to be data-invariant for FFNNs and, to a certain extent, RNNs. Instead, deviations in the observed behaviour of networks are likely due to differences in the network weight matrices originating from stochastic network initialisations. Initial conditions of the network appear to significantly influence the behaviour of transient chaos, which in turn causes different network dynamics to be observed beyond the transient phase. Behaviour associated with an increasing $\sigma_{b\text{eff}}$ causing greater asymptotic correlation values, either through subsequent noise or sequential data input, appears to occur in observed networks.

The exact nature of the observed periodicity of signal propagation through the networks remains unexplained. It is theorised, however, that this periodicity is caused by the biased transitions of neuron-level signals giving rise to specific patterns. Due to the suspected contributions of specific neuron values to the overall emerging patterns, it will be difficult to model this behaviour outside of simulation (i.e. just observing signal propagation of the network). It appears, however, that the envelope of the observed correlation aligns sufficiently well with predicted correlations that such details may be ignored.

Based on the observations made during this thesis, several avenues for future research are possible. Due to time constraints, it was not possible to perform tests on the statistical significance of the observations made. The lack of proper statistical testing is problematic in formally reasoning about these observations' probability and, therefore, the quality of the predictions. The relatively large size of the used datasets should allow for a statistically significant signal propagation representative of the network itself, and the repeated trials allow for some mitigation of the influence of outlier behaviour, but proper statistical testing and outlier detection may make observations more representative of the overall behaviour of its respective parameters. With regard to the theoretical limits of the prediction: The current experiments have been performed up to the maximum depth such that the theoretical $\mathcal{O}(T/N)$ corrections are at their most significant. As network dynamics scale proportionally with depth, future work may compare a range of network depths to observe potential trends of these $\mathcal{O}(T/N)$ corrections and their empirical observations. The networks have been configured to mitigate the significant influence of the effective bias, as predictions indicated this bias only shifts the correlation that the network converges towards. Empirical observations suggest that effective bias also influences the rate of decay. Therefore, an additional avenue of research is to explore the effects of RNN sequences and the decay of correlation further to formalise a better prediction of the correlation decay. Observed network behaviour shows that nonlinear networks appear to show behaviour that would be expected in the theoretical predictions; however, these predictions cannot be verified as the network's σ_w lie beyond the point where theoretical predictions are accurate. It would, therefore, be beneficial to pursue a correction of this divergence. Related to the previous point, it is necessary to account for the decorrelation in the chaotic regime to fully predict corrected depth scales, such as done in [10]. As an overarching conclusion, more observations may be necessary to conclusively reason about the correctness of the NN-QFT correspondence as is.

6 Acknowledgements

Special thanks to dr. Debabrata Panja for his insights and assistance on the subject of dynamical systems.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, conference Name: Proceedings of the IEEE.

[2] G. F. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the Number of Linear Regions of Deep Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/109d2dd3608f669ca17920c511c2a41e-Abstract.html

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," Sep. 2014, arXiv:1409.4842 [cs]. [Online]. Available: http://arxiv.org/abs/1409.4842

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3065386

[5] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994. [Online]. Available: https://ieeexplore.ieee.org/document/279181/

[6] J. Collins, J. Sohl-Dickstein, and D. Sussillo, "Capacity and Trainability in Recurrent Neural Networks," Mar. 2017, arXiv:1611.09913 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1611.09913

[7] D. Mishkin and J. Matas, "All you need is a good init," Feb. 2016, arXiv:1511.06422 [cs]. [Online]. Available: http://arxiv.org/abs/1511.06422

[8] J. Pennington, S. S. Schoenholz, and S. Ganguli, "Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice," Nov. 2017, arXiv:1711.04735 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1711.04735

[9] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, "Exponential expressivity in deep neural networks through transient chaos," Jun. 2016, arXiv:1606.05340 [cond-mat, stat]. [Online]. Available: http://arxiv.org/abs/1606.05340

[10] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, "Deep Information Propagation," Apr. 2017, arXiv:1611.01232 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1611.01232

[11] M. Chen, J. Pennington, and S. S. Schoenholz, "Dynamical Isometry and a Mean Field Theory of RNNs: Gating Enables Signal Propagation in Recurrent Neural Networks," Aug. 2018, arXiv:1806.05394 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1806.05394

[12] K. T. Grosvenor and R. Jefferson, "The edge of chaos: quantum field theory and deep neural networks," Jan. 2022, arXiv:2109.13247 [cond-mat, physics:hep-th, stat]. [Online]. Available: http://arxiv.org/abs/2109.13247

[13] C. G. Langton, "Computation at the edge of chaos: Phase transitions and emergent computation," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1, pp. 12–37, Jun. 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016727899090064V

[14] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks."

[15] H. Sompolinsky, A. Crisanti, and H. J. Sommers, "Chaos in Random Neural Networks," *Phys. Rev. Lett.*, vol. 61, no. 3, pp. 259–262, Jul. 1988, publisher: American Physical Society. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.61.259

[16] G. Yang, J. Pennington, V. Rao, J. Sohl-Dickstein, and S. S. Schoenholz, "A Mean Field Theory of Batch Normalization," Mar. 2019, arXiv:1902.08129 [cond-mat]. [Online]. Available: http://arxiv.org/abs/1902.08129

[17] "Mean-field theory," Sep. 2023, page Version ID: 1175947535. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Mean-field_theory&oldid=1175947535

[18] Y.-C. Lai and T. Tél, *Transient Chaos*, ser. Applied Mathematical Sciences. New York, NY: Springer New York, 2011, vol. 173. [Online]. Available: http://link.springer.com/10.1007/978-1-4419-6987-3

[19] P. Verzelli, L. Livi, and C. Alippi, "A CHARACTERIZATION OF THE EDGE OF CRITICALITY IN BINARY ECHO STATE NETWORKS," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. Aalborg: IEEE, Sep. 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8516959/

[20] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images."

[21] L. Zhang, L. Feng, K. Chen, and C. H. Lai, "Edge of chaos as a guiding principle for modern neural network training," Jul. 2021, arXiv:2107.09437 [nlin, physics:physics]. [Online]. Available: http://arxiv.org/abs/2107.09437

[22] R. Hardstone, S.-S. Poil, G. Schiavone, R. Jansen, V. V. Nikulin, H. D. Mansvelder, and K. Linkenkaer-Hansen, "Detrended Fluctuation Analysis: A Scale-Free View on Neuronal Oscillations," *Front Physiol*, vol. 3, p. 450, Nov. 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3510427/

[23] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, and H. Qu, "Understanding Hidden Memories of Recurrent Neural Networks," in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2017, pp. 13–24.

[24] B. Hanin, "Which Neural Net Architectures Give Rise to Exploding and Vanishing Gradients?"

A Training results

Figure 1 show the resulting accuracies of training the network using the ADAM optimizer with a learning rate of 1e−4, a batch size of 512 (100 training steps/epoch), across 10 epochs.

σ_w	accuracy
0.5	0.11
0.75	0.11
0.9	0.40
0.95	0.58
1.0	0.72
1.05	0.71
1.1	0.55
1.2	0.20
1.5	0.11

Table 1: Training accuracies